Story-level multimodal generativeAI: from understanding to generating visual data using multiple modalities

Vicky Kalogeiton

Ecole Polytechnique
Institut Polytechnique de Paris

Computer Science Department and Laboratory

Habilitation à diriger des recherches

Vicky Kalogeiton

Story-level multimodal generativeAI: from understanding to generating visual data using multiple modalities

---

Reviewers

Raoul de Charette	INRIA
Dimitris Samaras	Stony Brook University
Josef Sivic	Czech Technical University

Jury members

Matthieu Cord	Sorbonne University
Juergen Gall	University of Bonn
Vincent Lepetit	ENPC
Elisa Ricci	University of Trento
Jakob Verbeek	FAIR, Meta

Acknowledgements

This work would not have been possible without a very large number of people, starting from the many collaborators who directly or indirectly contributed to this work: Manuel Marin-Jimenez, David Picard, Marc Christie, Maria Vakalopoulou, Ivan Laptev, Andrew Zisserman, Stéphane Lathuilière, Slim Essid, Subhankar Roy, Victoria Fernández Abrevaya, Steve Oudot. It has been a real pleasure and lots of fun working with you.

I would also like to thank the reviewers of this manuscript Raoul de Charette, Dimitris Samaras, and Josef Sivic as well as the jury members Matthieu Cord, Juergen Gall, Vincent Lepetit, Elisa Ricci and Jakob Verbeek.

I am also extremely grateful for the support of my colleagues at the LIX laboratory of École Polytechnique including members of the VISTA/GeomeriX teams, Jessica Gameiro, Fanny Sabatier, Frederic Ayrault, Samuel Mimram, Gilles Schaeffer, and special thanks to Maks Ovsjanikov, Marie-Paule Cani and Damien Rohmer.

Over the years, I had the chance to interact with many excellent students and post-docs. I would like to specially mention Robin Courant, Nicolas Dufour, Julie Mordacq, Léo Milecki, Xi Wang, and Zhisong Liu who significantly contributed to the research presented in this manuscript.

Finally, I would like to thank my family and Ketchup, Lizi and especially my partner, my rock, Quentin.

Contents

Chapter 1

Introduction

“Once upon a time in a faraway land, there was a dragon called Zoe and she was different…’ I started the bedtime story while tucking my daughter into bed.

As the words left my lips, I imagined how magical it would be if the walls of her room could come alive with the scenes I was describing, adapting in real time to the story’s twists and turns. The colors would change with the mood, the characters would look just as she imagines them, and the music would sync with the emotions I wanted to evoke. More than just a simple projection, this system would understand the story I am telling including its tone, rhythm, and emotional beats and would create an immersive experience personalized to both our preferences, from my love for Ketchup to her love of orange. This vision has been the core of my research: to create an automatic, real-time visual storytelling system that seamlessly aligns with spoken narratives, bringing stories to life in a deeply personal and engaging way.

However, achieving this goal is far from simple. Computers, after all, cannot feel. Every day, video processing systems like those behind YouTube, Netflix, and Amazon handle countless hours of content, decoding, streaming, and encoding without truly understanding any of it. They process videos as mere sequences of frames, oblivious to their rich narratives and emotions. Yet, just like how a well-directed movie is more than a series of scenes, the stories we tell –whether in a movie or a bedtime story– are more than just words. They are complex patterns with emotions, intentions, and carefully considered details, from the choice of colors and camera angles to the pacing of dialogue and the interplay of light and shadow. My work aims to bridge this gap between raw processing and true understanding, pushing AI to understand and create the narration and emotional escalation that filmmakers achieve.

To do this, my research connects two critical components: long-term story-level movie understanding and dynamic text-to-image generation. The first component involves enabling the system to grasp the overarching narrative structure, understanding not just isolated events but how they contribute to the story as a whole. This requires the system to leverage multiple modalities, such as text, audio, and visual cues, to interpret the story in a way that considers the continuity and development of characters, themes, and emotions – much like how a director uses various cinematic techniques to guide the audience’s emotional journey. The second component focuses on the generation of visual content that is contextually relevant and emotionally aligned with the story being told. This is inspired by how a filmmaker carefully edits scenes, selects camera angles and adjusts lighting to evoke specific emotions. By merging these two components, my objective has been to create a system that not only generates images or sequences but also tells a coherent, engaging story that resonates on multiple levels.

This task is challenging because current systems excel primarily in low-level video understanding, such as recognizing objects or analyzing short clips. They can identify what is happening in a scene but often fail to grasp why it is happening or how it fits into a broader narrative. Most AI models treat videos as linear sequences of frames, missing the intricate storytelling decisions that make films compelling. For example, in traditional filmmaking, the placement of a camera or the timing of a cut can significantly alter the viewer’s perception and emotional response. My research focuses on these challenges, aiming not only at following the story but also at understanding and replicating the narrative techniques that make storytelling such a powerful tool for communication.

Ultimately, the objective of this work is to push the boundaries of AI-driven content generation, creating systems that can tell stories: stories that are not only visually impressive but also emotionally engaging and personally meaningful. By combining multimodal understanding with advanced text-to-video generation techniques, we hope that storytelling is experienced not just through words but through a rich, dynamic interplay of visuals, sounds, and emotions. In doing so, we aim to create a world where every bedtime story can be as vivid and magical as the imagination that inspires it.

1.1 Objective

The primary objective of this work is to advance the understanding and generation of visual content through the integration of multiple modalities. We aim to bridge the gap between traditional low-level video processing and high-level semantic understanding, moving beyond mere frame-by-frame analysis to comprehending and generating coherent, contextually rich visual stories. The work is twofold: first, to enhance the understanding of human-centric interactions and scene comprehension in videos by leveraging multimodal data; and second, to utilize this understanding to generate visually compelling content that captures the desired style of the human, such as artistic or photorealistic.

Specifically, our first objective is to develop approaches that improve the understanding of long-form visual data, particularly edited multimodal videos (i.e. films). We focus on understanding the underlying narrative structures, the motivations and intentions of characters, and the emotional content conveyed through cinematic techniques. By integrating different modalities—such as visual cues, audio, and text—we aim to create a holistic approach to movie understanding that can recognize relationships between characters and the story they collectively tell.

The second objective is to leverage the insights gained from multimodal movie understanding for the generation of visual content. This involves creating methods that can produce images and ideally video sequences that are not only realistic but also coherent with the intended narrative. We seek to develop techniques that allow for conditional generation based on textual descriptions, semantic information, and other modalities, ensuring that the generated content is contextually appropriate and emotionally resonant. By doing so, we aim to push the boundaries of AI-driven content creation, enabling systems to generate visual stories that are rich in narrative depth and cinematic quality.

A key aspect of this objective is the emphasis on character and scene evolution within generated content. We aim to create systems capable of generating dynamic, evolving narratives where characters and scenes develop in a way that is consistent with the overall story. This includes the generation of complex camera movements and scene compositions that reflect the emotional state of characters, thereby enhancing the audience's engagement and emotional connection with the content.

Finally, we aim to apply these advanced multimodal techniques to specialized domains such as medical imaging, where the integration of multiple data types can lead to significant improvements in diagnostic accuracy and treatment planning. By extending our work into the medical field, we seek to demonstrate the broad applicability and impact of multimodal content generation, from entertainment and storytelling to critical applications in healthcare.

1.2 Motivation

There are several applications of our work, especially in our society.

Multimodal video understanding. The potential applications of video understanding and visual content generation are numerous and diverse, and continued research in these fields could lead to a wide range of exciting new capabilities and possibilities. Below, we include some societal impacts for both video understanding and generation, as improving one helps improve the other. Note that in most cases, the research impact rely on multimodal data, i.e., audiovisual and textual signals.

The primary application of the research on video understanding presented in Chapter 2 is to improve accessibility for individuals with disabilities. Improving the ability to understand videos could lead to the creation of tools and algorithms that can automatically generate captions, audio descriptions, or other accessibility video features and more generally to the creation of content specifically tailored to the needs of individuals. This could make a wide range of content more accessible to individuals who are deaf or hard of hearing (by automatically generating captions), blind or low vision (by automatically generating audiovisual content), or have other disabilities or cognitive impairments (by abstracting or simplifying visual concepts). For instance, videos generated from audio descriptions or captions could be accessed by individuals who are blind or have low vision [178, 179]; also, if simplified or abstracted, they can become more accessible to individuals with cognitive impairments.

In addition, accurate and reliable visual content analysis could be used in entertainment applications. For instance, the work presented in Chapter 2 could be used to automatically generate storylines in videos [26, 86] or to predict the evolving nature of characters in movies [194, 195, 86]. This could make visual content more user-friendly for a wide range of audiences.

Multimodal visual content generation. In parallel, the work of this thesis on generation (Chapter 3) could be used to create videos from written news reports, stories, or other text-based content [181, 73], given that it often relies on other modalities, such as text, semantic masks or other data sources [72]. For example, a generation method could take a written description of an event or a scene[178, 51] and generate the corresponding visual content [326, 72].

Another societal application is to create personalized visual experiences [73, 180]. Video analysis and generation could lead to content tailored for individual users based on their interests, preferences, or other characteristics. For instance, a generation method could analyse a user's browsing history, social media activity, or other data to create a video that is personalized to their interests [326]. This could be used to create personalized news reports, entertainment experiences or educational content.

Multimodal Medical imaging. The use of computer vision in medical imaging can have a significant positive impact on society. Accurate and reliable video analysis is a valuable source in medical imaging, for automatically analyzing medical images or videos [204, 205] to identify abnormalities or to automatically generate reports from visual data that can assist medical professionals. Specifically, the technology developed in this thesis could potentially contribute to benefiting individuals and communities around the world in: (a) Improved diagnosis and treatment of diseases: By using computer vision techniques to analyze medical images, doctors and other healthcare providers could potentially identify diseases and other medical conditions more accurately and quickly [205]. This could lead to earlier and more effective treatment of diseases, which could improve patient outcomes and save lives; (b) Improved patient experience with more personalized and tailored care by using interpretable techniques [99], thus leading to a better experience for patients; and (c) Improving rare medical issues: Modern AI techniques require big data, which is unrealistic when it comes to medical imaging due to patients privacy and rarity of diseases. For this, the techniques she developed require no or few data for prediction and instead exploit the underlying structure of medical imaging for compact and informative representations [204, 205].

Economic impact. The research work of this thesis is a step towards marketing complex AI systems. The use cases of our work are numerous, including improving content-based recommendation video systems and helping videographers or impaired people. Moreover, her findings can help convert videos to audio or textbooks for disabled people (impaired vision or hearing) or can help develop language learning techniques by watching movies in other languages. Furthermore, by improving the automatic understanding and generation of systems, both the need for annotated data and consequently, the deployment cost of deep learning technologies reduce. This crucial step towards marketing may have important economic consequences in terms of job creation (design, production, retail, repair) and investments (further development, targeted conception startups).

Environmental impact. All work presented here requires training models in GPUs. In addition to local equipment (mostly at Polytechnique), we have relied on the French Jean Zay GPU cluster. France relies on nuclear energy, having greener energy than average, with 50-80g CO_2 for each kWh produced. We tried to optimise experiments and use more than 80% of the GPUs power.

1.3 Challenges

The goals of this thesis are accompanied by several challenges, coming from both the complexity of multimodal data and the demanding nature of visual content generation.

Balancing modalities. One of the primary challenges is the integration of multiple modalities in a way that enhances understanding without overwhelming the system. Different modalities, such as visual, textual, and auditory data, carry different types of information and often operate at different levels of abstraction. Balancing these modalities to create a cohesive understanding of a scene, while ensuring that no single modality dominates or distorts the final interpretation, is a complex and ongoing challenge.

Long-term movie understanding. Achieving high-level reasoning and long-term video understanding is an open challenge in videos. Up to the recent evolution of Vision Language Models [305], most systems used to excel at short-term, low-level tasks, such as action recognition or object detection, but struggled with understanding the broader narrative context within which these actions occur. Edited videos, with their deliberate cuts, transitions, and scene compositions, require an understanding that goes beyond recognizing what is happening in a single frame or shot. The challenge is to develop methods that can interpret these high-level cinematic techniques and understand the motivations, intentions, and emotional arcs that drive the story.

Controllable and user-friendly conditional generation. Ensuring that the generation covers fine appearance details and backgrounds following the input condition while at the same time being controllable and user-friendly remains a critical challenge.

Handling noisy conditions for content generation. Generative models and especially diffusion models typically struggle when the conditional information is noisy or unreliable [160, 361], often leading to suboptimal image generation. The difficulty lies in ensuring that the model can still generate high-quality visual content without discarding valuable, albeit noisy, data points.

Camera-character relation for visual content generation. Generating camera trajectories that are both realistic and artistically coherent with the narrative and character actions is highly complex. It requires the model to understand the input edited video (e.g. movie) and synchronize multiple modalities, including text, motion, and spatial dynamics. Also, generalizing these capabilities to broader contexts (e.g. bedtime story) and varied storytelling styles (e.g. different directors) remains an open question.

Missing modalities or missing data. In domains like medical imaging, additional challenges arise due to the need for high accuracy and the limited availability of annotated data. In such contexts, the main challenge is to develop multimodal systems that can handle incomplete or noisy data while still providing reliable and clinically relevant results.

1.4 Contributions and Outline

This thesis addresses the above challenges, and here we outline our main contributions.

I. Multimodal video understanding (Chapter 2). Chapter 2 focuses on fundamental aspects of multimodal recognition, which play a crucial role in understanding human-centric interactions and scene comprehension. The goal is to leverage multiple modalities to enhance the understanding and interpretation of complex visual data, split in two directions: (1) Human-centric video understanding (Section 2.1). We first work multimodal human clustering in videos (**BMVC 2020** [137] and **ICCV-W 2021** [26]) and then, we work on multimodal human interactions (**CVPR 2019** [194] and **TPAMI 2021** [195]). (2) Scene understanding (Section 2.2). Building upon the previous subsection, in Section 2.2, we discuss scene understanding, where the focus lies on comprehending the overall context and content of a scene. Specifically, we first focus on the cinematic perspective by learning funny moments in videos (**Oral, best honorable paper award ACCV 2022** [178], and **IJCV 2024** [179]). Then, we present our work on visual question answering in videos (currently *under submission*).

II. Multimodal visual content generation (Chapter 3). The goal in Chapter 3 is to leverage multiple modalities to enhance the generation of complex visual scenes, split in two directions: (1) Conditional image generation (Section 3.1). We present our work on multimodal conditional image generation, either using semantic information with GANs or with emphasis on text-to-image generation with diffusion. First, we present our work on semantically conditioned image generation (**ECCV 2022** [73]). Second, we present our Coherence-Aware Diffusion (CAD) work, a novel method that integrates coherence in conditional information into diffusion models, allowing them to learn from noisy annotations without discarding data (**CVPR 2024** [72]). Third, we present a comprehensive analysis and insights into Classifier-Free Guidance (CFG) weight schedulers, the default way for image generation (**TMLR 2024** [326]). (2) Multimodal motion generation (Section 3.2). We present our work on the Exceptional Trajectories (E.T.) dataset for generating complex camera motion trajectories from textual captions that describe the relation and synchronisation between the camera and characters (**ECCV 2024** [50]).

III. Multimodal medical applications (Chapter 4). Multimodality has gained attention in the medical domain, by integrating imaging or video modalities with biomedical signals or health records. Yet, two challenges remain: balancing the contribution of modalities, especially in cases with a limited amount of data available, and tackling missing modalities. Here, we address these challenges in two directions. (1) Renal transplant failure prediction (Section 4.1). We first introduce the masked-transformer based CosEmb [204, 205] for forecasting transplant rejections from MRI post-transplantation exams (**MICCAI 2022** [204], **MIDL 2022** [205]). Furthermore, inspired by the success of LLMs, in [206] we introduced the MEDIMP model that learns meaningful multi-modal representations of renal transplants by also incorporating structural clinicobiological data after translating them into text prompts. Note, that this has been one of the first works on medical imaging that exploited and explored text prompting (**MIDL 2023** [206]). (2) Multimodal learning for detecting physiological changes under missing modalities (Section 4.2). In this work, we introduce the multimodal ADAPT with two components: aligning all modalities in the space of the strongest, richest modality to learn a joint embedding, and a Masked Multimodal Transformer that handles missing modalities. We experiment on detecting stress physiological changes in individuals and outperform the state of the art (**MIDL 2024** [208] and **CVPR-W 2024** [209]).

The work presented here is separated from the work I did during my PhD, which included only spatiotemporal action recognition in videos and finished in 2017. It has been produced by the brilliant students I co-supervise(d): Robin Courant, Nicolas Dufour, Julie Mordacq, Léo Milecki, Andrew Brown, and Ridouane Ghermi and my amazing collaborators: postdocs X. Wang, and Z.S Liu and researchers M. Marin-Jimenez, D. Picard, M. Christie, M. Vakalopoulou, S. Oudot, I. Laptev, and A. Zisserman.

Other works. Although this material represents a large part of my work during the past 6 years, there is a number of articles that I have published that will not be mentioned below, either due to the lack of space or due to the difference in content. The articles, written after I started my position at École Polytechnique (2020) and **not** discussed in this document, include:

1. Your diffusion model is an implicit synthetic image detector. Xi Wang, Vicky Kalogeiton. In **ECCV-W** 2024
2. Bridging Text and Image for Artist Style Transfer via Contrastive Learning. Zhi-Song Liu, Li-Wen Wang, Jun Xiao, Vicky Kalogeiton. In **ECCV-W** 2024
3. Conditional Gradient-based Textual Inversion. Xi Wang, Vicky Kalogeiton. In **ECCV-W** 2024
4. Collaborating Foundation models for Domain Generalized Semantic Segmentation. Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, Stéphane Lathuilière. In **CVPR** 2024 [Project page], [*arXiv*:2312.09788]
5. LEAD: Latent Realignment for Human Motion Diffusion. Nefeli Andreou, Xi Wang, Victoria Fernández Abrevaya, Marie-Paule Cani, Vicky Kalogeiton. In **CVPR-W** 2024
6. Learning the What and How of Annotation in Video Object Segmentation. Thanos Delatolas, Vicky Kalogeiton, Dim P. Papadopoulos. In **WACV 2024** [Project Page] [*arXiv*:2311.04414v2]
7. BluNF: Blueprint Neural Field. Robin Courant, Xi Wang, Marc Christie, Vicky Kalogeiton. In **ICCV-W** 2023 [Project page]
8. EVA-VOS: Efficient Video Annotation for Video Object Segmentation. Thanos Delatolas, Vicky Kalogeiton, Dim P. Papadopoulos. In **ICCV-W** 2023 [Project page]
9. Name Your Style: Text-Guided Artistic Style Transfer. Zhi-Song Liu, Li-Wen Wang, Wan-Chi Siu, Vicky Kalogeiton. In **CVPR-W** 2023
10. One-shot Unsupervised Domain Adaptation with Personalized Diffusion Models. Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, Stéphane Lathuilière. In **CVPR-W** 2023 [Project page]
11. Reward Function Design for Crowd Simulation via Reinforcement Learning. Ariel Kwiatkowski, Vicky Kalogeiton, Julien Pettré, and Marie-Paule Cani. In **MIG** 2023
12. Transformers and Visual Transformers. Robin Courant, Maika Edberg, Nicolas Dufour, Vicky Kalogeiton. **Book Chapter** In *Machine Learning for Brain Disorders* (editor: O. Colliot), Springer, 2023
13. Understanding reinforcement learned crowds. Ariel Kwiatkowski, Vicky Kalogeiton, Julien Pettré, and Marie-Paule Cani. In **MIG** 2022
14. A Survey on Reinforcement Learning Methods in Character Animation. Ariel Kwiatkowski, Eduardo Alvarado, Vicky Kalogeiton, C Karen Liu, Julien Pettré, Michiel van de Panne, Marie-Paule Cani. In **Eurographics STAR** 2022
15. UGaitNet: Multimodal gait recognition with missing input modalities Manuel J. Marin-Jimenez, Francisco M. Castro Payán, Vicky Kalogeiton, Nicolas Guil. In **TIFS** 2021
16. High-Level Features for Movie Style Understanding Robin Courant, Christophe Lino, Marc Christie, Vicky Kalogeiton. In **ICCV-W** 2021 (**best paper award**)
17. Multiple Style Transfer Via Variational Autoencoder. Zhi-Song Liu, Vicky Kalogeiton, Marie-Paule Cani. In **ICIP** 2021. [Project page]
18. Multimodal Gait Recognition Under Missing Modalities. Ruben Delgado-Escano, Francisco Castro, Nicolas Guil, Vicky Kalogeiton, Manuel Marin-Jimenez. In **ICIP** 2021 [Project page]
19. Real-Time Active SLAM and Obstacle Avoidance for an Autonomous Robot Based on Stereo Vision. Vicky Kalogeiton, Kostas Ioannidis, Georgios Ch. Sirakoulis, Elias B. Kosmatopoulos. ***Cybernetics and Systems*** 2019 [while at Oxford]
20. Smoothing the Path Towards Retrieval. Andrew Brown, Weidi Xie, Vicky Kalogeiton, Andrew Zisserman. In **ECCV** 2020 [Project page] [while at Oxford]

Chapter 2

Multimodal Video Understanding

This Chapter focuses on fundamental aspects of multimodal recognition, which play a crucial role in understanding human-centric interactions and scene comprehension. The goal is to leverage multiple modalities to enhance the understanding and interpretation of complex visual data.

Human-centric video understanding (Section 2.1). In Section 2.1.1 we present our work on person-clustering in videos, i.e. grouping characters according to their identity [137, 26]. Previous methods focused on the narrower task of face-clustering, and ignored cues such as the person's voice, their overall appearance, and the editing structure of videos. Instead, we introduce methods for automatic video face and person-clustering [26] using multiple modalities (face, body, and voice). We also propose two large-scale datasets, Friends for face clustering and VPCD for person clustering in videos. The code and annotations are available online. This work has been published at BMVC 2020, ICCV-W 2021 [137, 26].

Capturing 'mutual gaze' is essential for understanding social interactions. To this end, Section 2.1.2 presents our LAEO-Net++ work for detecting people *Looking At Each Other (LAEO)* in videos by reasoning about the whole track. Moreover, we introduce two new LAEO datasets: UCOLAEO and AVA-LAEO. Our experiments demonstrate the capabilities of LAEO-Net++ which sets the new state of the art. We also apply LAEO-Net++ to a social network, where we automatically infer the social relationship between pairs of people based on the frequency and duration that they LAEO, and show that LAEO can be a useful tool for guided search of human interactions in videos. The code and datasets are available online. This work has been published in CVPR 2019, TPAMI 2021 [194, 195].

Scene understanding (Section 2.2). Automatically understanding funny moments is challenging, as they relate to various features, such as body language, dialogues and culture. In Section 2.2.1, we propose FunnyNet-W, a multimodal transformer-based model that predicts funny moments in videos. Unlike most methods that rely on ground truth data in the form of subtitles, in this work, we exploit modalities that come naturally with videos: (a) frames as they contain visual information, (b) audio as it contains higher-level cues associated with funny moments, such as intonation, pitch and pauses and (c) text automatically extracted with a speech-to-text model as it can provide rich information when processed by a Large Language Model. Extensive analysis show that FunnyNet-W successfully exploits visual, auditory and textual cues, while setting the new state of the art. Our code, models, and dataset are available online. This work has been published in ACCV 2022, IJCV 2024 [178, 179].

Existing multimodal video understanding datasets and tasks have notable limitations. Most datasets contain only short videos with limited events and narrow narratives, such as datasets with instructional and egocentric videos. Although movie datasets offer richer content, they are often limited to short-term tasks, lack publicly available videos and often present *data leakage* (i.e. LLMs have prior knowledge about them). To address these, in Section 2.2.2, we propose the Short Film Dataset (SFD) with over 1k publicly available amateur movies, a wide variety of genres and minimal data leakage issues. SFD offers long-term *story-oriented* video tasks with multiple-choice and open-ended question answering. Our dataset and code are publicly available. This work is currently under submission.

2.1 Human-centric video understanding

In this section, we address the problem of human-centric understanding in videos using multiple modalities split into two directions: person-clustering (Section 2.1.1), and detecting people Looking At Each Other (LAEO) in videos (Section 2.1.2).

2.1.1 Human clustering

Clustering people by identity in videos has numerous real-world applications, such as person-specific browsing, video organization, automatic cast listing, and story understanding, all without explicit identity labeling [48, 78, 303, 333, 135, 137, 302]. Effective person-clustering can greatly reduce annotation costs. Most methods rely solely on faces, which has two drawbacks: (i) they overlook cues like voice, appearance, and editing structure, and (ii) they limit utility for story understanding, where knowledge of all characters is needed, not just those with visible faces. For example, voice can clarify poor-resolution faces, and hair or clothing can link a person seen from behind to one speaking in another shot.

To this end, in this work our goal is to cluster person-tracks showing the entire body by identity in movies and TV shows to support story understanding, using all available cues (face, voice, body appearance, editing structure), including tracks without visible faces. Modalities from the same person are both redundant and complementary, helping solve two clustering challenges: achieving pure clusters (tracks from a single person) and merging clusters without contamination. Agreement across multiple modalities can ensure purity, while a common modality, like voice, can merge otherwise unmergeable clusters. Methods that rely on a single modality for merging inevitably sacrifice purity.

Related work. *I. Video face clustering* has been a research focus for years [303, 48, 78, 333]. Early methods used handcrafted features and video continuity. For example, [48] applied metric learning with automatically generated face pairs, and[333, 332] used Hidden Markov Random Fields for clustering face tracklets. WBSLRR [338] incorporated prior knowledge in a block-sparse low rank representation. Later methods relied on CNN-based face features [275, 278, 302, 303]. For instance, TSiam and SSiam [276, 277] mined pairs by sorting distances. FINCH [266] used hierarchical clustering via first-neighbor relations, while CCL [278] used pseudo-labels from pure tracks. Most methods required domain-specific training and used hierarchical agglomerative clustering (HAC). Our method, by contrast, requires no training, as it integrates must-link and cannot-link constraints directly in the clustering.
II. Related Datasets. Existing face-clustering datasets [76, 259, 137, 226, 85, 49] share common limitations: (a) small size, usually a movie or a few TV episodes; (b) under-representation of many demographics; and (c) only face annotations, making them unsuitable for multimodality. Story understanding [117, 11] and person-search [116] datasets with body or face annotations exist but lack audio [116, 117] or have only partial annotations [117, 11]. None include labeled voice utterances. Our dataset, by contrast, includes six TV shows and movies with diverse characters and multi-modal annotations for all characters.

Approach overview. We introduce two new methods for clustering in videos: video face clustering which uses a single modality and video person-clustering, which builds on top of the face one by using multiple modalities – face, voice, and body appearance.

For face clustering, we propose **C1C** (Figure 2.1): a hierarchical agglomeration clustering (HAC) approach [266] that imposes must link and cannot link constraints [48, 333] acquired in a self-supervised manner: instances from the same track *must* be linked as they represent the same character, while concurrent tracks *cannot* be linked, as they represent different characters. Unlike standard HAC methods where each partition Γ merges only one instance with existing clusters, it groups several instances using first Nearest Neighbor (NN) relations at the same partition. In the first partition, it links samples through first NN relations, while all following partitions link the clusters from the previous step.

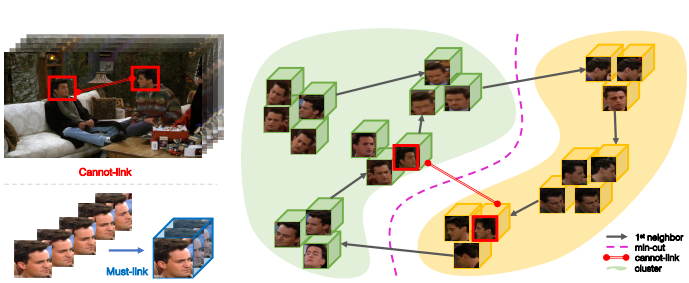


Figure 2.1: **C1C: Constrained video face clustering**. (Left) Must-link constraint: instances from the same track must be linked as they represent the same character. Cannot-link constraint: concurrent tracks appearing in the same frame cannot be linked, as they represent different characters. (Right) C1C Clustering given first NN relations and constraints.

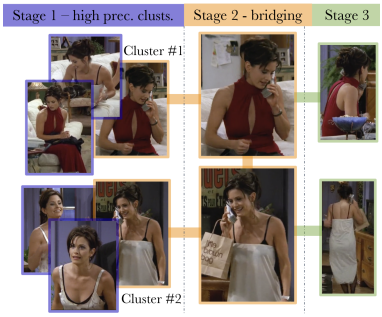


Figure 2.2: **MuHPC video-face clustering.** Stage 1 clusters of the same character; the face modality cannot merge them as they depict frontal faces (below) and profiles (top). Voice in Stage 2 merges them; Stage 3 merges face-less bodies.

(a) Friends Face clustering

(b) About Last night

(c) TBBT Person clustering

(d) Friends

Figure 2.3: **Proposed video face and person clustering datasets.** (a) Examples from the Friends dataset, captured in several scenes and viewpoints, including back of heads. (b-d) VPD dataset examples. VPD contains face, body and voice tracks annotated for many characters. A diverse set of characters are captured in a variety of scenes, showing backs of bodies, and over-the-shoulder shots (magenta, cyan). When speaking, we also include a voice-track (blue signal below body-tracks).

For person clustering we propose the **MuHPC** (Figure 2.2), a three-stage approach by exploiting three modalities; but, our approach can scale to any number of modalities. Stage 1 creates high-precision clusters using a single modality, here face. We group person-tracks that share a first nearest neighbour(NN) using multiple iterations of HAC, as described previously [266, 137, 126]. We follow this trend subject to two additional *constraints*: a cannot-link constraint for concurrent tracks (as in [137] based on [16, 48]), and a threshold on the maximum NN distance. Stage 2 exploits multimodality to *bridge* clusters that were otherwise unmergeable by the single face modality with a conservative threshold; in particular, by requiring that different modalities (i.e. face and voice) concur on the merge. Stage 3 clusters tracks without visible faces, which are therefore not yet clustered by the first two stages. Constraints from the editing structure (neighboring shots) and a threshold on body features (so that they depict the same person with the same clothing) are used to link face-less person-tracks to clusters with faces.

Proposed Datasets. The near-perfect performances [278, 302] of DL methods for video face clustering seemed, at first glance, to have solved the problem. However, when this work was proposed, the set of datasets was limited; for example, they only considered the principal characters and ignored the secondary or background characters [259, 76, 229]; this was hiding some of the shortcomings of existing clustering methods [277, 369, 276, 278]. To address these, we introduced a new video face clustering dataset with approximately 18k annotated heads for 49 characters (Figure 2.3(left)). Compared to previous datasets at that time, ours was larger, contained many secondary characters, and was more challenging as it also contained back-of-the-head detections.

To evaluate the multi-modal person-clustering task, we require a dataset with person-level annotations. However, there were very few such datasets due to the previous emphasis on face-clustering and moreover, most face-clustering and labelling datasets, such as Buffy [76], TBBT [259] and Ours [137], were based on TV material with limited diversity in skin color. For these reasons, we also introduced a new video person clustering dataset, as shown in Figure 2.3 (right), where we: (i) re-purposed multiple existing face datasets by adding person-level multi-modal annotations (e.g. all person-tracks and voice utterances); and (ii) included different TV shows and films to address this lack of diversity. Our dataset consists of visually disparate videos, and includes body-tracks, face-tracks when visible; and voice utterances when speaking, for all annotated characters. More details in [26].

Experiments. **Metrics**. We measure each metric at the episode level and then average the results over all episodes. We use Weighted Cluster Purity (WCP) that weights the purity of a cluster by the number of tracks belonging to it and and Normalized Mutual Information (NMI) [193] that measures the trade-off between clustering quality and number of resulting clusters.
Quantitative results. *Face clustering.* We compare the state of the art at that time, i.e. FINCH [266] and BCL [302] to our proposed C1C [137], tailored to faces, and MuHPC [26], both for face only setting, i.e. comparable to the other methods, and multimodal setting. Our MuHPC significantly outperforms the state of the art in all metrics, as it avoids incorrect merges, hence maintaining cluster purity. For instance, both NMI and WCP increase by over +10% averaged across all datasets. *Person clustering.* We compare against two strong baselines, one inspired by person Re-ID [375, 376, 164] that uses C1C to cluster body rather than face features, and one that uses C1C to cluster people. For all metrics, our full method significantly outperforms the strongest baseline by on average 6.1% in WCP and 11.8% in NMI. This validates that using all available video cues, such as multiple modalities and the editing structure of videos aids video person-clustering substantially.

Discussion. We proposed two novel methods for face clustering and multi-modal person-clustering in videos. For evaluation, we introduced two datasets, one for face and one for person clustering, the largest and most diverse datasets of their kind at that time. We showed that using all available video cues is essential for person clustering, leading to significant improvements in person clustering and state-of-the-art performances for face clustering. Preliminary experiments in [26] show that person clustering can help identify concurrencies of characters and their potential interactions. We hope this can support downstream story understanding tasks such as learning interactions and relationships [151].

Impact. The work described here could enhance video analysis capabilities in various fields, such as security, entertainment, and social media. Improved person clustering can lead to more efficient surveillance systems, enabling quicker identification and tracking of individuals in crowded or complex environments, thus improving public safety. In entertainment, it could revolutionize content personalization by accurately identifying characters across scenes, enhancing viewer experiences. However, it also raises privacy concerns, as the ability to cluster and track individuals in video footage may be misused for unauthorized profiling, highlighting the need for ethical considerations in its application.

2.1.2 Human interactions

Eye contact or 'mutual gaze' is an important part of the non-verbal communication between two people [182]. The duration and frequency of eye contact depend on the nature of the relationship and reflect the power relationships, the attraction, or the antagonism between the participants [1]. Therefore, to understand and interpret the social interactions that are occurring, it is important to capture this signal accurately. The importance of detecting people Looking At Each Other (LAEO) has already been recognized in a series of computer vision approaches [197, 227] as well as in other papers that study human gaze [44, 246, 248, 25]. LAEO is complementary to other forms of human non-verbal communication

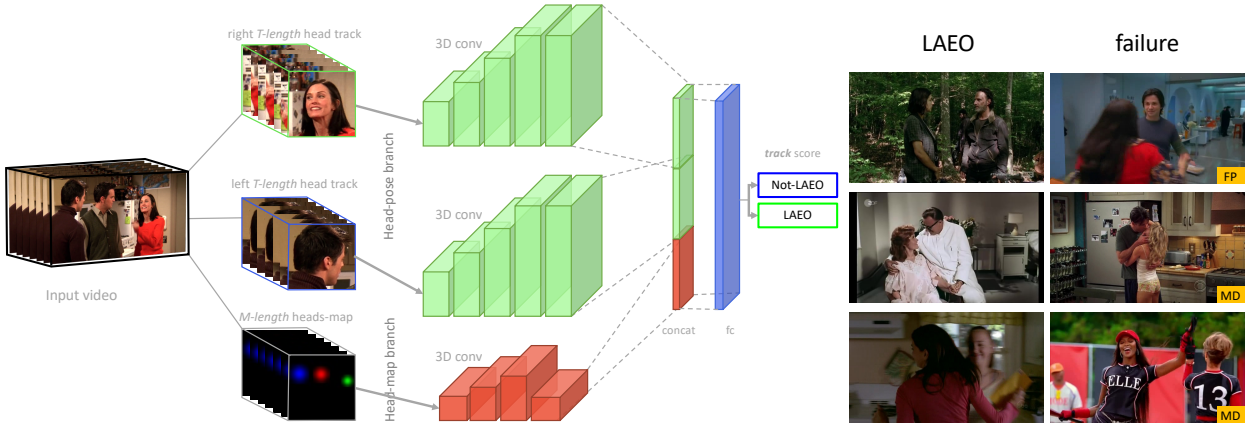


Figure 2.4: **(left) LAEO-Net++ and (right) Results**: (left) LAEO-Net++ consists of the head branches (green), the head-map branch (red) and a classification block, which scores the track sequence as LAEO or not-LAEO. (right) Results on UCOLAEO (first) and AVA-LAEO (second) TVHID (third row). (left column) correct LAEO results when the ground truth is LAEO. LAEO-Net++ successfully detects people LAEO in several situations (illuminations, scales, clutter). (right column) Failure cases for false positive LAEO detections (first example) and missed detections (two last examples). Most failures are missing people LAEO in ambiguous scenes; e.g. in the last red frame the characters are LAEO, even though the character on the left has closed eyes.

such as facial expressions, gestures, proxemics (distance), body language and pose, paralanguage (the tone of the voice, prosody), and interactions (e.g. hugging, handshake). Many of these have been the subject of papers, recent at the time of this work [196, 317, 97, 151].

Here, we introduce LAEO-Net++ [195] (and the first version LAEO-Net [194]), a new network for determining LAEO in videos. Unlike previous works at that time that only considered single frames, our approach answers whether two characters are LAEO over time by using a spatio-temporal model. The problem with frame-wise LAEO is that when characters blink or momentarily move their head, they are incorrectly considered non-LAEO. Our model considers head tracks over multiple frames and determines if two characters are LAEO for a time period based on the head pose and their relative position.

Related work. Gaze [246] and head pose [70] help determine the *visual focus of attention* (VFoA). A special VFoA case is *mutual gaze* or *looking at each other* (LAEO), where two subjects' VFoA are each other, often preceding or following physical interactions like a handshake. In *Behaviour Imaging*, detecting LAEO is key for understanding social interactions, as in autism [249]. [4] shows children with autism have increased eye contact with parents. [91, 182] note that willingness to LAEO demonstrates social interest.

The problem of detecting LAEO in videos was first introduced in [197], which used human head detection and tracking, modeling yaw and pitch angles with Gaussian Process regression. A LAEO score was calculated based on estimated angles and the relative head positions per frame and aggregated over a shot. Our LAEONet++ differs by estimating LAEO over a temporal window rather than a single frame. [253] detected conversational groups in social scenes by combining body orientation, head pose, and relative position, while [227] tackled the problem with two calibrated cameras in a two-person scenario. Recent integration of LAEO with 3D gaze estimation [68] has enhanced representation and highlighted the importance of 3D gaze in understanding relations, bridging 2D and 3D mutual gaze detection.

Looking at a person is a dominant category in human interactions in videos [231, 97]. [175] capture temporal cues, [185] classify relationships, and [151] jointly learn interactions and relations. Here, we use mutual gaze to identify interactions and determine friendship levels. A LAEO model can impact applications like detecting cartoons, animals (e.g., cats, chimpanzees [268]), or other objects (e.g., cars) looking at each other.

Approach overview. Given a video clip, we aim to determine if any two humans are *Looking At Each Other* (LAEO). To this end, we introduce three-branch LAEO-Net++ (Figure 2.4(left)). It takes as input two head tracks that determine the head poses and a head-map representing the relative position and scale between the two heads. Specifically, we depict as 2D Gaussians all the heads detected at each frame of the $\mathcal{T}$-frames track. The different Gaussian sizes encode the relative 3D arrangement (depth) of people in the scene, i.e. smaller sizes indicate that people are further from the camera compared to those with bigger size. This branch also encodes information for other people in the scene. Depending on its size and scale, a third person could cut the *gaze ray* between the two side people. Including this information helps LAEO-Net++ to distinguish such cases (red 'Chandler' Gaussian in Figure 2.4 bottom stream). Finally, LAEO-Net++ determines a confidence score for LAEO (a Softmax layer with Cross-Entropy) by also identifying the frames where it occurs. The network uses spatio-temporal 3D convolutions and is applied exhaustively over all pairs of simultaneous head tracks in the video.

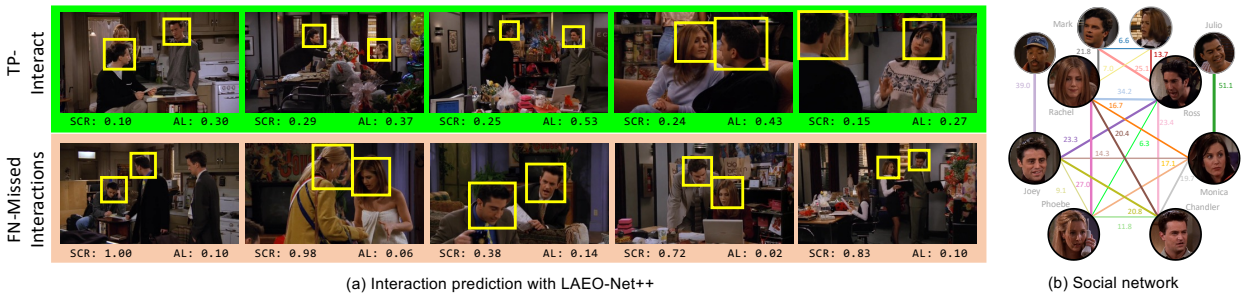
Proposed Datasets. In addition to the existing TVHID [231], i.e. the only video dataset with LAEO annotations, we propose two new datasets: UCOLAEO and AVA-LAEO[1]. **UCOLAEO** consists of 129 (3-12 seconds long) shots from four popular TV shows, with heads bounding box annotations in each frame and an additional label for every pair of heads in a frame as LAEO or not-LAEO. **AVA-LAEO** consists of movies from the training and validation sets of the AVA v.2.2 dataset [97]. We enhance the labels of the existing person bounding-boxes with LAEO annotations, resulting in $\sim$ 19k LAEO and $\sim$ 118k not-LAEO pairs for the training set and $\sim$ 5.8k LAEO and $\sim$ 28k not-LAEO pairs for the val set.

Experiments. **Metrics.** We use LAEO-classification Average Precision (AP).
Quantitative Results. We evaluate LAEO-Net++ on UCOLAEO, AVA-LAEO and TVHID and compare it against the state of the art at that time. When training and testing on UCOLAEO, and AVA-LAEO, the performance is 86.7% and 68.7%, respectively, revealing the significant performance gap between the two datasets, due to the different nature of AVA-LAEO: (1) no head annotations (just human bounding boxes); (2) it contains challenging visual concepts, such as low-resolution movies, crowded scenes, blurry, small heads, and particular clothing styles, e.g. people wearing turbans. Yet, LAEO-Net++ achieves AP=68.7%. LAEO-Net++ tested on TVHID achieves AP= 92.3% when training on UCOLAEO, outperforming training on AVA-LAEO that achieves AP=87.4%. This is because the domain of TVHID is closer to the one of UCOLAEO than to AVA-LAEO, given that UCOLAEO and TVHID consist of TV shows, whereas AVA-LAEO contains movies. Finally, the model trained on UCOLAEO outperforms all methods by $1-3\%$, reaching 92.3% vs 89.0% for Fine-head orientation [199] and 87.6%for Fully auto+HB [197].
Qualitative results. When applying LAEO-Net++ on UCO, AVA and TVHID we obtain the results of Figure 2.4 (right). Our model successfully detects people LAEO in several situations and scenarios, such as different illuminations, scales, cluttered background. By examining the remaining 8% error, we note that in most cases, the ground truth label is ambiguous, e.g. last frame in Figure 2.4 (right).

Social Network & Interaction prediction. One principal way of signalling interest in social interaction is the willingness of people to LAEO [91, 182]. The duration and frequency of eye contact reflect the power relationships, the attraction, or the antagonism between people [1]. We present here two applications of LAEO-Net++ in analysing social interactions in TV material.

I. Interaction-ness. At the shot level, we show that LAEO is an indicator of whether two characters are *interacting*. Here, we define two characters as interacting if they are directly involved (e.g. kiss, hug), or the actions of one influence the actions of the other (e.g. show something on a screen), or they communicate (e.g. talk to each other), or if they perform an activity together (e.g. shopping). Two characters are not interacting within a shot if they do not refer to each other (e.g. both characters listen

[1]Both datasets are available online `http://www.robots.ox.ac.uk/~vgg/research/laeonet/`.



(a) Interaction prediction with LAEO-Net++ (b) Social network

Figure 2.5: **(a) Interaction prediction with LAEO-Net++.** Examples of true positives (TP) and false negatives (FN) for Average-LAEO score (AL); note, in all examples the SCR results are reversed: i.e. the green rows are wrongly predicted as not-interacting; the orange row is correctly predicted as interacting. **(b) Social network using the *Average-LAEO* (AL) on Friends.** We depict the %AL between character pairs with the edges in the graph: the thicker the edge, the more dominant the relationship e.g. Ross and Rachel.

to a third person talking), or they do not influence each other, or they perform different tasks (e.g. one character is watching TV while the other is reading a book).

For every character pair, we measure the 'average-LAEO score' (AL) over the frames where the two characters co-exist and the Shots-Coexistence-Ratio (SCR) (i.e the ratio between the number of frames that two characters co-exist in a shot over the total number of frames of the shot). The examples in Figure 2.5(a) showcase the superiority of AL compared to SCR: the green are correctly predicted as interacting by AL, but wrongly predicted as not-interacting by SCR; the orange ones are missed interactions by AL, but correctly predicted as interacting by SCR. AL is suitable for predicting the presence or absence of interactions between characters, whereas the SCR is incapable of differentiating them (e.g. Monica and Joey in the last green example). Moreover, AL fails to determine interactions where people are not LAEO (e.g. Ross and Chandler or Mark and Rachel in orange). In most real-life cases, however, an interaction typically involves gazing; hence, AL is suitable for automatically capturing characters interacting.

II. Friend-ness. At the episode level, we show that LAEO is an indicator of the extent of social interactions between two characters, we term this *friend-ness*. We measure LAEO at the episode level as the average of AL over all shots in which the two characters appear and depict it in Figure 2.5(b): the thicker the edge, the higher the score and the stronger the relations. AL captures the dominant relationships, e.g. Ross and Rachel, against more distant one, e.g. Phoebe and Chandler. Our study reveals all prominent pair relations, demonstrating that the more people are LAEO, the stronger their *social relationship*.

Discussion. In this work, we focused on people *looking at each other (LAEO)* in videos. We proposed LAEO-Net++, which takes as input head tracks and determines if the people in the track are LAEO. This is the first work that uses *tracks* instead of bounding-boxes as input to reason about people on the whole track. We showed the generality of our model by applying it to a social case scenario, where we automatically infer the *social relationship* between two people based on the frequency they LAEO i.e. *friend*-ness, and showed that our metric can be useful for a guided search of interactions between characters in videos (i.e. interaction prediction).

Impact. Detecting mutual gaze and interactions in videos could help improve social interaction analysis and relationship recognition in various contexts. In social media and entertainment, for instance, it could enable more nuanced content recommendations and more interactive experiences by recognizing and understanding the nature of interpersonal connections. In mental health and therapy, it could aid in assessing social behaviors and interactions, offering insights into conditions like autism. However, this capability also raises privacy concerns, as accurate recognition of interactions and friendships might lead to invasive tracking or profiling if not managed with care.

2.2 Scene understanding

Here, we address scene understanding in cinematic content using multiple modalities in two directions: in Section 2.2.1, we present a method for understanding funny moments in movies, while in Section 2.2.2 we propose the Short Film Dataset with question-answering for story-level understanding.

2.2.1 Multimodal learning of funny moments in videos

We perceive the world through our senses, particularly in multimedia, where all signals can evoke emotions and reactions. Funniness is universal, yet while humans easily recognize humor across cultures and eras, machines struggle with it. Despite growing human-machine interactions, identifying funniness remains a challenge, hindering spontaneity. Humor is complex, often involving visual, auditory, or mixed cues, with no formula for the perfect joke. Recent studies exploring humor and funny moments [6, 331] mostly rely on text, with few incorporating videos [230, 143]. Their limitation is their dependence on external transcripts, which are unavailable in raw video data. Advances in speech-to-text now allow accurate transcript extraction from raw audio, aiding contextual understanding. Audio is vital for detecting humor, providing complementary cues like tone, pauses, and pitch [358, 29]. Visual cues, such as facial expressions and gestures, also impact how humor is perceived. Therefore, we exploit all these modalities and introduce FunnyNet-W, a multimodal model for predicting funny moments in videos.

Related work. *I. Sarcasm and Humor* share traits like irony but differ in representation. Sarcasm is often detected through dialogue analysis using language and acoustic patterns [56, 255, 307]; humor is usually identified by audio cues before laughter [29, 107]. Methods typically focus on text [6, 331] or multimodal cues [107, 106]. [143] uses vision and language attention, while [29] and [107] leverage LSTMs.
II. Sound Event Detection and Laughter Detection. Sound event detection involves identifying and timestamping sound events, often using Mel spectrograms [201, 321, 218, 219, 263]. We focus on laughter detection, using it as pseudo-labels to train FunnyNet-W. Laughter detection literature is limited, with some methods using physiological sensors [15, 281] and others employing supervised learning with annotated datasets [261, 87]. Our unsupervised method leverages multichannel audio data.
III. Multimodal tasks: *Audio+Video:* Methods use face movements to separate voices [81, 3] or align audio and video for speaker localization [272, 309]. *Video+Language:* Tasks include video captioning [322, 169], question answering [167, 350, 381], text-to-video generation [283]. *Audio+Language:* Tasks cover speech emotion recognition [355, 237], audio-text retrieval [183], audio captioning [147]. *Video+Language+Audio:* improves over unimodal approaches for video captioning [119] or emotion recognition [58].
IV. Modality Alignment. Recent efforts [240, 101, 88] focus on shared embeddings. CLIP [240] advanced text-image representation, with similar successes in audio-vision [210]. Attention connects multimodal signals [88, 346], and our method integrates cross-attention to capture token correlations [330, 297, 124, 154], using modality projections as bottlenecks [214] to prevent bias from a dominant modality.

Approach overview. FunnyNet-W (Figure 2.6) consists of: (i) the visual, audio, text encoders with videos, audio and subtitles (we use an automatic speech recognition system [241]) as inputs; (ii) the proposed Cross-Attention Fusion (CAF) module that explores cross and intra-modality correlations:

- **Cross-attention** models relationships among vision, audio, and text features. We stack all features as $F_S \in \mathbb{R}^{3\times512}$, and then feed F_S into three cross-attention modules to attend to $i=\{V,T,A\}$ for {vision, text, audio}. Next, the scaled attention per modality is computed as $\sigma\left(\frac{Q_S K_i^T}{\sqrt{d}}\right)V_i$, with σ the softmax. The query Q comes from the stacked features: $Q_S = F_S W^{Q_S}$, while the key K and value V come from a single modality as $K_i = F_i W^{K_i}$, and $V_i = F_i W^{V_i}$. Next, we obtain three cross-attentions and sum them to a unified feature F_U as: $F_U = \sum_{i\in\{V,F,A\}} \sigma\left(\frac{Q_S K_i^T}{\sqrt{d}}\right)V_i$.

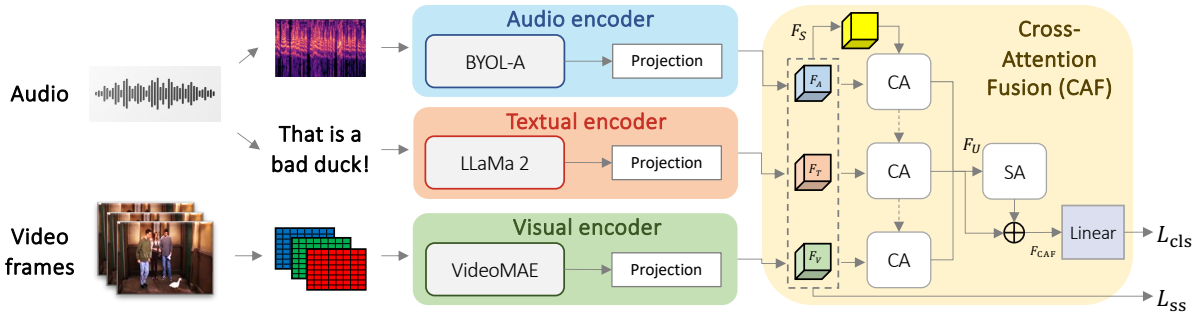


Figure 2.6: **Architecture of FunnyNet-W.** Given audio-visual clips, FunnyNet-W predicts funny moments in videos. The outputs of the audio (*blue*), textual (*red*), and visual (*green*) encoders pass through the Cross Attention Fusion (*CAF*), which consists of cross-attention (*CA*) and self-attention (*SA*) for feature fusion. It is trained to embed all modalities in the same space via self-supervision (L_{ss}) and to classify clips as funny or not-funny (L_{cls}).

- **Self-attention** computes the intra-correlation of the F_U features, which are further summed with a residual F_U as: $F_{CAF} = F_U + \sigma\left(\frac{Q_U K_U^T}{\sqrt{d}}\right) V_U$, where $Q_U = F_U W^{Q_U}$, $K_U = F_U W^{K_U}$, $V_U = F_U W^{V_U}$.

Finally, we average F_{CAF} tokens and feed it to a classification layer. Note, CAF differs to existing methods [207, 330] in the cross attention computation. Using stacked features F_S to attend to each modality Q_S brings three benefits: (a) it is order-agnostic: for any modality pair we compute cross-attention once, instead of twice by interchanging queries and keys/values; resulting in reduced compute; (b) each modality serves as a query to search for tokens in other modalities; bringing rich feature fusion; and (c) it generalizes to any number of modalities, resulting in scalability.

Training. FunnyNet-W is trained with $L = \lambda_{ss} L_{ss} + \lambda_{cls} L_{cls}$, where λ_{ss}, λ_{cls} the weighting parameters that control the importance of each loss. The two individual losses are: First, Softmax loss L_{cls} to predict if the input is funny or not. Second, to capture 'mutual' audiovisual information, we solve a self-supervised synchronization task [46, 149, 225]: we encourage visual features to be correlated with true audios and uncorrelated with audios from other videos. Given the i-th pair of visual v^i and true audio features a^i and N other audios from the same batch: $a_1, ..., a_N$ we minimize the loss [40, 47, 220]: $L_{cotrs} = -\log \frac{\exp(S(v^i, a^i)/\tau)}{\sum_{j=1}^{N} \exp(S(v^i, a^j)/\tau)}$, where S the cosine similarity and τ the temperature factor. Here, we compute the contrastive loss between all three modalities, i.e., visual-audio, text-audio, and visual-text. Thus, our self-supervised loss is: $L_{ss} = -\frac{1}{3}\left(L_{cotrs}^{v^i,a^i} + L_{cotrs}^{v^i,t^i} + L_{cotrs}^{t^i,a^i}\right)$.

Positive and negative samples. For training, we exploit the laughter that naturally exists in TV Shows: we define as 'funny' for any audiovisual snippet followed by laughter; and 'not-funny' any audiovisual snippet not followed by laughter. To detect funny moments, we also propose an unsupervised laughter detector that separates voices from background audio, described in [178, 179].

Proposed Datasets. The datasets for funny moment detection are: The Big Bang Theory [143], Multimodal Humor Dataset [230] MUStARD [29] and UR-Funny [107]. Here, we also enrich the Friends [26, 137] dataset by manually annotating 3.5k laughter time codes at the start and the end of all laughters.

Experiments. **Metrics.** To evaluate FunnyNet-W, we use classification accuracy (Acc) and F1 score (F1).
Quantitative results. We compare FunnyNet-W [179] to the state of the art at that time: MUStARD [29], MSAM [230], MISA [108], HKT [106] and LaughM [143]. Our results show that first, FunnyNet-W outperforms all methods in both metrics by approximately 1-3% Acc, confirming its effectiveness, and, second, the performance in the out-of-domain UR-Funny is significantly high (80.3% Acc vs 77.4% for HKT). These highlight that FunnyNet-W is an effective model for funny moment detection.
Qualitative results. To visualize the impact of modalities, we compute the average attention values on the three CA modules (CA boxes in Figure 2.6) and then, show the average weights for each modality

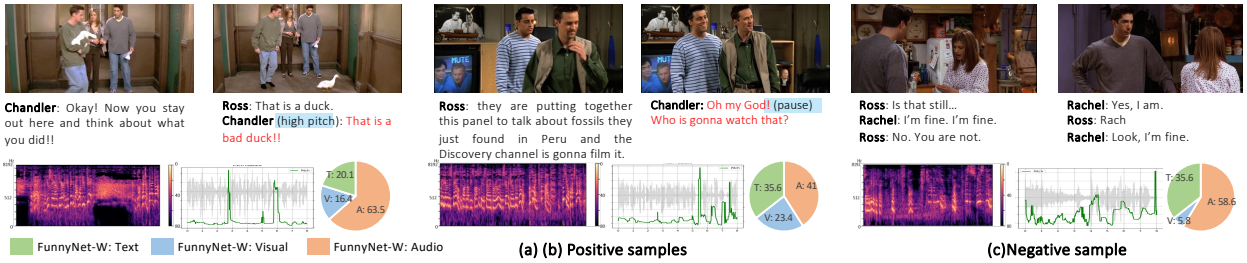


Figure 2.7: **Examples of (*a*,*b*) funny, and (*c*) non-funny predictions** on the Friends test set. We show the audio, visual and text inputs, the learned average weights of cross-attentions from CAF (*pie chart*), and the subtitles.

in the pie chart of each example in Figure 2.7. For this, we show (a-b) two positive and one (c) negative samples on Friends with frames, subtitles and audio spectrogram (left) and pitch (right). The contribution of each modality varies; the commonality though is that audio contributes more than half, followed by text and visual features. When there is a strong audio signal, the contribution of audio increases significantly. This is shown when the character yells ('Chandler' in (a)), or pauses the speech ('Chandler' in (b)). In contrast, in (c), the tone, volume, pitch or rhythm do not change greatly, so the text starts to play a bigger role in determining them as non-funny scenes. Furthermore, in (c), the visual feature plays very little role in the final prediction probably because the scenes do not capture the whole character's bodies and their movement, so the visual model can offer only little information.

FunnyNet-W against LLM chatbot. Here, we compare FunnyNet-W [179] against a Large Language Model (LLM) [223, 313] chatbot to assess its performance relative to these general models. Specifically, we evaluate the LlaMa-2 [313] chatbot on the Friends dataset with prompt training (few-shot setting: we give some examples to the LLMs) and without prompt training (zero-shot setting: we prompt the LLM with the question and subtitles). Our results show that with prompt training, the chatbot's performance increases from 53.2% to 55.9% in Acc. Overall, FunnyNet-W outperforms all examined cases with the chatbot, reaching 85.2% in Acc. Interestingly, we note that the performance of FunnyNet-W using text only is 68.1% in Acc, close to the one of the LlaMa-2 chatbot (55.9%), thus showcasing the impressive representation power of LLM chatbots, who perhaps have already some knowledge of popular edited films, such as Friends TV-show (see also the following Section 2.2.2).

Discussion. We introduced FunnyNet-W [179] and FunnyNet [178], an audiovisual model for funny moment detection. In contrast to works that rely mainly on text, FunnyNet-W also exploits audio that comes naturally with videos and contains high-level cues (pauses, tones). Our findings show audio is the dominant cue for signaling funny situations, while audio and text offer complementary information. Our results show the effectiveness of FunnyNet-W, which sets the new state of the art on five datasets. Future work includes analyzing the contribution of audio cues (pitch, tone) and different languages.

Impact. FunnyNet-W has several applications. It can help cognitive researchers study humor at scale, assist artists in editing films without needing a live audience, and enhance human-machine interactions by adding humor to conversational agents. However, its deployment requires caution due to its potential misuse, such as aiding identity fraud by mimicking a victim's sense of humor. FunnyNet-W, trained mainly on Western, especially U.S. materials, may struggle to generalize across cultures due to thematic and expressive differences. Additionally, it faces language bias, as it is trained in English, limiting its effectiveness across different languages and cultural contexts. *Environmental.* All experiments are done on NVIDIA RTX4090 and A100 GPUs, with each of them requiring 215W in power supply. For this project, we use approximately 800 GPU hours. Training a FunnyNet-W model with all three modalities requires around 6 GPU hours on NVIDIA RTX4090, which amounts to 1.29 kWh and 300.75g of CO_2 emitted.

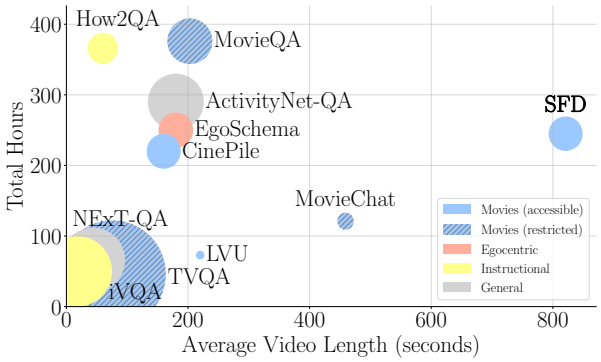


Figure 2.8: **Comparison of SFD to other VQA datasets.** The circle size indicates the number of QA pairs in each dataset.

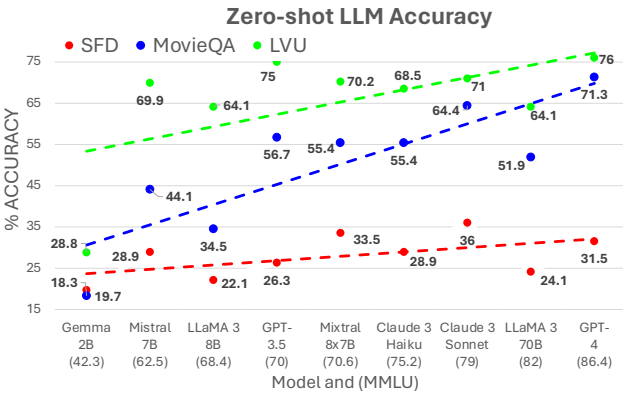


Figure 2.9: **Data leakage.** When given *only* the movie title, higher zero-shot accuracy in question-answering by LLMs indicates greater data leakage. LLMs are ranked by MMLU.

2.2.2 Movie Question Answering

Recent advances in vision-language models [168, 162, 38, 39, 327, 318, 80, 79, 52, 344] have shown promise in enhancing machine perception. However, existing datasets face limitations. Most consist of short videos [97, 2, 62, 345, 348, 94, 290, 142, 163, 96, 337], focusing on short-term tasks like action recognition and video retrieval, requiring only a few seconds of content [192, 156]. For long-form video understanding, three main categories exist: (i) *Egocentric videos* capture continuous actions and excel in video duration but lack narrative depth [95, 54, 127, 170, 192, 282, 273]; (ii) *Instructional videos*[202, 347] focus on procedural aspects but are brief and lack storytelling; (iii) *Movies* provide complex narratives and extended duration, making them ideal for testing long-form video comprehension [256, 11, 304, 157, 285, 103, 334, 117, 316, 342, 262, 198, 244, 35, 23, 190].

Despite their benefits, movie datasets have three major limitations: (i) *Accessibility*: they often include copyrighted content, limiting public access; (ii) *Clip duration*: they consist of short, incomplete clips [304, 157, 245, 334], hindering long-form understanding and narrative focus [286]; (iii) *Data Leakage*: as these datasets comprise well-known commercial movies, modern Large Language Models (LLMs) and Vision-Language Models (VLMs) have likely been exposed to some form of movie information (i.e., synopses, reviews, discussions, subtitles, blog posts).

To overcome these issues, we introduce the Short Film Dataset (SFD), featuring 1,078 short films totaling over 243 hours, with an average duration of 13 minutes per film –longer than existing datasets. SFD includes publicly accessible amateur films from YouTube, offering complex narratives and minimal exposure to LLMs, reducing data leakage risks (see Figure 2.9).

Related work. *I. VideoQA benchmarks* evaluate reasoning, memory, and comprehension through various datasets covering visual descriptions [345, 37], temporal reasoning [337], compositional reasoning [96], social intelligence [357], instructional [347], egocentric [192, 127, 170], and movie videos [304, 157, 190]. These datasets typically use short videos and require reasoning on only a few frames [192].
II. VideoQA methods use large-scale pre-trained models through multimodal contrastive learning [79, 162, 327, 360, 359, 39, 38] or visually-conditioned large language models [160, 52]. For example, [344] enhances LLMs with visual capabilities, [349] excels in zero-shot VideoQA using lightweight adapters, while Video-LLaVA [168] connects a visual encoder and language model via a simple projection layer.
III. Long-form video understanding tests models on long-term reasoning [250]. Datasets like those for egocentric videos [192] and movies [334, 285, 103] focus on complex tasks, although some, like LVU [334], have limited tasks. Datasets like [304] and [286] focus on movie plots and characters but lack video data due to copyright concerns.

Dataset	Venue	Annotation	Avg. Length (s)	#QA Pairs	Multimodal	Long-Term	Accessible	Unknown To LLMs	Full Movies
General VideoQA datasets									
MSRVTT-QA [343] (test)	ACM 2017	Auto	15	72,820	✗	✗	✓	✓	-
MSVD-QA [37] (test)	ACM 2017	Auto	10	13,156	✗	✗	✓	✓	-
TGIF-QA [125] (test)	CVPR 2017	Auto	3	25,751	✗	✗	✓	✓	-
ActivityNet-QA [356] (test)	AAAI 2019	Manual	180	8,000	✗	✗	✓	✓	-
How2QA [265] (test)	EMNLP 2020	Manual	60	4,400	✗	✗	✓	✓	-
NeXT-QA [337] (test)	CVPR 2021	Manual	44	9,178	✗	✗	✓	✓	-
iVQA [173]	ICCV 2021	Manual	18	10,000	✗	✗	✓	✓	-
EgoSchema [192]	NeurIPS 2023	Manual + Auto	180	5,000	✗	✓	✓	✓	-
MovieQA datasets									
MovieQA [304] (test)	CVPR 2016	Manual	203	6,462	✓	✓	✗	✗	✗
TVQA [157] (test)	EMNLP 2018	Manual	76	15,253	✓	✗	✗	✓	✗
LVU [334] (test)	CVPR 2021	Manual	220	1,223	✓	✓	✓	✗	✗
MovieChat [286] (test)	CVPR 2024	Manual	459	2,417	✓	✓	✗	✓	✗
CinePile [245] (test)	arXiv 2024	Manual + Auto	160	4,940	✓	✓	✓	✓	✗
SFD (Ours)		Manual + Auto	821	4,885	✓	✓	✓	✓	✓

Table 2.1: **Comparison of VideoQA and MovieQA datasets.**

Proposed Dataset. Short films, typically 5-20 minutes long, are motion pictures that span various styles (e.g., narrative fiction, documentary, animation) and genres (e.g., action, drama, comedy, horror) and are often experimental rather than commercial [86]. The recent rise of publicly available short films allows us to create the Short Film Dataset (SFD) to advance research in story-level video understanding.

We create SFD from the Omeleto YouTube channel[2], which features high-quality, award-winning short films. We downloaded videos, subtitles, and metadata, including titles, loglines, synopses, and details like genre, release year, region, and language and filled in missing transcripts.

SFD supports two question-answering tasks: (a) Multiple-Choice Question answering (MCQ) [304, 245] and (b) Open-Ended Question answering (OEQ). All questions and answers, generated by LLMs from movie descriptions, went through manual curation to ensure they accurately reflect the films' settings, characters, storylines, and themes. SFD includes 4,885 MCQs and OEQs, averaging 4.53 questions per film. Compared to existing video and movie question-answering datasets, SFD has the longest average video duration (821 seconds) and offers benefits in multimodality, public accessibility, data leakage prevention, and narrative coherence.

Experiments. **Data leakage.** Modern LLMs, pre-trained on vast internet data, risk being exposed during training to information about commercial films, such as synopses, reviews, scripts, and transcripts. This issue may result in models recalling answers directly without analyzing the video content, leading to biased benchmarks. In this section, we quantitatively assess the extent of data leakage on datasets. Specifically, we prompt LLMs to answer open-ended questions using *only* the movie title and compute the accuracy of responses, following [189]. We experiment with 3 movie datasets: MovieQA [304], LVU [334] and our SFD, 5 open-sourced models: Gemma 2B [306], Mistral 7B/8x7B [129], LLaMA-3 8B/70B [312] and 4 commercial models Claude 3 Haiku/Sonnet, and GPT-3.5/4 [221]. Figure 2.9 reports the results of the data leakage experiment and further details can be found in [86].

We observe that both MovieQA and LVU suffer from data leakage, reaching up to 71.3% and 76.0% accuracies. In contrast, thanks to the low presence of amateur film on the internet, SFD exhibits a maximum accuracy of 36.0%, indicating a low leakage issue. Furthermore, as expected, the extent of data leakage correlates with the knowledge level reflected by MMLU score [77], indicating that LLMs with more knowledge exhibit higher levels of memorization, leading to worse leakage issues. For instance, for both MovieQA and LVU, the zero-shot accuracy increases from approximately random level (20%) to more than 70% as the model size rises. Meanwhile, SFD maintains stable and low accuracies ranging between 19.7% and 36.0%, regardless of LLM knowledge variation, further indicating its low data leakage.

[2] https://www.youtube.com/@Omeleto

This experiment reveals that relying solely on existing datasets to evaluate new methods is insufficient. Instead, our SFD offers a more objective and reliable test bed for long-term video understanding.

Quantitative results. We benchmark 7 state-of-the-art methods on SFD: FrozenBiLM [13], mPLUG-Owl2 [351], Video-LLaVA [168], LLoVi [363], LangRepo [136], MovieChat [286] and TimeChat [252]. These models are tested in a zero-shot video question-answering setting.
User Study. To verify the answerability of our MCQ and assess the upper limit of SFD, we conducted three user studies: (1) *(Vision-Language)*,—full video with audio and subtitles; (2) *(Vision-only)*-muted videos; (3) *(Language-only)*-plain text subtitles. For each question, all participants were asked to select the correct answer. Our results show that when provided with the full multimodal information, participants answer questions with high accuracy (89.8%). As expected, removing modalities lowers accuracy: when using only subtitles the performance is 70.9%, whereas the vision-only performance drops to 59.0%.
Question Answering. For MCQ, model performance is generally poor (max 55.6% for LLoVi), with LLoVi being a notable exception in the language-only setting (64.2%). Compared to human performance, there is a significant gap in the multi-modal setting, where human performance reaches 89.8% and the best model reaches 55.6%. This can be explained by the vision-only setting, where the gap reaches 20.7%. The OEQ task proves even more challenging, with lower accuracy across all models, ranging from 3.5% to 40.3%. Subtitles-only models, particularly LLoVi, lead in performance, highlighting the dominance of language processing in understanding and answering complex questions.

Our results and analysis suggest that (1) text (subtitles) is a stronger cue than visual frames for movie question-answering; (2) modern multimodal methods (max Acc at 60.0%) fall behind human evaluation; and (3) there is a large room for improvement in the visual aspect, where the gap in performance between modern methods (average Acc 38.3%) and user study (59.0% Acc) is still very high.

Discussion. We introduced the Short Film Dataset (SFD), a long-form video understanding benchmark featuring complete and narrative-driven amateur short movies. SFD includes multiple-choice and open-ended question-answering tasks. Compared to other datasets, SFD stands out by offering richer, story-oriented content with tasks specifically tailored for long-term reasoning. Unlike most movie datasets that use copyrighted commercial films prone to data leakage with LLMs, SFD relies on amateur films, which are publicly accessible and have a limited online presence. Furthermore, our analysis indicates that while language-based methods achieve performance levels comparable to humans, state-of-the-art vision-based and multimodal methods fall behind human evaluation and they still have considerable room for improvement, highlighting the need for further advancements. In conclusion, we believe that our SFD paves the way for accessible, and comprehensive long-term movie understanding that is not known a priori by LLMs, thus helping the community develop robust methods.

Impact. The proposed SFD has several implications regarding copyright, licensing, and potential social impacts. We do not distribute raw video content; instead, we provide URLs that link directly to YouTube, ensuring that creators' copyrights are respected in accordance with YouTube's Terms of Service.[3] The dataset is intended solely for academic research and is licensed under CC BY-NC-SA 4.0 (Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International) to protect the metadata and associated information. However, there are potential negative social impacts. The dataset primarily includes English-speaking films from North America and Europe, which could lead to biased video understanding models due to the under-representation of non-English-speaking cultures. Additionally, advances in movie understanding may encourage filmmakers to tailor their work to optimize for what is favoured by algorithms, potentially stifling creative freedom and depriving innovation.

[3] `https://www.youtube.com/static?template=terms` (Section: License to Other Users)

Chapter 3

Multimodal Visual Content Generation

This Chapter focuses on multimodal visual content generation, where the goal is to leverage multiple modalities to enhance the generation of complex visual scenes.

Conditional image generation (Section 3.1). First, Section 3.1.1 presents our semantically conditioned image generation work. Most such methods focus on the narrower task of pose transfer and ignore the more challenging task of subject transfer which consists in not only transferring the pose but also the appearance and background. Instead, here, we introduce the GAN-based SCAM that encodes rich and diverse information in each semantic image region, achieving precise generation with emphasis on fine details. SCAM successfully encodes the diversity of appearance in each semantic region and sets the new state of the art on subject transfer. The code is available online. This work has been published at ECCV 2022 [73].

Second, Section 3.1.2 presents our Coherence-Aware Diffusion (CAD) method that integrates coherence in conditional information into diffusion models, allowing them to learn from noisy annotations without discarding data. This is useful in scenarios where conditional information may be noisy or unreliable due to human annotation errors or weak alignment. We assume that each data point has an associated coherence score that reflects the quality of the conditional information. We then condition the diffusion model on both the conditional information and the coherence score. In this way, the model learns to ignore or discount the conditioning when the coherence is low. We show that CAD is theoretically sound and empirically effective on various conditional generation tasks. The code is available online online. This work has been published at CVPR 2024 [72].

Third, Section 3.1.3 presents a comprehensive analysis and insights into Classifier-Free Guidance (CFG) weight schedulers, the default way for image generation. CFG operates by combining the conditional and unconditional predictions using a fixed weight. However, varying the weights throughout the diffusion process leads to superior results. In this work, we analyze this behaviour and provide valuable insights into CFG weight schedulers. Our findings suggest that (a) simple, monotonically increasing weight schedulers consistently lead to improved performances, requiring merely a single line of code and (b) more complex parametrized schedulers can improve results, but do not generalize across models and tasks. The one-line code is published at TMLR 2024 [326].

Multimodal motion generation (Section 3.2). Stories and emotions in movies emerge through well-thought-out directing decisions, in particular camera placement and movement. Crafting compelling camera trajectories remains a complex iterative process. To tackle this, here, we present our proposed dataset called the Exceptional Trajectories (E.T.) with camera trajectories along with character information and textual captions encompassing descriptions of both camera and character. To our knowledge, this is the first dataset of its kind. We also propose the diffusion-based Director, which generates complex camera trajectories from textual captions that describe the relation and synchronisation between the camera and characters. Our model outperforms the state of the art and our experiments reveal its effectiveness. The code and dataset are available online. This work has been published at ECCV 2024 [50].

3.1 Conditional image generation

Here, we address image generation conditioned on various input modalities, i.e. semantic information (Section 3.1.1) and class or text (Sections 3.1.2-3.1.3).

3.1.1 Semantically conditioned image generation

Subject transfer between images is crucial for applications like film and art industries. For example, in filmmaking, a stunt performer could be seamlessly replaced by the main actor, eliminating the need for a look-alike and offering more freedom. The goal of subject transfer is for the source subject to replace the target subject in the target image while preserving the background, object interactions, and spatial configuration. Unlike faces or landscapes, human bodies are highly diverse and difficult to model.

Most methods address either pose transfer [365, 30, 323] or style transfer [380, 228], but they have limitations: (1) they only work on simple backgrounds (PISE [365], SEAN [380],[323]), and (2) they are expensive as they require intensive training [323] or one model per subject (Everybody Dance Now [30]). Subject transfer, however, involves changing both pose and style/identity simultaneously.

The related semantic editing task controls a network's output by using a segmentation mask. It can be adapted for subject transfer by applying the target's mask with the source's style. However, modern methods struggle with complex, real-world scenes. For instance, SPADE [228] fails at independent region style control, while SEAN [380] cannot handle detailed scenes with multiple background objects.

To address these, we propose **SCAM** (Semantic Cross Attention Modulation), a semantic editing model for subject transfer. SCAM captures fine details within semantic regions by using multiple latents per semantic region, enabling better handling of coarse labels like backgrounds. It generates more complex backgrounds and outperforms SEAN [380] in both subject transfer and semantic reconstruction.

Related work. *I. Image to Image Synthesis with GANs.* StyleGAN [140, 141] modulates the feature map at each resolution using a style vector, improving GANs. Pix2Pix [121, 324] introduces control through paired data, but this is challenging to obtain. CycleGAN [379] circumvents the need for paired data by using cycle consistency with unpaired data. In our case, paired data are unavailable, as creating ground-truth images with identical pose and occlusion is infeasible. We address this by training on a reconstruction proxy task and performing subject transfer at test time.
II. Semantic Image Generation. Pix2Pix [121, 324] uses segmentation masks but loses semantic detail. SPADE [228] improves this with layer-wise semantic conditioning. CLADE [299] reduces SPADE's complexity. Other methods enhance SPADE [328, 174, 165, 269, 301, 299, 74, 98], but do not focus on image re-generation for editing. SEAN [380] uses style vectors per semantic label, while GroupDNet [383] and INADE [298] encode each semantic region separately, but struggle with coarse labels and single vectors per region. Our SCAM approach introduces the SAT-Encoder for rich representations and multiple latents per region, with pixel-wise modulation enabling unsupervised semantic structure. Diffusion methods like [200] are also used for editing but are computationally expensive.
III. Attention. Transformers, despite their success [59, 242, 243, 27], have quadratic complexity. Vision methods [69, 310, 177] mitigate this by using image patches, which lose information. Perceiver [124, 123] addresses this with cross attention, using fewer learned tokens. Attention has also advanced in GANs [364, 75, 134, 372, 155, 362]. GANsformer [118] uses cross attention for style codes but focuses on unconditional generation, not subject transfer. Our SCA module improves GANsformer's attention for semantically constrained generation.
IV. Pose Transfer. Keypoints for pose transfer [186, 382, 300, 161] result in coarse body representations. Semantic masks [104, 67, 366, 365] improve this but focus on pose transfer, not subject transfer, and often do not preserve backgrounds. [30] overfit GANs to videos, unsuitable for dynamic scenes, while [323] tie subjects to backgrounds. Our focus is on subject transfer, changing both pose and background.

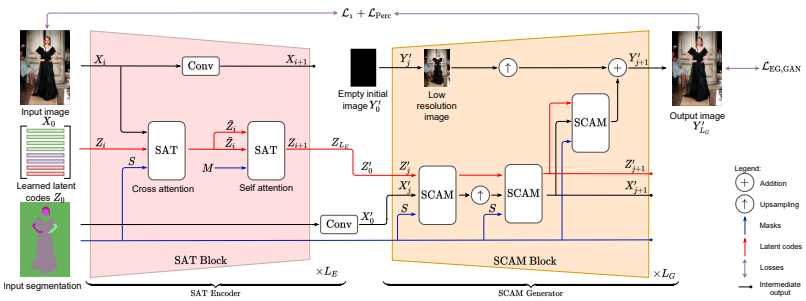


Figure 3.1: **Training setup of SCAM**. It consists of the SAT-Encoder (pink) and the SCAM-Generator (yellow). The SAT-Encoder allows the latents to retrieve information from an image, exploiting both the raw image and the convolution feature maps. Once the image is encoded, the latents are fed to the SCAM-Generator, which captures top-down and bottom-up interactions with a semantic constraint, allowing to easily alter the desired regions thanks to the latents that are dedicated to a given region.

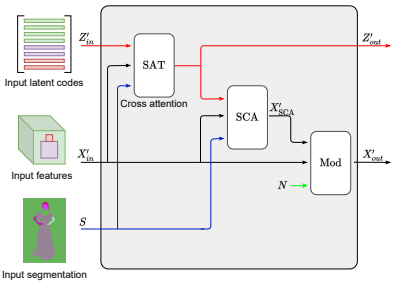


Figure 3.2: **SCAM-Operation**. It modulates a feature map according to a segmentation map, allowing each pixel to retrieve information from a semantically restricted set of latents. It enables both top-bottom (latents retrieve information from the feature map) and bottom-top interactions (the map gets information from latents).

Approach overview. Our goal is to perform semantic editing with a focus on subject transfer. We propose the **SCAM** method (Semantic Cross Attention Modulation, Figure 3.1). It relies on three items.

First, **SCA** (Semantic Cross Attention), i.e. a novel mechanism that masks the attention according to segmentation masks, thus encoding semantically meaningful latent variables. The goal of SCA is two-fold depending on what is the query and what is the key. Either it allows to give the feature map information from a semantically restricted set of latents or, respectively, it allows a set of latents to retrieve information in a semantically restricted region of the feature map. It is defined as:

$$\text{SCA}(I_1, I_2, I_3) = \sigma\left(\frac{QK^T \odot I_3 + \tau\left(1 - I_3\right)}{\sqrt{d_{in}}}\right)V \quad , \tag{3.1}$$

where I_1, I_2, I_3 the inputs, with I_1 attending I_2, and I_3 the mask that forces tokens from I_1 to attend only specific tokens from I_2[1], $Q{=}W_Q I_1$, $K{=}W_K I_2$ and $V{=}W_V I_2$ the queries, keys and values, and d_{in} the internal attention dimension.

We use three types of SCA. *(a) SCA with pixels X attending latents Z*: $\text{SCA}(X, Z, S)$, where $W_Q{\in}\mathbb{R}^{n\times d_{in}}$ and $W_K, W_V{\in}\mathbb{R}^{m\times d_{in}}$. The idea is to force the pixels from a semantic region to attend latents that are associated with the same label. *(b) SCA with latents Z attending pixels X*: $\text{SCA}(Z, X, S)$, where $W_Q{\in}\mathbb{R}^{m\times d_{in}}$, $W_K, W_V{\in}\mathbb{R}^{n\times d_{in}}$. The idea is to semantically mask attention values to enforce latents to attend semantically corresponding pixels. *(c) SCA with latents Z attending themselves*: $\text{SCA}(Z, Z, M)$, where $W_Q, W_K, W_V{\in}\mathbb{R}^{n\times d_{in}}$. We denote $M \in \mathbb{N}^{m\times m}$ this mask, with $M_{\text{latents}}(i, j){=}1$ if the semantic label of latent i is the same as the one of latent j; 0 otherwise. The idea is to let the latents only attend latents that share the same semantic label.

Second, SCAM consists of **SAT-Encoder** (Semantic Attention Transformer) that relies on cross attention to decide which information to gather in the image and for which latent. Third, SCAM consists of the **SCAM-Generator** (Semantic Cross Attention Modulation) that captures rich semantic information in an unsupervised way. It consists of a series of SCAM-Operation blocks, which aim at exchanging information between pixels/features and latents of the same semantic label (see Figure 3.2).

Experiments. **Metrics.** We use PSNR, and swap FID (S-FID). S-FID is computed as the FID between the test set and a set of subject transfer images computed on the test set.
Datasets. We use iDesigner [254], CelebAMask-HQ [153], and ADE20K [377].

[1]The attention values requiring masking are filled with $-\infty$ before the softmax. (In practice $\tau{=}-10^9$)

Figure 3.3: **Subject Transfer results** on the test set of (top) iDesigner [254] and (bottom) CelebAMask-HQ[153]. Note the hard case in the first row, where only SCAM rotates the subject.

Quantitative results. At the time of publication, SCAM outperformed all previous state-of-the-art approaches for all metrics. These major boosts in PSNR (approx. 20 vs 10-14 for other methods on all datasets) showed that our reconstructed images better preserve the details of the initial images, meaning the representation power of SCAM is higher than that of other approaches. The difference is also notable for S-FID (19.8 vs 22.8 for SEAN [380]), showing that our method performs better subject transfer on datasets with coarse semantic labels than other approaches.
Qualitative results. Figure 3.3 shows the subject, background images, and segmentation mask of the pose, followed by subject transfer results from various approaches. The top row shows samples from iDesigner, and the second row is from CelebAMask-HQ. Among all methods, SCAM best preserves all components of subject transfer: subject appearance, background, and pose. In the challenging first row, where the subject has a different pose from the reference, SCAM successfully rotates the subject, unlike other methods. In the bottom row, SCAM recovers more details in the transferred image, such as skin color and facial expression. Notably, it captures the bicolor hair, whereas SEAN, SEAN-CLADE, and INADE show averaged hair color. Most approaches struggle with person generation; only SCAM produces a coherent human with the background reference.

Discussion. We introduced SCAM that performs semantic editing and in particular subject transfer in images. The architecture contributions of SCAM are: first, the semantic cross attention (SCA) mechanism, performing attention between features and a set of latents under the constraint that they only attend to semantically meaningful regions; second, the Semantic Attention Transformer Encoder (STA) retrieving information based on a semantic attention mask; third, the Semantic Cross Attention Modulation Generator (SCAM) performing semantic-based generation. SCAM set the new state of the art at that time by leveraging multiple latents per semantic region and by providing a finer encoding of the latent vectors both at encoding and decoding stages.

Impact. This work has several applications, such as image editing where one person can be swapped for another. This can have a negative impact if used with bad intentions, for instance, it could be misused to create fake news. Even though today this is not vital as detection of deep fakes still remains feasible, it will become more challenging as research advances; hence, regulation will be necessary to manage these risks. The method could also benefit film production by allowing actors to be replaced, which is useful if an actor is unavailable or to reduce costs by using superstars' likenesses without their physical presence. However, this may raise legal issues if people's images are used without permission.
Environmental impact. We used 42.3 thousands GPUs hours on Nvidia V100-32g. We used the French Jean Zay cluster, with 50-80g CO_2 for each kWh produced. Our experiments used 80% of the GPUs maximum power of 250Wh, which amounts to 10.6 MWh of energy used for the whole project. Considering only the CO_2 for the production of the electricity used, this results in 528-846kg of CO_2 emitted for this project. Training a single SCAM model requires 50 GPUs hours, which amounts to 10kWh and 500-800 g of CO_2 emitted. As a comparison, the world average per capita CO_2 emission is 4.7 ton/year.

3.1.2 Coherence-Aware Diffusion

Conditional Diffusion models excel in image generation while affording greater user control over the generation process by integrating additional information [264, 14]. This extra data enables the model to guide the generated image towards a specific target, leading to improved various applications including high-quality text-to-image generation [257], as well as other modalities such as depth or human body pose [367]. Furthermore, the accessibility of open-source models like Stable Diffusion has democratized diffusion, already causing significant shifts in various domains such as design, art, and marketing.

Training conditional diffusion models requires substantial volumes of paired data comprising the target image and its corresponding condition. In text-to-image generation, this pairing involves an image and a descriptive caption that characterizes both the content and the style of the image. Similarly, for class conditional generation, the pair consists of an image and its corresponding class label. Besides the technical challenges associated with acquiring extremely large quantities of paired data, ensuring accurate alignment between image and text conditions is still an open research question, as attested by the large amount of work in this [160, 361]. In practice, large web-scraped datasets, such as LAION-5B [270] or CC12M [32], contain abundant noisy pairs due to their collecting process. To clean the pairs, hence ensuring alignment of higher quality, the prevailing strategy filters out samples that fail to meet an arbitrarily chosen criterion, often done through techniques like thresholding the CLIP-score [240, 234, 236]. This approach, however, has two main drawbacks: first, it is challenging to adjust the criterion accurately and more importantly, it discards many high-quality samples that could potentially enhance generation quality irrespective of the condition. For instance, out of the 50B initially-collected text-image pairs, only 10% were left in LAION-5B [270], thus discarding 90% of the samples, i.e. 45B images.

Instead of discarding the vast majority of training samples, in this work, we leverage them to learn simultaneously conditional and unconditional distributions. Specifically, for each sample, we introduce the *coherence score*, which measures how coherent the conditioning is with respect to the data, i.e. how well the condition corresponds to its associated image. We incorporate this *coherence score* into the training process by embedding it into a latent vector, which is subsequently merged with the condition. We then condition the diffusion model on this coherence score in addition to the original condition. This additional information enables the model to determine the extent to which the condition should influence the generation of a target image, as shown in Figure 3.4. By doing so, the model learns to discard the low-coherence conditions and focus on the high-coherence ones. Consequently, our model can behave as either a conditional or an unconditional model. Low-coherence samples, lead to unconditional sampling, while high-coherence samples lead to conditional samples. During inference, our method has the flexibility to take as input the coherence score, thereby allowing users to vary the impact of the condition on the generation process. Building on this, we redesign the Classifier-Free-Guidance (CFG) [112] to rely on coherence conditioning instead of dropping out the conditioning randomly.

Related work. *I. Conditional generation.* Previous attempts to condition generative models were focused on GANs [93]. Recently, diffusion models have made significant advances in image generation [284, 288, 289, 111]. Compared to GANs, they have better coverage over the data distribution, are easier to train, and outperform them in terms of image quality [61]. Architecture-wise, diffusion models rely mostly on modified versions of a U-Net [111, 288, 61]. Recent works have shown that other architectures are possible [232, 113]. ControlNet [367] has shown that fine-tuning these models allows for very fine-grained control over the output with various conditioning modalities.
II. Learning with noisy conditioning has been widely explored in classification. For binary classification,[215] study machine learning robustness with noisy labels, while[120] train a DNN with exclusively positive labels and confidence scores. [18] introduced instance-dependent noise scored by confidence. The negative impact of noisy labels has been mitigated with changes in architecture [92, 43], in the loss [251], or filtering noisy samples [102]. Recently, [139] propose conditioning an image captioner model by the CLIP-score to mitigate the impact of text misalignment. Instead, we focus on image synthesis, conditioning the diffusion model with a coherence score.

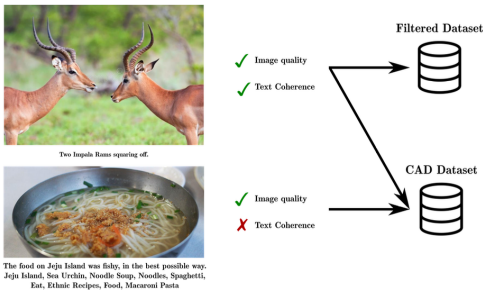


Figure 3.4: **Motivation for Coherence-aware diffusion**.

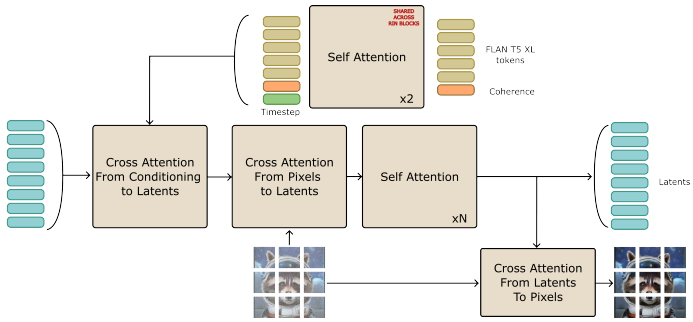


Figure 3.5: Architecture of the proposed Text RIN Block used in CAD.

Approach overview. We assume that for every datapoint (X, y) we have an associated condition c, the coherence score of y where $c \in [0, 1]$. Our goal is to incorporate label coherence into the diffusion model to discard only the conditioning that contains low levels of coherence while continuing to train on the image. A value of c=1 indicates that y is the best possible annotation for X, while c=0 suggests that y is a poor annotation for X. To achieve this, we modify the conditioning of the diffusion model ϵ_θ to include both y and c, using the following loss:

$$L_{\text{simple}} = \mathbb{E}_{(X,y,c)\sim p_{\text{data}}, t\sim\mathcal{U}[0,1]}[\|\epsilon - \epsilon_\theta(X_t, y, c, t)\|] \quad . \tag{3.2}$$

We refer to this kind of models as coherence-aware diffusion (CAD) models. By informing the diffusion model of the coherence score associated with samples, we avoid filtering out low-confident samples and let the model learn by itself what information to take into account. Avoiding the filtering allows us to still learn X even in the presence of noisy labels. In practice, to enable text-conditioned image generation, we propose a modification to the RIN architecture, coined Text RIN Block (see Figure 3.5). First, the text tokens are mapped with the coherence with 2x self-attention layers initialized with LayerScale [310] (top part in the figure) and 16 registers [55]. This mapping is the same for every Text RIN block.

Test-time prompting. After training with different levels of coherence, we can also prompt it with varying degrees of coherence. Prompting with minimal coherence leads to an unconditional model. When we prompt with maximal coherence, we get a model that is very confident about the provided label. To strengthen the use of the label, we propose a modification to the Classifier Free Guidance (CFG) method [112] that leverages the coherence. CFG uses both a conditional and unconditional model to improve the quality of generated samples. To learn such models, a conditional diffusion model is used and the conditioning is dropped out for a portion of the training samples. The CFG formulation is:

$$\hat{\epsilon}_\theta(x_t, y) = \epsilon_\theta(x_t, y) + \omega(\epsilon_\theta(x_t, y) - \epsilon_\theta(x_t, \emptyset)) \quad , \tag{3.3}$$

with ω the guidance rate. Instead, we propose a coherence-aware version of CFG (CA-CFG):

$$\hat{\epsilon}_\theta(x_t, y) = \epsilon_\theta(x_t, y, 1) + \omega(\epsilon_\theta(x_t, y, 1) - \epsilon_\theta(x_t, y, 0)) \quad . \tag{3.4}$$

This modification removes the need to dropout the conditioning. Instead, we directly use the noise in the conditioning to drive the guidance.

Experiments. For text-conditional image generation, we use a modified version of RIN [122]. To map the text to an embedding space, we use a frozen FLAN-T5 XL [45]. We then map the embedding with 2 self-attention transformer layers initialized with LayerScale [311]. We finally add the conditioning to the latent branch of RIN at each RIN Block with a cross-attention layer.

Datasets. We train CAD on a mix of datasets composed of CC12M [33] and LAION Aesthetics 6+ [270].

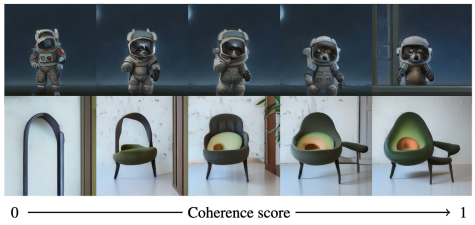


Figure 3.6: Images generated with *coherence score* between the prompt and the target image. The score varies from 0 (no coherence) to 1 (maximum coherence). Higher coherence scores tend to generate images that adhere more to the prompt.

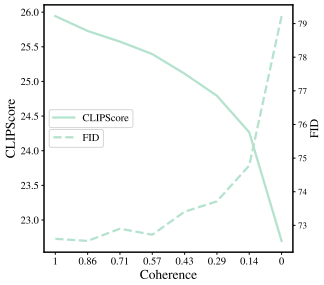


Figure 3.7: Increasing the coherence from 0 to 1, CLIPScore increases and FID decreases.

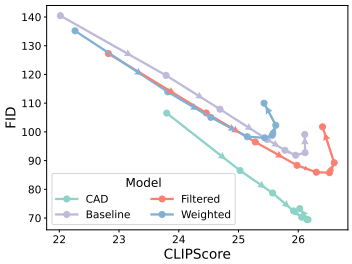


Figure 3.8: FID vs CLIP score for varying guidance ω. CAD achieves a better trade-off with a much lower FID for the same CLIP score.

Top prompt: *"a raccoon wearing an astronaut suit. The racoon is looking out of the window at a starry night; unreal engine, detailed, digital painting,cinematic,character design by pixar and hayao miyazaki, unreal 5, daz, hyperrealistic, octane render"*, bottom prompt: *"An armchair in the shape of an avocado"*

Metrics. We evaluate image quality with the Frechet Inception Distance [110] (FID), and CLIP Score [240] and evaluate metrics on CLIP features on a 10K samples subset of COCO [171] in a zero-shot setting.
Quantitative results. Here, we explore the behavior of our proposed coherence-aware diffusion model at test time. For the text conditional setting, Figure 3.7 shows that the coherence and the quality of the generated image increase as the coherence increases. Indeed, FID decreases and the CLIPScore increases. In Figure 3.8, we observe that our method achieves a significantly better FID/CLIP tradeoff than the other methods. We corroborate these results with a user study, where we generate images for randomly sampled captions in COCO with various methods and ask users to vote for the image with the best quality, and for the one with the most coherence to the prompt. Users prefer the image quality of our images in 95% of the cases and find our images better aligned with the prompts by 89%.
Qualitative results. We prompt the model with varying coherence scores from 0 to 1 and display results in Figure 3.6. When the coherence increases, the outputs are close to the prompt. In the bottom figure, the generated image displays an avocado armchair, where avocado and armchair are well mixed. Even the raccoon generation at the top follows closely the complex textual prompt. The raccoon does wear an astronaut suit and is looking through the window at a starry night. Similarly, as the coherence decreases, the images start to diverge from the original prompt. The avocado chair starts to first lose the "avocado" traits until there is only a chair and at the end an object that does not look like an avocado or a chair. At the top, we first lose the window, then the raccoon. Note that we do not converge to a totally random image. Instead, some features from the prompt are preserved, such as the astronaut suit and the starry night. This is highly linked to the CLIP network biases, which may pay less attention to less salient parts of an image such as the background, and are more sensitive to the main subject.

Discussion. We proposed a novel method for training conditional diffusion models with additional coherence information. By incorporating coherence scores into the conditioning process, our approach allows the model to dynamically adjust its reliance on the conditioning. We also extend the classifier-free guidance, enabling the derivation of conditional and unconditional models without the need for dropout during training. We have demonstrated that our method, called condition-aware diffusion (CAD), produces more diverse and realistic samples on various text-to-image generation tasks.

Impact and Limitations. The main limitation of CAD lies in the extraction of coherence scores, as unreliable coherence scores can lead to biases. Future research includes focusing on more robust and reliable methods for obtaining coherence scores to further improve the generalizability of CAD.

3.1.3 Analysis of Classifier-Free Guidance Weight Schedulers

Diffusion models have shown strong generative capabilities across various domains: images [111], acoustic signals [138], videos [184]. Conditional generation with diffusion, such as text-conditioned image generation, is often achieved by adding an extra condition input [216] and has been explored in numerous works [264, 260, 14]. Classifier Guidance [61] combines gradients from a separately trained classifier with a diffusion model, while Classifier-Free Guidance (CFG) [112] uses a Bayesian implicit classifier to achieve condition reliance without an external classifier. In both cases, a weighting parameter ω controls the generative and guidance terms, applied at all timesteps. Varying ω is a trade-off between fidelity and condition reliance. Recent works have explored dynamic guidance: MUSE [31] suggests a linearly increasing guidance weight can enhance performance and diversity, an approach adopted in Stable Video Diffusion [19] and discussed in [83] through a parameterized cosine-based curve (pcs4). However, these studies lack empirical validation for dynamic guidance weight schedulers. For instance, MUSE [31] briefly mentions linear guidance and pcs4 [83] is only discussed in the appendix.

This work bridges this gap by exploring the behaviour of diffusion guidance and systematically examining the impact of dynamic schedulers on visual generation. We explore various heuristic and parametrized dynamic schedulers across tasks and datasets.

Related work. Recently, diffusion models have excelled in image synthesis [287, 111], text-to-image [61, 257, 236, 234], and text-to-motion [42]. Conditioned diffusion models use additional input for control, with two main methods: Classifier Guidance (CG)[61] and Classifier-Free Guidance (CFG)[112]. CFG has improved text-conditional generation significantly, employing text encoders like CLIP [238] for models such as Stable Diffusion [257] and SDXL. For CG, [374] used entropy to rescale gradients, while [65] employed multiple class conditions and [64] suggested gradient projection. In CFG, [158] developed a zero-shot classifier, and [31, 83] proposed dynamic guidance schedulers. Instead, we extensively study dynamic guidance for various tasks.

Approach overview. Following DDPM [111], diffusion consists in training a network ϵ_θ to denoise a noisy input to recover the original data at different noise levels. A common practice nowadays in conditional diffusion is the *Classifier-Free Guidance (CFG)* [112], which is controlled by the guidance ω, the condition c and is defined as:

$$\hat{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t, c) + \omega\left(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t)\right) \quad . \tag{3.5}$$

We can reformulate the above equation into two terms: a *generation* term $\epsilon_\theta(x_t) \propto \nabla_{x_t} \log p(x_t)$ and a *guidance* term $\nabla_{x_t} \log p(c|x_t)$. The guidance term can be derived either from a pre-trained classifier or an implicit one, with the guidance ω balancing between generation and guidance.

Dynamic guidance. We observe from [152] that removing the guidance at certain timesteps improves the performance over using a *static* weight ω for CFG like in [112, 61]. Therefore, we ask the question whether we can replace static guidance with other options. To this end, we investigate *dynamic* guidance scheduler that evolves throughout the generation process, also in line with some empirical schemes mentioned in recent literature [19, 31, 66]. In that case, the CFG Equation 3.5 is rewritten as follows:

$$\hat{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t, c) + \omega(\mathbf{t})\left(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t)\right) \quad . \tag{3.6}$$

To identify an effective dynamic scheduler $\omega(\mathbf{t})$, we analyse two types of function in subsequent paragraphs: parameter-free heuristic schedulers and single-parameter parameterized ones.

Dynamic Guidance: 1. Heuristic Schedulers. We use six simple heuristic schedulers as dynamic guidance $\omega(t)$, split into three groups depending on the shape of their curve: (a) increasing functions

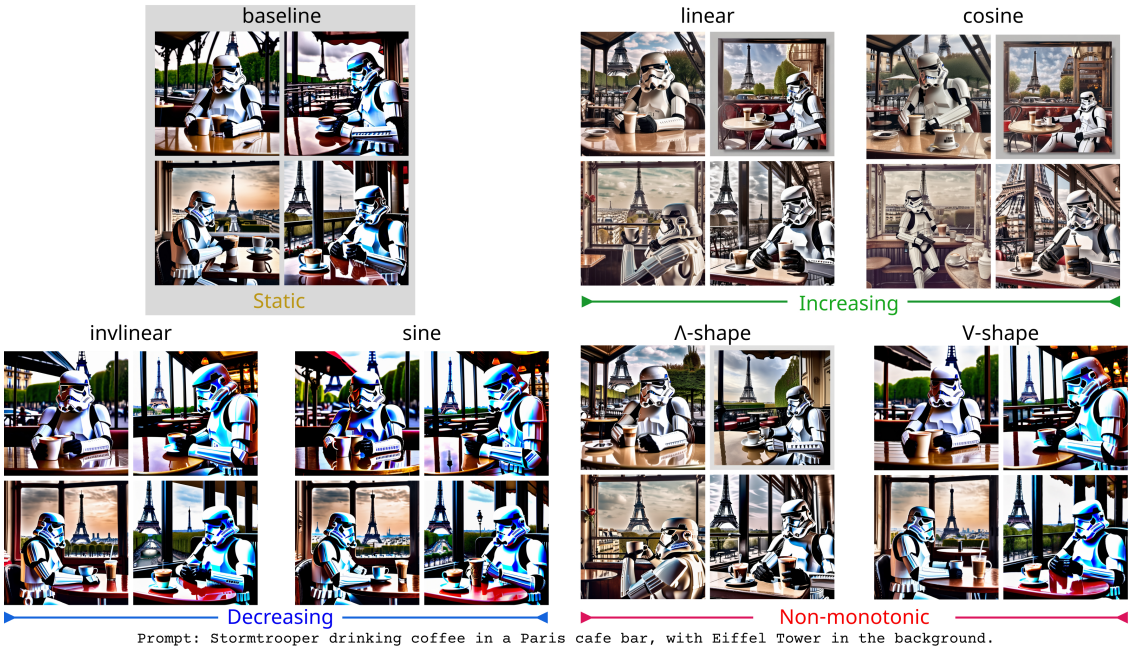


Figure 3.9: **Examples of all heuristics** on SDXL. Increasing ones (*linear* and *cosine*) enhance fidelity, textual adherence and diversity.

(linear, cosine); (b) decreasing functions (inverse linear, sine); (c) non-monotonic functions (linear V-shape, linear Λ-shape), defined as:

$$\text{linear: } \omega(t) = 1 - t/T, \qquad \text{invlinear: } \omega(t) = t/T,$$
$$\text{cosine: } \omega(t) = \cos(\pi t/T) + 1, \qquad \text{sine: } \omega(t) = \sin(\pi t/T - \pi/2) + 1,$$
$$\text{V-shape: } \omega(t) = \text{invlinear}(t) \textit{ if } t < T/2, \qquad \Lambda\text{-shape: } \omega(t) = \text{linear}(t) \textit{ if } t < T/2,$$
$$\text{linear}(t) \textit{ else}, \qquad \text{invlinear}(t) \textit{ else}.$$

To allow for a direct comparison between the effect of these schedulers and the static guidance ω, we normalize each scheduler by the area under the curve: $\int_0^T \omega(t)dt = \omega T$.

Experiments with heuristic schedulers. **Metrics.** To assess the performance, we use the Frechet Inception Distance (FID) and Inception Score (IS) metrics, over $50,000$ inference from 50-step DDIM [287]. For text, we also use CLIP score (CS).

Class-conditional image generation. We study the 6 heuristic schedulers $\omega(t)$ on the CIFAR-10-DDPM setting for class-conditional synthesis. In this experiment, we evaluate the effect of various guidance weight on the image quality vs class adherence trade-off. The results show that both increasing schedulers (linear and cosine) significantly improve over the static baseline, whereas decreasing schedulers (invlinear and sine) are significantly worse than the static. The V-shape and Λ-shape schedulers perform respectively better and worse than the static baseline, but only marginally.

Text-to-image generation. In FID, both the linear and cosine schedulers achieve better FID-CS than the baseline [236]. In Diversity, linear is slightly lower than cosine, and they are both better than static baseline. Additionally, unlike the baseline where higher guidance typically results in compromised FID, heuristic schedulers counter this. Figure 3.9 depicts the text-to-image generations with all heuristic schedulers from SDXL.

Findings with heuristic schedulers. Combining with our initial observation that removing the beginning stage improves the performance, they point to the same conclusion: **monotonically increasing guidance schedulers** achieve improved performances, revealing that the static CFG primarily may overshoot the guidance in the initial stages. In summary, we make the following observations: monotonically increasing heuristic schedulers (e.g., linear and cosine) (a) improve generation performances (IS/CS vs. FID) over static baseline on different models; (b) improve image fidelity (texture, details), diversity (composition, style) and quality (lighting, gestures). We note that this gain is achieved without hyperparameter tuning, retraining or extra computational cost.

Dynamic Guidance: 2. Parametrized Schedulers. We also investigate two parameterized schedulers with a tunable parameter to maximize performance: a power-cosine curve family (introduced in MDT [83]) (**pcs**) and two clamping families (linear and cosine), with controllable parameter s and c, respectively. They are defined:

$$w_t = \frac{1 - \cos \pi \left(\frac{T-t}{T}\right)^s}{2} w \qquad \textbf{(pcs)} \tag{3.7}$$

$$w_t = \max(c, w_t) \qquad \textbf{(clamp)} \tag{3.8}$$

Experiments with parametrized schedulers. **Text-to-image generation** with the clamp-linear and pcs schedulers. For SD1.5 [257], the pcs struggles to achieve low FID, except when $s = 1$. Conversely, the clamp family exhibits optimal performance around c=2. For SDXL [236], the pcs shows the best performance at $s = 0.1$. Clamp-linear achieves optimum at $c = 4$ (FID 18.2), largely improving FID across the entire CS range compared to the baseline (FID 24.9, about 30% gain) and the linear scheduler. The optimal parameters of clamp-linear (resp. pcs) are not the same for both models, i.e. c=2 for SD1.5 and c=4 for SDXL (resp. s=1 and s=0.1 for pcs). This reveals the lack of generalizability of this family.

Findings with parametrized schedulers. Our observations are: (a) tuning the parametrized functions improves the performance for both generation tasks, (b) tuning clamps seems easier than pcs family, as its performance shows fewer variations, and (c) the optimal parameters for one method do not generalize across different settings. Thus, specialized tuning is required for each model and task, leading to extensive grid searches and increased computational load.

Discussion. We analyzed dynamic schedulers for the weight parameter in Classifier-Free Guidance by systematically comparing heuristic and parameterized schedulers. We experiment on two tasks (class-conditioned generation and text-to-image generation), several models (DDPM, SD1.5 and SDXL) and various datasets (more details in [326]). Our findings are: (1) a simple monotonically increasing scheduler systematically improves the performance compared to a constant static guidance, at no extra computational cost and with no hyper-parameter search. (2) parameterized schedulers with tuned parameters per task, model and dataset, improve the results. They, however, do not generalize well to other models and datasets as there is no universal parameter that suits all tasks.

For practitioners who target state-of-the-art performances, we recommend searching or optimizing for the best clamping parameter. For those not willing to manually tune parameters per case, we suggest using heuristics, specifically linear or cosine.

Impact. By enhancing the efficiency and performance of generative models, our work contributes to advancements in areas such as content creation, where AI-generated images and media are increasingly used. This can democratize access to high-quality content production, enabling creators and industries to produce innovative work at lower costs. Same as in Section 3.1.2, a notable concern is to the potential misuse of AI for creating realistic deepfakes or misleading content or violating one's privacy.

3.2 Camera motion generation

Cinematography is a collaborative process mixing technical, artistic, and storytelling skills to convey a distinct message through scene layout, lighting, and camera placement. The camera plays a critical role in conveying the director's intention, using a common language known as film grammar. However, mastering camera placements and motions remains challenging, especially for novices. To ease this, several methods have been proposed to compute camera trajectories, including geometric [20, 172], optimization- and control-based [71, 82], and deep learning [131, 21, 115, 71] approaches. These either use cinematic-rule-based control [115, 21, 82] or example-based imitation [131, 130, 325] but often require designing specific geometric models or cost functions for each motion, limiting creative combinations.

Recent advances in video generation [329, 373] allow for more creative camera motion in generated videos. [132] addressed camera trajectory generation using diffusion models with high controllability but faced limitations due to a character-centric coordinate system and oversimplified evaluation metrics. In other domains, generative techniques use large datasets with textual descriptions (such as language-motion [235, 100]), but cinematography lacks such datasets with crucial cinematic information. Most recent approaches use synthetic data [131, 130, 132] or general videos without cinematic features [115, 378]. Example-based approaches for cinematic transfer from real film clips[325, 133] offer limited control and variability, without encoding cinematographic knowledge.

Here, we propose the *E.T. the Exceptional Trajectories* camera trajectory dataset from real movie clips with camera and character trajectories and textual descriptions. We also introduce Director, a diffusion-based model that generates camera trajectories using text descriptions and character information.

Related work. *I. Camera control* evolved from geometric and rule-based controls [20, 172, 71] to deep learning. [131] introduced a Mixture-of-Experts model for 3D animations, with keyframing [130]. [325] optimized trajectories using Neural Radiance Field [203]. Example-based methods struggle with generalization due to the need for selected reference videos. In drone cinematography, cinematic-rule-based methods use Deep Reinforcement Learning (DRL) and Imitation Learning (IL). [115, 21] used IL and DRL for drone control, while [340] used aesthetic score-based RL. RL-based methods require environment-specific training and limit trajectory diversity. We use diffusion models for better generalization.

II. Camera diffusion. Generative models, especially diffusion models, have advanced in image, video, and motion generation [258, 236, 217, 283, 19, 308, 42, 368]. Diffusion models produce high-fidelity samples [339, 61], making them suitable for camera trajectory generation. The Cinematographic Camera Diffusion (CCD) [132], based on MDM [308], uses synthetic data and a character-centric coordinate system. Our E.T. dataset uses a global coordinate system with a rich vocabulary and extensive data.

III. Camera trajectory datasets. Large multimodal datasets are common in generative methods, like LAION [271] for text-to-image generation and KIT and HumanML3D [235, 100] for human motion synthesis. Few datasets exist for camera control [378, 132]. The RealEstate10K dataset [378] focuses on smooth 3D reconstruction movements, lacking cinematic complexity. Jiang et al.[132] introduced a synthetic dataset that oversimplifies cinematic dynamics. SLAHMR [90] offers robust 3D human pose estimation, motivating our E.T. dataset with enhanced captions for camera and character trajectories.

Proposed dataset. The key properties of E.T. are: **Cinematic content.** E.T. features realistic and cinematic camera trajectories from real movies, unlike RealEstate10k's [378] smooth trajectories and CCD's [132] synthetic data. **Scale.** Built on $16,210$ scenes from CMD [12], E.T. includes 115K samples, 11M frames, and 120 hours of footage. It surpasses KIT [235] and HumanMl3D [100] in scale and exceeds CCD [132] in hours, frames, and samples. Though comparable to RealEstate10k [378], E.T. adds character trajectories and captions from real movies. **Controllability.** E.T. offers camera and character trajectories with captions, providing semantic information and a user-friendly format. In contrast, RealEstate lacks captions, and CCD's limited captions focus only on camera data. E.T.'s rich captions match the vocabulary size of human motion datasets like KIT and HumanML3D.

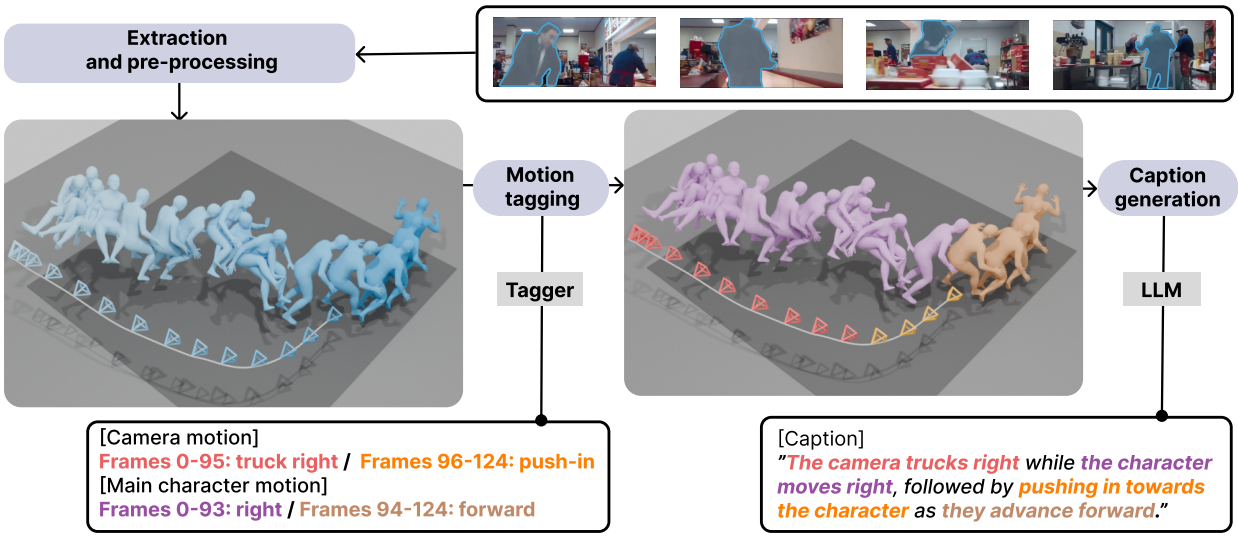


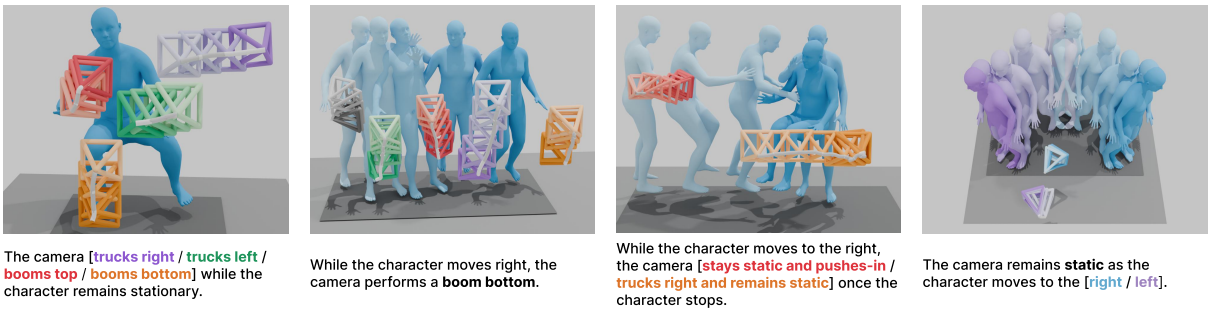
Figure 3.10: **E.T. creation.** Given RGB frames, we first extract and pre-process camera and character poses, then tag resulting camera and character trajectories (sequence of poses) to obtain rough descriptions (middle), which we finally, we translate into rich textual captions, aligning the camera trajectory with that of the character (right).

Dataset	#Samples	#Frames	#Hours	Domain	Character		Camera		#Vocabulary
					Traj	#Captions	Traj	#Captions	
KIT Motion-Language [235]	4K	0.8M	11.23	Mocap	✓	6K		-	1,623
HumanML3D [100]	14K	2M	28.59	Mocap	✓	45K		-	**5.371**
RealEstate10k [378]	79K	**11M**	**121**	Youtube		-	✓	-	-
CCD [132]	25K	4.5M	50	Synthetic		-	✓	25K	48
E.T. (Ours)	**115K**	**11M**	120	**Movie**	✓	**115K**	✓	**230K**	1,790

Table 3.1: **Dataset comparison.** We compare the E.T. dataset to (i) two human motion datasets KIT [235] and HumanML3D [100]; and (ii) camera trajectory datasets RealEstate10K [378] and CCD [132]. Here the notion of sample is common across all datasets and corresponds to data associated with a continuous temporal sequence.

Dataset creation pipeline. E.T. is created through a three-step process (Figure 3.10). (1) We extract and refine 3D coordinates of cameras and characters into uniform trajectories. (2) We perform *motion tagging* by dividing each trajectory into segments that represent pure camera movements, which we then label. (3) We generate captions describing both camera and character trajectories over time. We use SLAHMR [352] for extracting camera and character poses, and apply pre-processing steps like alignment, filtering, smoothing, and cropping (up to 300 frames) as described in [100]. For *camera trajectory tagging*, we utilize rigid body velocity $\in SE(3)$ to distinguish between different camera movements, such as 'trucking' and 'depth'. For *character trajectory tagging*, we assume characters face their movement direction, identify the main character per shot following [314], and generate captions for both camera and character trajectories. Lastly, inspired by [57], we use the LLM Mistral-7B [129] to convert motion tagging descriptions into detailed textual annotations.

Approach overview. We introduce the *DiffusIon tRansformEr Camera TrajectORy* (Director) method for camera trajectory generation. Director takes as input the character trajectory with the camera-character caption and generates a camera trajectory. Additionally, we present the *Contrastive Language-Trajectory* embedding (CLaTR) that serves as a basis for creating a common space between text and trajectories, enabling the computation of evaluation metrics. The base of Director is a pre-norm Transformer [315, 341]. We condition the transformer on the diffusion timestep, the character trajectory, and a textual description that describes the relative movement between the camera and character trajectories. Inspired by the DiT architecture variants [233], we explore three distinct ways to include the conditioning in the denoising process: Director A adds the conditioning to the context of the transformer input; Director B



(a) Controllability (b) Diversity (c) Complexity (d) Character-aware

Figure 3.11: **Qualitative results.** Generated camera trajectories with corresponding prompts and character trajectories, highlighting (a) controllability, (b) diversity, (c) complexity, and (d) character awareness. Darker shades indicate later frames.

concatenates the conditionings into a single token for an AdaLN operation; and Director C uses the full sequence length of the conditioning with cross-attention.

Experiments. **Metrics.** We use two sets of metrics. For camera trajectory quality, we use the Frechet CLaTr Distance (FD_{CLaTr}), Precision (P), Recall (R), Density (D), and Coverage (C)[213]. For text-camera coherence, we use the CLaTr-Score (CS)[109].
Datasets. We train and evaluate two subsets of the E.T. dataset: the pure camera trajectory subset and the mixed camera trajectories subset. We compare Director with concurrent methods on both subsets.
Quantitative results. Director outperforms CCD [132] and MDM [308] on all metrics and both subsets. It achieves a margin of -3.0 FD_{CLaTr} against MDM and -32.1 against CCD, and a substantial improvement of $+3.6$ ClaTR-Score against MDM and $+15.7$ against CCD.
Ablation of Director Architectures. Director C outperforms other variants, followed by Director A. Director B excels in text-camera coherence on the pure trajectory subset but struggles on the mixed trajectory subset due to AdaLN's limitations in capturing sequential complexity.

Qualitative results. Figure 3.11 shows generated camera trajectories from Director (architecture C), including keyframes, character meshes, and captions. We highlight four key strengths: (1) Controllability. Director allows precise control; modifying two words in the caption generates diverse camera movements like *trucks right*" and *booms bottom*". (2) Diversity. It produces varied camera trajectories from the same input conditions, offering a range of creative outputs. (3) Complexity. Director handles complex scenarios, including varied character movements and camera descriptions. (4) Character-awareness. It generates camera movements that align with the character's trajectory.

Discussion. We introduced E.T., a dataset of camera and character trajectories from movies, with accompanying text captions. We demonstrated how E.T. enables training Director, a diffusion-based method that sets a new state-of-the-art in camera trajectory generation. Future work will enhance caption expressiveness with more detail on modifiers and character positioning.

Impact. *Creative Integrity:* It is a fine line between using an AI tool to enhance human creativity and allowing it to deprive the human creative process. Under misusage, the proposed method could diminish the artistic expression instead of supporting or complementing it. *Intellectual Property*: The use of AI-generated content raises questions about ownership and copyright. The Intellectual Property ownership of the generated content can be debatable. *Job Displacement or Creation*: The automation of certain aspects of filmmaking could lead to concerns about job displacement within the industry, or under proper usage, may also help to create new types of jobs in the domain.

Chapter 4

Multimodality in Medical Applications

Here, our goal is to propose novel, and generalizable computer vision methods for medical imaging, with focus on forecasting issues in individuals by leveraging imaging and clinicobiological data.

In Section 4.1, we first describe our CosEmb [204, 205] masked-transformer-based model for forecasting organ transplant rejections through the serum creatinine prediction from follow-up exams of MRI data post-transplantation. Then, inspired by LLMs, in [206] we extended CosEmb and introduced the MEDIMP model that learns multi-modal representations of renal transplants by also incorporating structural clinicobiological data after translating them into text prompts; being one of the first works on medical imaging that exploited text prompting. Our code and models are available online at `https://github.com/leomlck/renal_transplant_imaging` and `https://github.com/leomlck/MEDIMP`. This work has been published at MICCAI 2022 [204], MIDL 2022 [205] and MIDL 2023 [206].

Section 4.2 details our work for multimodal learning for detecting physiological changes under missing modalities. Multimodality has recently gained attention in the medical domain, where imaging or video modalities may be integrated with biomedical signals or health records. Yet, two challenges remain: balancing the contributions of modalities, especially in cases with a limited amount of data available, and tackling missing modalities. To address both issues, in this work, we introduce ADAPT, a multimodal, scalable model with two key components: (i) aligning all modalities in the space of the strongest, richest modality (called *anchor*) to learn a joint embedding space, and (ii) a Masked Multimodal Transformer, leveraging both inter- and intra-modality correlations while handling missing modalities. We focus on detecting stress changes in individuals. Our code and models are available online: `https://github.com/jumdc/ADAPT`. This work has been published at MIDL 2024 [208] and CVPR-W 2024 [209].

4.1 Renal transplant failure prediction

Renal transplantation is more cost-effective than long-term dialysis and significantly improves quality of life [294]. However, it carries a risk of chronic dysfunction that can lead to graft loss or patient death [105]. While blood tests and urine samples (e.g., serum creatinine) are primary indicators of kidney function, irregular results often require needle biopsy, an invasive surgical operation. Thus, the need for a non-invasive alternative is crucial. Medical imaging could play a crucial role in renal transplantation. Due to the typically few and small annotated datasets, the community has turned to learning robust medical representations, typically involving two stages: first, applying self-supervised or weakly-supervised methods [295, 150, 296] like contrastive or adversarial learning [291, 9, 24] to imaging datasets; second, fine-tuning these representations for specific tasks. Pre-trained models often outperform those trained on ImageNet but can still produce suboptimal results if they capture only spurious correlations [7].

Our work [204, 205, 206] addresses this by forecasting renal transplant function through serum creatinine prediction. We propose self-supervised techniques [204, 205] for analyzing Dynamic Contrast-Enhanced (DCE) MRI data post-transplantation and use contrastive learning with images and medical prompts [206] to improve DCE MRI representations. Our methods CosEmb and MEDIMP handle missing data robustly and result to superior performance compared to other imputation strategies, marking a significant advancement in forecasting serum creatinine from imaging and multimodal LLM data.

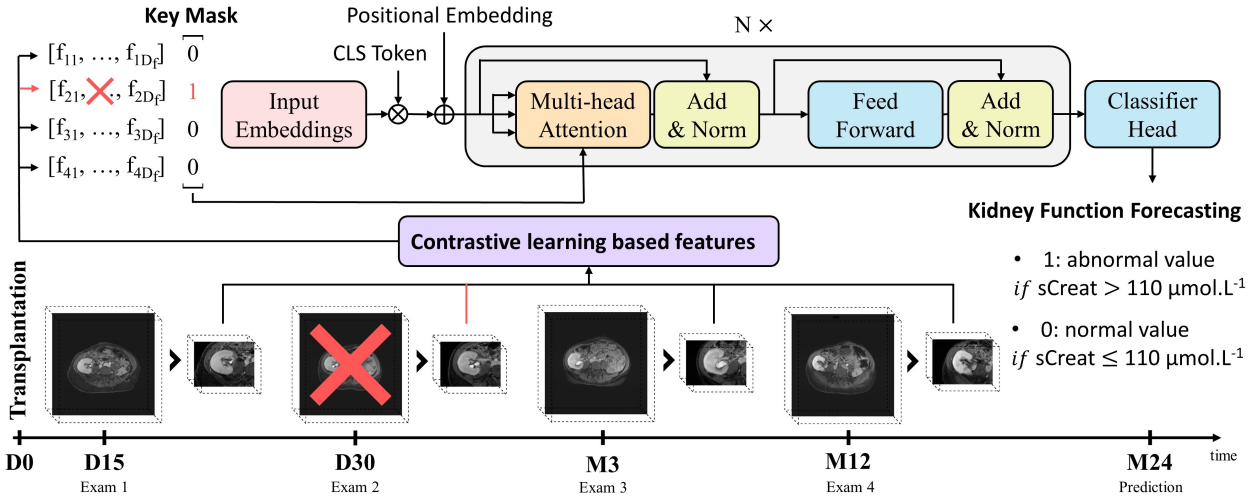


Figure 4.1: **Overview of the proposed CosEmb method.** Different contrastive schemes are used to represent the different MRIs. These features are used to train a sequential model coupled with a key mask tensor to mark the missing data.

Related work. *I. Renal transplant dysfunction.* Recent methods focus on detecting specific conditions like renal fibrosis [224] or acute rejection [146]. Studies, such as [279], utilize multi-modal MRI and clinical data to assess renal allograft status across various exams. These methods aim to non-invasively retrieve structural, functional, and molecular information to diagnose chronic kidney disease [5].
II. Missing data is a critical issue in clinical data curation, typically addressed by imputation methods. Traditional approaches include statistical methods and discriminative models like structured prediction [145]. Generative approaches, such as expectation-maximization algorithms [84] and Generative Adversarial Imputation Nets (GAIN)[353], also perform well in medical imaging [53, 336]. However, these models often require large datasets [144], which may be limited in clinical settings. Recent advancements with transformer-based attention mechanisms show promise for imputation in both structural [335] and trajectory data [17, 89], including using attention masks in models like BERT [60]. Our work [204, 205] is among the first to robustly handle high-dimensional missing data in long sequences.
III. Multimodal medical imaging. Advances in Natural Language Processing (NLP) have enhanced weakly-supervised tasks in computer vision. Multiview contrastive learning [10] has been applied to jointly train image and text encoders [370, 239, 128, 212]. For natural images, [239] used million image-text pairs to achieve strong performances. Studies like[370] used chest X-rays and radiology descriptions, while [212] extended this approach to tasks like semantic segmentation and object detection. These approaches primarily focus on 2D images and the MIMIC-CXR dataset, which is labor-intensive to curate and mostly contains information about imaging exams rather than comorbidities.

Approach overview. In this work, we propose three approaches.
I. Medical Contrastive Pre-training. In medical imaging, data is scarce and datasets are small; thus, meaningful pre-trainings are essential. To this end, in [205], we first propose a self-supervised scheme, where we learn meaningful features by solving the proxy task of determining if two MRI volumes belong to the same patient. For this, we create a two-stream model, with each ResNet-based stream taking as input MRI volumes and outputting features; then, a feature embedding head associates these features with the proxy task. From the embedded features (z'_1, z'_2), the optimization by the loss:

$$\text{CosEmbLoss}(z'_1, z'_2, y) = \begin{cases} 1 - \cos(z'_1, z'_2), & \text{if } y = 1, \\ \max(0, \cos(z'_1, z'_2)), & \text{if } y = 0 \end{cases}, \tag{4.1}$$

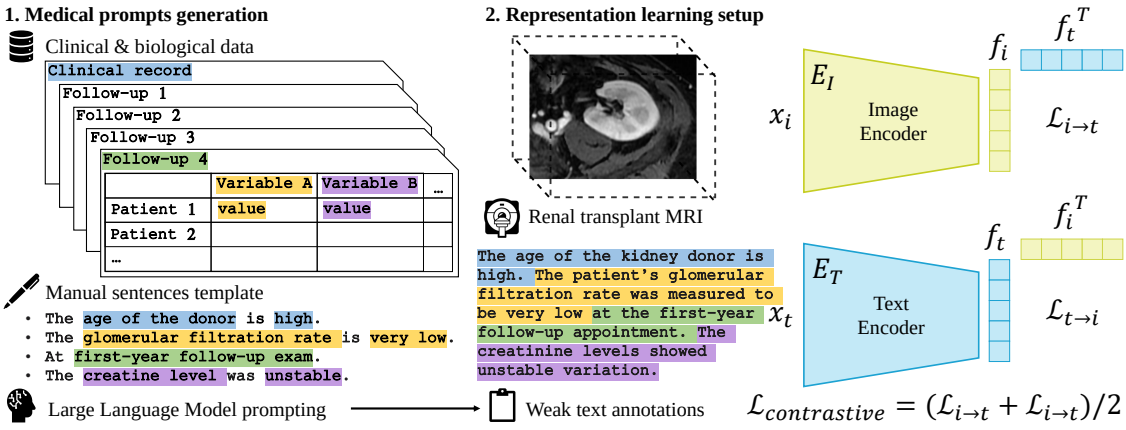


Figure 4.2: **Overview of our proposed method MEDIMP – Medical Images and Prompts.** 1. Medical prompts are generated from clinicobiological data using predefined templates of sentences, given as inputs to Large Language Models to produce augmented text data. 2. The medical prompts are used to learn multi-modal representations of renal transplants DCE MRI using contrastive learning from image-text pairs.

where cos refers to the cosine similarity and z_1', z_2' refer to the features from each stream in the embedding head. This loss enforces the model to build relevant features that express adequately the kidney transplant imaging and define the way to create strategies to label y each pair.

II. Model for Renal failure prediction with missing data. We propose the CosEmb model for renal transplant failure prediction that handles missing data (Figure 4.1). Following the transformer model [315], CosEmb takes as input features corresponding to the different follow-ups (patient exams), and through multi-head attention and normalization layers performs the classification by using the CLS token output. To deal with missing data, we propose to build a key mask tensor $m_k \in \mathbb{R}^T$ based on the availability of exams for each patient so that zero attention is given to missing data both during the training and inference times, i.e. $\forall t \in [\![1, T]\!]\ m_k[t] = -\infty$ if exam t is available else 0. Thus, our mask cancels the attention on missing exams by $\text{Attention}(Q, K, V) = \text{softmax}(\frac{(QK^t)}{\sqrt{d_k}} + M_k)V$ where query Q, key K and value V and $M_k = [\![m_k m_k ... m_k]\!] \in \mathbb{R}^{T \times d_k}$.

III. Multimodal model for Renal failure prediction. In this work [206], we exploit image-text pairing with contrastive learning, as well as the encoding capabilities of recent advances in LLMs [28, 222, 239]. For this, we propose a framework that generates textual data from structural clinicobiological data that describe variables used in clinical practice and linked to the graft survival. Then, we use these textual data together with existing medical imaging features in a contrastive manner for prediction.

Specifically, we propose the Medical Image with Prompts (MEDIMP) method, which consists of two components. First, we construct medical prompts from structural clinical and biological data as follows (left side of Figure 4.2): (a) based on guidance from medical experts, we create several *template sentences* per variable of interest, e.g., "the GFR of the patient is very low at the first-year follow-up exam" or "The transplant patient's GFR is assessed as very low at the date follow-up examination.". For this, we produce $N = 10$ textual data augmentations for each template sentence with the dialogue ChatGPT [222]. This textual augmentation offers richness and variation in descriptions, aiding the training. Second, following a CLIP-like scheme, we pre-train an image (with 3D MRI volumes) and a text encoder (with tokenized text). Specifically, we follow [41] and deploy the InfoNCE loss [220] that maximizes a lower bound on the mutual information between the two modalities (right part of Figure 4.2).

Experiments. **Metrics.** For evaluation we use F1 score and ROC AUC.

Datasets. Approved by the Institutional Review Board, we experimented on a dataset with 105 subjects, where each subject underwent up to 4 follow-up exams.

Quantitative results. By comparing CosEmb and MEDIMP with the same level of information, we observe that CosEmb outperforms MEDIMP by 1% F1 when the text information is not very rich due to

its smaller and more compact size. However, when more variables are integrated, MEDIMP prevails reaching 87.8% F1 vs 78.1% for CosEmb. Overall, MEDIMP with all medical prompts results in the best predictions at 2 (85.0% AUC) and 4 (75.7% AUC) years post-transplantation.

Discussion. We introduced three methods aiming at transplant function forecasting in the context of renal transplantation monitoring. First, we examined the significance of contrastive learning schemes by proposing two self and weakly supervised pre-training schemes tailored to medical imaging. Second, we proposed a CosEmb, a transformer encoder architecture with a custom method to handle missing data. Third, we proposed MEDIMP, where we extended CosEmb by including clinical or biological information in the learning process. For this, we leveraged recent advances in LLMs and translated, in a structured and systematic way, clinicobiological information into text. We then trained MEDIMP with text and imaging pairs in a CLIP-like manner. ThOur experiments showed improved representation learning for renal transplant MRI compared to previous state-of-the-art methods, particularly in forecasting renal transplant function. MEDIMP enhances representations by leveraging LLMs and integrating clinical and biological data, potentially advancing the understanding of complex medical phenomena. These promising results advocate using textual data from LLMs to assist in training robust medical models.

Impact and Limitations. Despite our state-of-the-art results, several limitations remain. First, while data scarcity is a challenge, additional test data would validate the generalizability of our methods. Currently, no public medical imaging dataset offers comprehensive longitudinal imaging and clinical data for prognosis. Still, our framework could be adapted to similar datasets. Second, this work is an initial effort to generate medical prompts from limited clinical and biological variables, essential for complex medical concepts. We plan to expand this by incorporating more variables to guide image encoder training, which can be done by defining additional templates for automatic text generation. Third, despite the remarkable results of LLMs, their reliability is a major concern, particularly with potentially inaccurate outputs. In this work, we used ChatGPT for robust text augmentations without leaking sensitive clinical information. Further exploration of LLMs, including specialized prompt engineering, could enhance their utility in medical contexts.

4.2 Multimodal learning for detecting physiological changes under missing modalities

Monitoring physiological changes to external stimuli is crucial for assessing individuals' well-being, particularly in contexts with medical and safety implications. Examples include stress, a response to emotional, mental, and physical challenges [267], and a triggering or aggravating factor for various pathological conditions [63]. High-performance environments, such as exposure to g-forces in aircraft, can lead to alterations in consciousness [211]. At the same time, drowsiness during driving poses a critical physiological response with safety implications, contributing to road accidents and fatalities [293]. Various sensors report physiological changes that may be detected visually (videos), acoustically (audio), or from biomedical signals (e.g., electrocardiograms). However, specific modalities may be missing during training and testing. Therefore, developing methods capable of handling missing modalities during both stages while balancing modalities' contributions is crucial to ensure robustness, notably when modalities with strong unimodal performances are severely missing.

Various methods address the challenge of missing modalities, each with notable limitations, including (1) bias towards the most available modalities leading to sub-optimal performance [148], (2) dependence on complete modalities during training [191, 36] (3) limited generalizability to more than two modalities [188, 187], and (4) utilization of a shared encoder tailored for modalities with inputs of the same dimensions which complicates extension to heterogeneous modalities like imaging and biomedical signals [148]. To address the above issues, we introduce the **A**nchore**D** multimod**A**l **P**hysiological

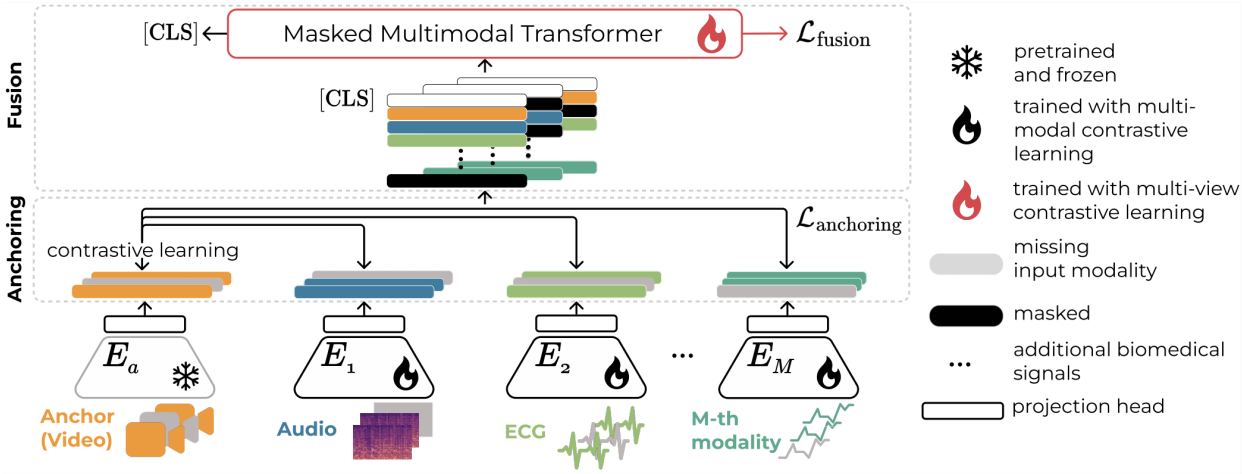


Figure 4.3: **Overview of ADAPT.** In each minibatch, ADAPT takes up to M modalities, including video, audio, and biosignals, as input to produce a modality-agnostic representation for downstream tasks. It is trained in two steps. (i) *Anchoring.* We align the representations of all modalities via contrastive learning to the one of an *anchor* modality, i.e., the strongest and richest modality; here the video. (ii) *Fusion.* The encoders' features are concatenated and fed into the Masked Multimodal Transformer. When a modality is unavailable, the transformer masks its corresponding feature representations. The final [CLS] token is used for downstream tasks.

Transformer (ADAPT) that is designed to operate effectively under missing modalities both during training and inference enabling robust real-life applicability [208, 209].

Related work. *I. Handling missing modalities.* Missing modalities pose a persistent challenge in Multimodal Learning, particularly in medical imaging, due to privacy concerns or impractical data acquisition [166, 8]. Various strategies have been explored to address this challenge. *Knowledge Distillation* [114, 191, 320] involves learning from a teacher network trained on complete modality data. *Generative modeling* aims to impute missing inputs by generating synthetic data [354, 274]. Both approaches rely on complete modality at training, which can be insufficient for robust training. Another line of work is *common space modeling*, which learns a shared latent space from partially available modalities [188, 148, 319]. SMIL [188] perturbs the latent feature space to approximate the embedding of missing modalities but is limited to bi-modal datasets, limiting its generalizability. ShaSpec [319] addresses more than two modalities by learning shared and specific features, but its use of a shared encoder complicates generalization to heterogeneous modalities of different dimensions. Additionally, shared latent space modeling may introduce biases toward the most available modalities [148]. Simultaneously, cross-modal contrastive learning has shown impressive results [371, 206, 159] by aligning multimodal data in a joint embedding space. Recently, ImageBind [88] aligned six modalities by relying on image-paired data and emphasized that aligning all pair combinations is unnecessary to bind more than two modalities together. Our proposed ADAPT advances this by training unimodal encoders solely with supervision from one modality, aligning them in a joint embedding space. It ensures that every modality contributes to the final representation, even if it is severely missing during training.
II. Multimodal transformer. Transformers [315] are the de facto approach for multimodal tasks [247, 292]. They rely on the attention mechanism to model long-range dependencies with the flexibility to account for incomplete samples. [205, 187] efficiently handle missing data in sequences and bimodal datasets through masked attention. ADAPT extends this to more than two modalities by leveraging attention to fuse them and exploring their inter- and intra-modal correlations while masking missing ones. We also perform a systematic study of missing modalities during training and testing, showing the versatility and potential of ADAPT for real-life scenarios.

Approach Overview. ADAPT consists of two key components and is illustrated in Figure 4.3. First, our goal is to embed all modalities in the same feature space. Instead of optimizing one loss per modality pair, which would result in quadratic growth of training time, we align each modality to one frozen modality, called *anchor*. It allows learning a joint embedding space with linear scalability and balancing each modality's contribution. We call this step the 'anchoring'. In this work, anchor is the video, as it can capture visually distinguishable physiological changes; however, any modality can be the anchor. Second, it comprises a Masked Multimodal Transformer that leverages inter- and intra-modality correlations to concatenate features from different modalities into a unified representation. Additionally, inspired by our previous work [205], we leverage the masked attention of transformers [315] to ensure flexibility in handling missing modalities similar to [187, 205]. When a modality is unavailable, its corresponding feature representation is masked. Furthermore, drawing inspiration from [280], we mitigate the model's over-reliance on a single modality while enhancing its robustness in the absence of modalities through an augmentation technique called *modality dropout*. We leverage the masking scheme at the attention level to randomly mask out input modalities. The transformer is trained using two objectives: self-supervised learning and the objective of the downstream task.

Experiments. **Metrics.** We use the Accuracy and weighted F1 score (F1).
Datasets. We use `StressID` [34] for stress identification, which contains physiological responses via electrocardiogram, electrodermal activity, respiration, audio, and videos. We use the train, val, and test splits from [34]. `StressID` has 18% and 46% of missing video and audio recordings, respectively.
Quantitative results. For `StressID`, we observe that ADAPT outperforms all methods from [34] by a notable margin, reaching F1 of 75.9% and Acc of 69.5%, while remaining highly competitive with 'feature fusion' and 'decision-level fusion' (rows 4 & 5)[34] (66.0% and 72.0% F1) and ShaSpec+ [319] (75.7% F1).
Robustness to missing modalities. We examine the performance of various methods when removing various modalities. When we remove audio and/or video, the most cumbersome modalities to acquire, and examine the performances: Acc: *no-audio* 68.3%, *no-video* 61.2% and *real-life* (i.e., no audio, no video) 60.0% vs 69.5% when using all modalities. We observe that even by removing modalities, ADAPT successfully detects stress with more than 60% Accuracy, highlighting its ability to handle missing modalities, in contrast to all other methods unable to address this.

Discussion. In this work [208, 209], we proposed ADAPT, a modality-agnostic representation framework designed to operate effectively under missing modalities during both training and testing. Our framework has been challenged on two different tasks targeting the detection of physiological changes, outperforming the current state of the art while showcasing its superiority for handling missing modalities. Our analysis highlights the robustness of our method in different scenarios and strategies. Future work includes applications to other medical tasks.

Impact. Using computer vision in medical imaging can have a positive societal impact. Accurate and reliable analysis is a valuable source in medical imaging, for automatically analyzing medical data [204, 205] to identify abnormalities or to automatically generate reports from visual data that can assist medical professionals. Specifically, the technology we develop could potentially contribute to: (a) Improved diagnosis and treatment of diseases: By using computer vision techniques to analyze medical images, doctors and other healthcare providers could potentially identify diseases and other medical conditions more accurately and quickly [208, 209]. This could lead to earlier and more effective treatment of diseases, which could improve patient outcomes and save lives; (b) Improved patient experience with more personalized and tailored care by using interpretable techniques [99], thus leading to a better experience for patients; and (c) Improving rare medical issues: Modern AI techniques require big data, which is unrealistic when it comes to medical imaging due to patients privacy and rarity of diseases. For this, the techniques she developed require no or few data for prediction and instead exploit the underlying structure of medical imaging for compact and informative representations [204, 205].

Chapter 5

Discussion

5.1 Perspectives

This thesis has touched upon several problems in multimodal visual content recognition and generation from the story perspective. The unifying theme of the work is to address the complexity of multimodality in settings where story or other artistic elements are crucial split into two directions: understanding scenes in videos (in our case edited videos, mostly movies) and conditional generation of visual content (mostly text-to-image generation). The research area of the latter is in its infancy, and our work contributes to its first steps rather than to a final solution. Our contributions are summarized below.

Chapter 2 explores multimodality for cinematic stories in edited videos both for humans and for scenes. For this, we first presented methods for identifying the interaction of mutual gaze between people [194] together with identifying underlying social relationships based on people's gazing [195] (*65 citations*). Then, our work and dataset on person-clustering in videos [137, 26] (*45 citations*) based on multiple modalities (face, body, voice) showed the importance of multiple modalities for person-level reasoning for downstream applications such as story understanding. Next, we discussed scene understanding, where the focus lied on comprehending the overall context and content of a scene. Specifically, we focused on multimodality for cinematography by learning funny moments in videos [178, 179] (*15 citations*) and by proposing a dataset for story-level question answering [86]. This body of work highlights the power of leveraging multiple modalities in videos, demonstrating the enhanced interpretative capacity for story-level reasoning across a range of applications.

Chapter 3 explores how to generate visual content by exploring multiple modalities as conditions, such as text-to-image generation. In [73] (*15 citations*) we explored semantically-conditioned image generation, where we showed that having multiple latents per semantic region improves fine image details and handling coarse semantic labels such as the background. Then, we explored conditional image generation with diffusion (including text-to-image generation) with methods and insights that lead to more diverse and realistic samples: by incorporating coherence scores into the conditioning, the diffusion model can dynamically adjust its reliance on the conditioning [72] and by exploring various weight schedulers we showed the best practices for improved performances [326]. Finally, we showcased the importance of text-camera trajectory pairs for cinematography [50]. This body of work highlights how multimodality can improve the performance and the creative potential of visual content generation, from detailed image generation to dynamic camera movement in cinematic storytelling.

Chapter 4 refers to a real-life application of multiple modalities in the medical domain, in particular for medical forecasting. Here, we explored two directions: first, how to combine imaging modalities with clinicobiological signals and second, how to make robust and reliable predictions, in such a critical domain as medical data, when data are scarce and when modalities are not always matched or paired (missing modalities). For the former, our work on renal transplant failure prediction [206] (*2 citations*) showed that translating clinical data to text can help the integration with visual data and our work on detecting physiological changes [208, 209] showed the importance of modality alignment for such integration. For the latter, our work identifying physiological changes under missing modalities in the low data regime [204, 205, 208, 209] (*5 citations*) showed that transformer masking is the most suitable methodology.

5.2 Future work

The work presented in this thesis attempts to bridge the gap between story-level understanding methods and visual content generation. But, now... Sora [176] is here! So, what's next?

Audio-to-Video Generation. One of the major advancements in generative models is the generation of video from audio. Currently, most approaches convert speech to text before generating videos, but this intermediate step may limit the expressive potential. A promising direction is generating video directly from speech, bypassing the text modality. This approach would allow models to incorporate essential speech elements, such as tone, pitch, emotion, and rhythm, which are critical for conveying narrative and mood. By moving beyond text, such systems could more accurately reflect the nuances of spoken dialogue, offering richer and more authentic audiovisual experiences.

Fundamental Research in Multimodality. The challenge of multimodal alignment remains an open problem. Text is often used as the central modality for aligning other forms of data such as images, video, and audio; yet, this reliance on text may not be optimal [88]. Future research should explore whether bypassing traditional modalities, such as text, could enhance multimodal alignment. This could involve developing an entirely artificial alignment mechanism not tied to a single existing modality, enabling more fluid and efficient cross-modal generation. Such research would enhance the capabilities of generative models and pave the way for more intuitive interaction between humans and AI.

Memorability of Stories: Emotion, Place, and Narrative. Memorability is a key factor in storytelling, particularly in visual media such as film. Three elements—emotion, setting (place), and narrative—have been shown to be fundamental to making stories memorable [22]. Investigating how these elements manifest in existing media, such as short films, can provide insights into their influence on audience retention and emotional engagement. A structured analysis of how variations in these factors impact memorability could help inform future models that generate content optimized for long-term viewer impact, tailoring narratives to evoke stronger emotional and cognitive responses.

Incorporating Memorability in Visual Content Generation. Building upon the understanding of story, emotion, and place, future research should explore how these elements can be directly integrated into conditional image and video generation. Current generative models often focus on visual coherence and aesthetics, but incorporating memorability as a core element could transform how content is produced. By embedding narrative depth, emotional tone, and contextual significance into generated visuals, models could create content that is not only visually compelling but also emotionally resonant and cognitively engaging, driving new applications in film, advertising, and beyond.

Bibliography

[1] Abele, A.: Functions of gaze in social interaction: Communication and monitoring. Journal of Nonverbal Behavior (1986)

[2] Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. In: arXiv (2016)

[3] Afouras, T., Chung, J.S., Zisserman, A.: The conversation: Deep audio-visual speech enhancement. In: Proc. INTERSPEECH (2020)

[4] Ajodan, E.L., Clark-Whitney, E., Silver, B., Silverman, M.R., Southerland, A., et al.: Increased eye contact during parent-child versus clinician-child interactions in young children with autism. PsyArXiv (2019)

[5] Alnazer, I., Bourdon, P., Urruty, T., Falou, O., Khalil, M., Shahin, A., Fernandez-Maloigne, C.: Recent advances in medical image processing for the evaluation of chronic kidney disease. Med. Image Anal. **69**, 101960 (2021)

[6] Annamoradnejad, I., Zoghi, G.: Colbert: Using bert sentence embedding for humor detection. arXiv preprint arXiv:2004.12765 (2020)

[7] Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. Eprint arXiv:1907.02893 (2019)

[8] Azad, R., Khosravi, N., Dehghanmanshadi, M., Cohen-Adad, J., Merhof, D.: Medical image segmentation on mri images with missing modalities: A review. arXiv preprint arXiv:2203.06217 (2022)

[9] Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Natarajan, V., Norouzi, M.: Big self-supervised models advance medical image classification. In: International Conference on Computer Vision (ICCV). pp. 3458–3468. IEEE (2021)

[10] Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. In: NeurIPS. vol. 32. Curran Associates, Inc. (2019)

[11] Bain, M., Nagrani, A., Brown, A., Zisserman, A.: Condensed movies: Story based retrieval with contextual embeddings. In: Proc. Asian Conf. on Computer Vision (2020)

[12] Bain, M., Nagrani, A., Brown, A., Zisserman, A.: Condensed movies: Story based retrieval with contextual embeddings. In: ACCV (2020)

[13] Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: ICCV (2022)

[14] Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arxiv (2022)

[15] Barral, O., Kosunen, I., Jacucci, G.: No need to laugh out loud: Predicting humor appraisal of comic strips based on physiological signals in a realistic environment. ACM Transactions on Computer-Human Interaction (TOCHI) (2017)

[16] Bauml, M., Tapaswi, M., Stiefelhagen, R.: Semi-supervised learning with constraints for person identification in multimedia data. In: CVPR (2013)

[17] Becker, S., Hug, R., Huebner, W., Arens, M., Morris, B.T.: Missformer: (in-)attention-based handling of missing observations for trajectory filtering and prediction. In: Advances in Visual Computing: 16th International Symposium (ISVC). p. 521–533. Springer International Publishing (2021)

[18] Berthon, A., Han, B., Niu, G., Liu, T., Sugiyama, M.: Confidence scores make instance-dependent label-noise learning possible. Proc. ICML (2021)

[19] Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)

[20] Blinn, J.: Where am I? what am I looking at? (cinematography). IEEE Computer Graphics and Applications (1988)

[21] Bonatti, R., Wang, W., Ho, C., Ahuja, A., Gschwindt, M., Camci, E., Kayacan, E., Choudhury, S., Scherer, S.: Autonomous aerial cinematography in unstructured environments with learned artistic decision-making. J. Field Robotics. (2020)

[22] Bondebjerg, I.: Documentary and cognitive theory: Narrative, emotion and memory. Media and Communication (2014)

[23] Bose, D., Hebbar, R., Somandepalli, K., Zhang, H., Cui, Y., Cole-McLaughlin, K., Wang, H., Narayanan, S.: Movieclip: Visual scene recognition in movies. In: Proc. WACV (2022)

[24] Boyd, J., Liashuha, M., Deutsch, E., Paragios, N., Christodoulidis, S., Vakalopoulou, M.: Self-supervised representation learning using visual field expansion on digital pathology. In: International Conference on Computer Vision (ICCV). pp. 639–647. IEEE (2021)

[25] Brau, E., Guan, J., Jeffries, T., Barnard, K.: Multiple-gaze geometry: Inferring novel 3D locations from gazes observed in monocular video. In: ECCV (2018)

[26] Brown, A., Kalogeiton, V., Zisserman, A.: Face, body, voice: Video person-clustering with multiple modalities. In: ICCV (2021)

[27] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: NeurIPS (2020)

[28] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: NeurIPS. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)

[29] Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., Poria, S.: Towards multimodal sarcasm detection (an *Obviously* perfect paper). In: ACL (2019)

[30] Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: ICCV (2019)

[31] Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. In: Proc. ICML (2023)

[32] Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. CVPR (2021)

[33] Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. CVPR (2021)

[34] Chaptoukaev, H., Strizhkova, V., Panariello, M., D'alpaos, B., Reka, A., Manera, V., Thümmler, S., Ismailova, E., Evans, N., Bremond, F.F., et al.: Stressid: a multimodal dataset for stress identification. In: NeurIPS (2023)

[35] Chen, B., Ziai, A., Tucker, R., Xie, Y.: Match cutting: Finding cuts with smooth visual transitions. In: Proc. WACV (2022)

[36] Chen, C., Dou, Q., Jin, Y., Liu, Q., Heng, P.A.: Learning with privileged multimodal knowledge for unimodal segmentation. IEEE transactions on medical imaging (2021)

[37] Chen, D.L., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (2011)

[38] Chen, S., He, X., Guo, L., Zhu, X., Wang, W., Tang, J., Liu, J.: Valor: Vision-audio-language omni-perception pretraining model and dataset. In: arXiv (2023)

[39] Chen, S., Li, H., Wang, Q., Zhao, Z., Sun, M., Zhu, X., Liu, J.: Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. In: NeurIPS (2023)

[40] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proc. ICML (2020)

[41] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (ICML). PMLR (2020)

[42] Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: CVPR (2023)

[43] Cheng, L., Zhou, X., Zhao, L., Li, D., Shang, H., Zheng, Y., Pan, P., Xu, Y.: Weakly supervised learning with side information for noisy labeled images. ECCV (2020)

[44] Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A., Rehg, J.M.: Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In: ECCV (2018)

[45] Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv (2022)

[46] Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Proc. Asian Conf. on Computer Vision (2016)

[47] Chung, S.W., Chung, J.S., Kang, H.G.: Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In: Proc. ICASSP (2019)

[48] Cinbis, R.G., Verbeek, J., Schmid, C.: Unsupervised metric learning for face identification in tv video. In: ICCV (2011)

[49] Cour, T., Sapp, B., Jordan, C., Taskar, B.: Learning from ambiguously labeled images. In: CVPR (2009)

[50] Courant, R., Dufour, N., Wang, X., Christie, M., Kalogeiton, V.: Et the exceptional trajectories: Text-to-camera-trajectory generation with character awareness. In: ECCV (2024)

[51] Courant, R., Lino, C., Christie, M., Kalogeiton, V.: High-level features for movie style understanding. In: ICCV-W (2021)

[52] Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. In: NeurIPS (2023)

[53] Dalca, A.V., Bouman, K.L., Freeman, W.T., Rost, N.S., Sabuncu, M.R., Golland, P.: Medical image imputation from image collections. IEEE Trans. Med. Imaging **38**, 504–514 (2019)

[54] Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV (2018)

[55] Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. ICLR (2024)

[56] Davidov, D., Tsur, O., Rappoport, A.: Semi-supervised recognition of sarcastic sentences in twitter and amazon. In: ACL (2010)

[57] Delmas, G., Weinzaepfel, P., Lucas, T., Moreno-Noguer, F., Rogez, G.: Posescript: 3d human poses from natural language. In: ECCV (2022)

[58] Deng, D., Zhou, Y., Pi, J., Shi, B.E.: Multimodal utterance-level affect analysis using visual, audio and text features. arXiv preprint arXiv:1805.00625 (2018)

[59] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL (2019)

[60] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL). vol. 1, pp. 4171–4186. Association for Computational Linguistics (2019)

[61] Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. NeurIPS (2021)

[62] Diba, A., Fayyaz, M., Sharma, V., Paluri, M., Gall, J., Stiefelhagen, R., Gool, L.V.: Large scale holistic video understanding. In: ECCV (2020)

[63] Dimsdale, J.E.: Psychological stress and cardiovascular disease. Journal of the American College of Cardiology (2008)

[64] Dinh, A.D., Liu, D., Xu, C.: Pixelasparam: A gradient view on diffusion sampling with guidance. In: Proc. ICML. PMLR (2023)

[65] Dinh, A.D., Liu, D., Xu, C.: Rethinking conditional diffusion sampling with progressive guidance. In: NeurIPS (2023)

[66] Donahue, C., McAuley, J., Puckette, M.: Adversarial audio synthesis. In: ICLR (2018)

[67] Dong, H., Liang, X., Gong, K., Lai, H., Zhu, J., Yin, J.: Soft-gated warping-gan for pose-guided person image synthesis. In: NeurIPS (2018)

[68] Doosti, B., Chen, C.H., Vemulapalli, R., Jia, X., Zhu, Y., Green, B.: Boosting image-based mutual gaze detection using pseudo 3D gaze. arXiv preprint arXiv:2010.07811 (2020)

[69] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)

[70] Drouard, V., Horaud, R., Deleforge, A., Ba, S., Evangelidis, G.: Robust head-pose estimation based on partially-latent mixture of linear regressions. IEEE Transactions on Image Processing (2017)

[71] Drucker, S.M., Galyean, T.A., Zeltzer, D.: Cinema: A system for procedural camera movements. In: Symposium on Interactive 3D graphics (1992)

[72] Dufour, N., Besnier, V., Kalogeiton, V., Picard, D.: Don't drop your samples! coherence-aware training benefits conditional diffusion. In: CVPR (2024)

[73] Dufour, N., Picard, D., Kalogeiton, V.: Scam! transferring humans between images with semantic cross attention modulation. In: ECCV (2022)

[74] Endo, Y., Kanamori, Y.: Diversifying semantic image synthesis and editing via class-and layer-wise vaes. In: Computer Graphics Forum (2020)

[75] Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR (2021)

[76] Everingham, M., Sivic, J., Zisserman, A.: Hello! my name is... buffy"–automatic naming of characters in tv video. In: BMVC (2006)

[77] Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: ICCV (2021)

[78] Fitzgibbon, A.W., Zisserman, A.: On affine invariant clustering and automatic cast listing in movies. In: ECCV (2002)

[79] Fu, T.J., Li, L., Gan, Z., Lin, K., Wang, W.Y., Wang, L., Liu, Z.: Violet : End-to-end video-language transformers with masked visual-token modeling. In: arXiv (2022)

[80] Fu, T.J., Li, L., Gan, Z., Lin, K., Wang, W.Y., Wang, L., Liu, Z.: An empirical study of end-to-end video-language transformers with masked visual modeling. In: CVPR (2023)

[81] Gabbay, A., Ephrat, A., Halperin, T., Peleg, S.: Seeing through noise: Visually driven speaker separation and enhancement. In: Proc. ICASSP (2018)

[82] Galvane, Q., Christie, M., Lino, C., Ronfard, R.: Camera-on-rails: automated computation of constrained camera paths. In: ACM Motion In Games (2015)

[83] Gao, S., Zhou, P., Cheng, M.M., Yan, S.: Masked diffusion transformer is a strong image synthesizer. In: Int. Conf. Comput. Vis. (2023)

[84] García-Laencina, P.J., Sancho-Gómez, J.L., Figueiras-Vidal, A.R.: Pattern classification with missing data: a review. Neural Comput. Appl. **19**(2), 263–282 (2010)

[85] Ghaleb, E., Tapaswi, M., Al-Halah, Z., Ekenel, H.K., Stiefelhagen, R.: Accio: A data set for face track retrieval in movies across age. In: Proc. ICMR (2015)

[86] Ghermi, R., Wang, X., Kalogeiton, V., Laptev, I.: Story-level video understanding in short movies. In: submitted to NeurIPS 2024 (2024)

[87] Gillick, J., Deng, W., Ryokai, K., Bamman, D.: Robust laughter detection in noisy environments. Proc. INTERSPEECH (2021)

[88] Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: CVPR (2023)

[89] Giuliari, F., Hasan, I., Cristani, M., Galasso, F.: Transformer networks for trajectory forecasting. In: 25th International Conference on Pattern Recognition (ICPR). pp. 10335–10342. IEEE (2021)

[90] Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa*, A., Malik*, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: ICCV (2023)

[91] Goffman, E.: Behavior in public places. Simon and Schuster (2008)

[92] Goldberger, J., Ben-Reuven, E.: Training deep neural-networks using a noise adaptation layer. ICLR (2016)

[93] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. NeurIPS (2014)

[94] Goyal, R., Kahou, S.E., Michalski, V., Materzyńska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The "something something" video database for learning and evaluating visual common sense. In: ICCV (2017)

[95] Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., Gonzalez, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolar, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Puentes, P.R., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhao, Z., Zhu, Y., Arbelaez, P., Crandall, D., Damen, D., Farinella, G.M., Fuegen, C., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR (2022)

[96] Grunde-McLaughlin, M., Krishna, R., Agrawala, M.: Agqa: A benchmark for compositional spatio-temporal reasoning. In: CVPR (2021)

[97] Gu, C., Sun, C., Ross, D., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., Malik, J.: AVA: A video dataset of spatio-temporally localized atomic visual actions. In: CVPR (2018)

[98] Gu, S., Bao, J., Yang, H., Chen, D., Wen, F., Yuan, L.: Mask-guided portrait editing with conditional gans. In: CVPR (2019)

[99] Guanghui, F., Wang, R., Li, J., Vakalopoulou, M., Kalogeiton, V.: Me-ndt: Neural-backed decision tree for visual explainability of deep medical models. In: MIDL (2021)

[100] Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: CVPR (2022)

[101] Guzhov, A., Raue, F., Hees, J., Dengel, A.: Audioclip: Extending clip to image, text and audio. In: Proc. ICASSP (2022)

[102] Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. NeurIPS (2018)

[103] Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W., Zisserman, A.: Autoad: Movie description in context. In: CVPR (2023)

[104] Han, X., Hu, X., Huang, W., Scott, M.R.: Clothflow: A flow-based model for clothed person generation. In: ICCV (2019)

[105] Hariharan, S., Israni, A.K., Danovitch, G.: Long-term survival after kidney transplantation. The New England Journal of Medicine **385**(8), 729–743 (2021)

[106] Hasan, M.K., Lee, S., Rahman, W., Zadeh, A., Mihalcea, R., Morency, L.P., Hoque, E.: Humor knowledge enriched transformer for understanding multimodal humor. In: AAAI (2021)

[107] Hasan, M.K., Rahman, W., Bagher Zadeh, A., Zhong, J., Tanveer, M.I., Morency, L.P., Hoque, M.E.: UR-FUNNY: A multimodal language dataset for understanding humor. In: Proc. EMNLP (2019)

[108] Hazarika, D., Zimmermann, R., Poria, S.: Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. ACM International Conference on Multimedia (2020)

[109] Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. In: Proc. EMNLP (2021)

[110] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS (2017)

[111] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020)

[112] Ho, J., Salimans, T.: Classifier-free diffusion guidance. NeurIPS (2022)

[113] Hoogeboom, E., Heek, J., Salimans, T.: simple diffusion: End-to-end diffusion for high resolution images. arxiv (2023)

[114] Hu, M., Maillard, M., Zhang, Y., Ciceri, T., La Barbera, G., Bloch, I., Gori, P.: Knowledge distillation from multi-modal to mono-modal segmentation networks. In: Internation Conference on Medical Image Computing and Computer Assisted Intervention (2020)

[115] Huang, C., Lin, C., Yang, Z., Kong, Y., Chen, P., Yang, X., Cheng, K.: Learning to film from professional human motion videos. In: CVPR (2019)

[116] Huang, Q., Liu, W., Lin, D.: Person search in videos with one portrait through visual and temporal links. In: ECCV (2018)

[117] Huang, Q., Xiong, Y., Rao, A., Wang, J., Lin, D.: Movienet: A holistic dataset for movie understanding. In: ECCV (2020)

[118] Hudson, D.A., Zitnick, C.L.: Generative adversarial transformers. In: Proc. ICML (2021)

[119] Iashin, V., Rahtu, E.: Multi-modal dense video captioning. In: CVPR-W (2020)

[120] Ishida, T., Niu, G., Sugiyama, M.: Binary classification from positive-confidence data. NeurIPS (2018)

[121] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)

[122] Jabri, A., Fleet, D., Chen, T.: Scalable adaptive computation for iterative generation. Proc. ICML (2022)

[123] Jaegle, A., Borgeaud, S., Alayrac, J.B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., Hénaff, O., Botvinick, M.M., Zisserman, A., Vinyals, O., Carreira, J.: Perceiver io: A general architecture for structured inputs & outputs. In: arXiv (2021)

[124] Jaegle, A., Gimeno, F., Brock, A., Zisserman, A., Vinyals, O., Carreira, J.: Perceiver: General perception with iterative attention. In: Proc. ICML (2021)

[125] Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In: CVPR (2017)

[126] Jarvis, R.A., Patrick, E.A.: Clustering using a similarity measure based on shared near neighbors. IEEE Trans. Computers (1973)

[127] Jia, B., Lei, T., Zhu, S.C., Huang, S.: Egotaskqa: Understanding human tasks in egocentric videos. In: NeurIPS (2022)

[128] Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning (ICML). PMLR (2021)

[129] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7B. arXiv preprint arXiv:2310.06825 (2023)

[130] Jiang, H., Christie, M., Wang, X., Liu, L., Wang, B., Chen, B.: Camera keyframing with style and control. ACM TOG (2021)

[131] Jiang, H., Wang, B., Wang, X., Christie, M., Chen, B.: Example-driven virtual cinematography by learning camera behaviors. ACM TOG (2020)

[132] Jiang, H., Wang, X., Christie, M., Liu, L., Chen, B.: Cinematographic camera diffusion model. Comput. Graph. Forum (2024)

[133] Jiang, X., Rao, A., Wang, J., Lin, D., Dai, B.: Cinematic behavior transfer via nerf-based differentiable filming. arXiv preprint arXiv:2311.17754 (2023)

[134] Jiang, Y., Chang, S., Wang, Z.: Transgan: Two transformers can make one strong gan. In: CoRR (2021)

[135] Jin, S., Su, H., Stauffer, C., Learned-Miller, E.: End-to-end face detection and cast grouping in movies using erdos-renyi clustering. In: ICCV (2017)

[136] Kahatapitiya, K., Ranasinghe, K., Park, J., Ryoo, M.S.: Language repository for long video understanding. In: arXiv (2024)

[137] Kalogeiton, V., Zisserman, A.: Constrained video face clustering using 1nn relations. In: BMVC (2020)

[138] Kang, M., Min, D., Hwang, S.J.: Grad-stylespeech: Any-speaker adaptive text-to-speech synthesis with diffusion models. In: ICASSP (2023)

[139] Kang, W., Mun, J., Lee, S., Roh, B.: Noise-aware learning from web-crawled image-text data for image captioning. arxiv (2022)

[140] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)

[141] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: CVPR (2020)

[142] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset. In: arXiv preprint arXiv:1705.06950 (2017)

[143] Kayatani, Y., Yang, Z., Otani, M., Garcia, N., Chu, C., Nakashima, Y., Takemura, H.: The laughing machine: Predicting humor in video. In: Proc. WACV (2021)

[144] Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., Mukhopadhyay, A.: Gans for medical image analysis. Artif. Intell. Med. **109**, 101938 (2020)

[145] Keshavan, R.H., Montanari, A., Oh, S.: Matrix completion from a few entries. IEEE Trans. Inf. Theory **56**(6), 2980–2998 (2010)

[146] Khalifa, F., Beache, G.M., El-Ghar, M.A., El-Diasty, T., Gimel'farb, G., Kong, M., El-Baz, A.: Dynamic contrast-enhanced mri-based early detection of acute renal transplant rejection. IEEE Trans. Med. Imaging **32**(10), 1910–1927 (2013)

[147] Koizumi, Y., Masumura, R., Nishida, K., Yasuda, M., Saito, S.: A transformer-based audio captioning model with keyword estimation. arXiv preprint arXiv:2007.00222 (2020)

[148] Konwer, A., Hu, X., Bae, J., Xu, X., Chen, C., Prasanna, P.: Enhancing modality-agnostic representations via meta-learning for brain tumor segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21415–21425 (2023)

[149] Korbar, B.: Co-training of audio and video representations from self-supervised temporal synchronization. In: CoRR (2018)

[150] Krishnan, R., Rajpurkar, P., Topol, E.J.: Self-supervised learning in medicine and healthcare. Nature Biomedical Engineering **6**(12), 1346–1352 (Dec 2022)

[151] Kukleva, A., Tapaswi, M., Laptev, I.: Learning interactions and relationships between movie characters. In: CVPR (2020)

[152] Kynkäänniemi, T., Aittala, M., Karras, T., Laine, S., Aila, T., Lehtinen, J.: Applying guidance in a limited interval improves sample and distribution quality in diffusion models. arXiv preprint arXiv:2404.07724 (2024)

[153] Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: CVPR (2020)

[154] Lee, J.T., Jain, M., Park, H., Yun, S.: Cross-attentional audio-visual fusion for weakly-supervised action localization. In: ICLR (2020)

[155] Lee, K., Chang, H., Jiang, L., Zhang, H., Tu, Z., Liu, C.: Vitgan: Training gans with vision transformers. ArXiv (2021)

[156] Lei, J., Berg, T.L., Bansal, M.: Revealing single frame bias for video-and-language learning. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (2022)

[157] Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. In: Proc. EMNLP (2019)

[158] Li, A.C., Prabhudesai, M., Duggal, S., Brown, E., Pathak, D.: Your diffusion model is secretly a zero-shot classifier. In: Int. Conf. Comput. Vis. (2023)

[159] Li, J., Liu, C., Cheng, S., Arcucci, R., Hong, S.: Frozen language model helps ECG zero-shot learning. In: MIDL (2023)

[160] Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. Proc. ICML (2023)

[161] Li, K., Zhang, J., Liu, Y., Lai, Y.K., Dai, Q.: Pona: Pose-guided non-local attention for human pose transfer. In: IEEE Transactions on Image Processing (2020)

[162] Li, K., Wang, Y., Li, Y., Wang, Y., He, Y., Wang, L., Qiao, Y.: Unmasked teacher: Towards training-efficient video foundation models. In: ICCV (2023)

[163] Li, L., Chen, Y.C., Cheng, Y., Gan, Z., Yu, L., Liu, J.: Hero: Hierarchical encoder for video+language omni-representation pre-training. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (2020)

[164] Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proc. CVPR (2014)

[165] Li, Y., Li, Y., Lu, J., Shechtman, E., Lee, Y.J., Singh, K.K.: Collaging class-specific gans for semantic image synthesis. In: ICCV (2021)

[166] Liang, P.P., Lyu, Y., Fan, X., Wu, Z., Cheng, Y., Wu, J., Chen, L., Wu, P., Lee, M.A., Zhu, Y., et al.: Multibench: Multiscale benchmarks for multimodal representation learning. arXiv preprint arXiv:2107.07502 (2021)

[167] Liang, Z., Jiang, W., Hu, H., Zhu, J.: Learning to contrast the counterfactual samples for robust visual question answering. In: Proc. EMNLP (2020)

[168] Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: arXiv (2023)

[169] Lin, K., Li, L., Lin, C.C., Ahmed, F., Gan, Z., Liu, Z., Lu, Y., Wang, L.: Swinbert: End-to-end transformers with sparse attention for video captioning. In: CVPR (2022)

[170] Lin, K.Q., Wang, A.J., Soldan, M., Wray, M., Yan, R., Xu, E.Z., Gao, D., Tu, R., Zhao, W., Kong, W., Cai, C., Wang, H., Damen, D., Ghanem, B., Liu, W., Shou, M.Z.: Egocentric video-language pretraining. In: NeurIPS (2022)

[171] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. ECCV (2014)

[172] Lino, C., Christie, M.: Intuitive and efficient camera control with the toric space. ACM TOG (2015)

[173] Liu, F., Xiang, T., Hospedales, T.M., Yang, W., Sun, C.: ivqa: Inverse visual question answering. In: CVPR (2018)

[174] Liu, X., Yin, G., Shao, J., Wang, X., Li, h.: Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In: NeurIPS (2019)

[175] Liu, X., Liu, W., Zhang, M., Chen, J., Gao, L., Yan, C., Mei, T.: Social relation recognition from videos via multi-scale spatial-temporal reasoning. In: CVPR (2019)

[176] Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., et al.: Sora: A review on background, technology, limitations, and opportunities of large vision models. In: arXiv (2024)

[177] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)

[178] Liu, Z.S., Courant, R., Kalogeiton, V.: Funnynet: Audiovisual learning of funny moments in videos. In: Proc. Asian Conf. on Computer Vision (2022)

[179] Liu, Z.S., Courant, R., Kalogeiton, V.: Funnynet-w: Multimodal learning of funny moments in videos in the wild. IJCV (2024)

[180] Liu, Z.S., Kalogeiton, V., Cani, M.P.: Multiple style transfer via variational autoencoder. In: ICIP. IEEE (2021)

[181] Liu, Z.S., Wang, L.W., Siu, W.C., Kalogeiton, V.: Name your style: Text-guided artistic style transfer. In: CVPR-W (2023)

[182] Loeb, B.K.: Mutual eye contact and social interaction and their relationship to affiliation. MT (1972)

[183] Lou, S., Xu, X., Wu, M., Yu, K.: Audio-text retrieval in context. In: Proc. ICASSP (2022)

[184] Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: Videofusion: Decomposed diffusion models for high-quality video generation. In: CVPR (2023)

[185] Lv, J., Liu, W., Zhou, L., Wu, B., Ma, H.: Multi-stream fusion model for social relation recognition from videos. In: International Conference on Multimedia Modeling (2018)

[186] Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Gool, L.V.: Pose guided person image generation. In: NeurIPS (2017)

[187] Ma, M., Ren, J., Zhao, L., Testuggine, D., Peng, X.: Are multimodal transformers robust to missing modality? In: CVPR (2022)

[188] Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., Peng, X.: Smil: Multimodal learning with severely missing modality. In: Proceedings of the AAAI Conference on Artificial Intelligence (2021)

[189] Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: arXiv (2023)

[190] Maharaj, T., Ballas, N., Rohrbach, A., Courville, A., Pal, C.: A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In: CVPR (2017)

[191] Mallya, M., Hamarneh, G.: Deep multimodal guidance for medical image classification. In: MICCAI. Springer (2022)

[192] Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. In: NeurIPS (2023)

[193] Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge university press (2008)

[194] Marin-Jimenez, M.J., Kalogeiton, V., Medina-Suarez, P., Zisserman, A.: Laeo-net: revisiting people looking at each other in videos. In: CVPR (2019)

[195] Marin-Jimenez*, M.J., Kalogeiton*, V., Medina-Suarez, P., Zisserman, A.: Laeo-net++: revisiting people looking at each other in videos. In: IEEE TPAMI (2021)

[196] Marín-Jiménez, M.J., Muñoz Salinas, R., Yeguas-Bolivar, E., Pérez de la Blanca, N.: Human interaction categorization by using audio-visual cues. Machine Vision and Applications (2014)

[197] Marín-Jiménez, M.J., Zisserman, A., Eichner, M., Ferrari, V.: Detecting people looking at each other in videos. IJCV (2014)

[198] Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR (2009)

[199] Massé, B., Ba, S., Horaud, R.: Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction. IEEE TPAMI (2018). doi: 10.1109/TPAMI.2017.2782819

[200] Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: ICLR (2022)

[201] Mesaros, A., Heittola, T., Benetos, E., Foster, P., Lagrange, M., Virtanen, T., Plumbley, M.D.: Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge. ACM Transactions on Audio, Speech, and Language Processing (2017)

[202] Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: ICCV (2019)

[203] Mildenhall, B., Srinivasan, P., Tancik, M., Barron, J., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)

[204] Milecki, L., Kalogeiton, V., Bodard, S., Anglicheau, D., Correas, J.M., Timsit, M.O., Vakalopoulou, M.: Constrative learning for kidney transplant analysis using mri data and deep convolutional networks. In: MIDL (2021)

[205] Milecki, L., Kalogeiton, V., Bodard, S., Anglicheau, D., Correas, J.M., Timsit, M.O., Vakalopoulou, M.: Contrastive masked transformers for forecasting renal transplant function. In: MICCAI. Springer (2022)

[206] Milecki, L., Kalogeiton, V., Bodard, S., Anglicheau, D., Correas, J.M., Timsit, M.O., Vakalopoulou, M.: Medimp: 3d medical images with clinical prompts from limited tabular data for renal transplantation. In: MIDL (2023)

[207] Mohla, S., Pande, S., Banerjee, B., Chaudhuri, S.: Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. In: CVPR (2020)

[208] Mordacq, J., Milecki, L., Oudot, S., Kalogeiton, V.: Modality-agnostic representations for detecting loss of consciousness of fighter pilots. In: MIDL (2024)

[209] Mordacq, J., Milecki, L., Oudot, S., Kalogeiton, V.: Multimodal learning for detecting stress under missing modalities. In: CVPR-W (2024)

[210] Morgado, P., Vasconcelos, N., Misra, I.: Audio-visual instance discrimination with cross-modal agreement. In: CVPR (2021)

[211] Morrissette, K.L., McGowan, D.G.: Further support for the concept of a g-loc syndrome: a survey of military high-performance aviators. Aviation, space, and environmental medicine (2000)

[212] Müller, P., Kaissis, G., Zou, C., Rueckert, D.: Joint learning of localized representations from medical images and reports. In: European Conference on Computer Vision (ECCV). pp. 685–701. Springer (2022)

[213] Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: Proc. ICML (2020)

[214] Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. In: NeurIPS (2021)

[215] Natarajan, N., Dhillon, I.S., Ravikumar, P.K., Tewari, A.: Learning with noisy labels. NeurIPS (2013)

[216] Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Proc. ICML (2021)

[217] Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: Proc. ICML (2022)

[218] Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., Kashino, K.: Byol for audio: Self-supervised learning for general-purpose audio representation. In: International Joint Conference on Neural Networks (IJCNN) (2021)

[219] Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., Kashino, K.: Byol for audio: Exploring pre-trained general-purpose audio representations. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023)

[220] Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)

[221] OpenAI: ChatGPT: Conversational ai powered by GPT-3.5. OpenAI Blog (2021)

[222] OpenAI: Chatgpt: Optimizing language models for dialogue (2022), https://openai.com/blog/chatgpt/

[223] OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

[224] Orlacchio, A., Chegai, F., Del Giudice, C., Anselmo, A., Iaria, G., Palmieri, G., Di Caprera, E., Tosti, D., Costanzo, E., Tisone, G., Simonetti, G.: Kidney transplant: Usefulness of real-time elastography (rte) in the diagnosis of graft interstitial fibrosis. Ultrasound Med. Biol. **40**(11), 2564–2572 (2014)

[225] Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: ECCV (2018)

[226] Ozerov, A., Vigouroux, J.R., Chevallier, L., Pérez, P.: On evaluating face tracks in movies. In: ICIP (2013)

[227] Palmero, C., van Dam, E.A., Escalera, S., Kelia, M., Lichtert, G.F., Noldus, L.P., Spink, A.J., van Wieringen, A.: Automatic mutual gaze detection in face-to-face dyadic interaction videos. In: Proceedings of Measuring Behavior (2018)

[228] Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. CVPR (2019)

[229] Parkhi, O.M., Rahtu, E., Cao, Q., Zisserman, A.: Automated video face labelling for films and tv material. IEEE TPAMI (2020)

[230] Patro, B.N., Lunayach, M., Srivastava, D., Sarvesh, S., Singh, H., Namboodiri, V.P.: Multimodal humor dataset: Predicting laughter tracks for sitcoms. In: Proc. WACV (2021)

[231] Patron-Perez, A., Marszałek, M., Zisserman, A., Reid, I.D.: High five: Recognising human interactions in TV shows. In: BMVC (2010)

[232] Peebles, W., Xie, S.: Scalable diffusion models with transformers. ICCV (2022)

[233] Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)

[234] Pernias, P., Rampas, D., Aubreville, M.: Wuerstchen: Efficient pretraining of text-to-image models. arXiv preprint arXiv:2306.00637 (2023)

[235] Plappert, M., Mandery, C., Asfour, T.: The KIT motion-language dataset. Big Data (2016)

[236] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)

[237] Priyasad, D., Fernando, T., Denman, S., Sridharan, S., Fookes, C.: Attention driven fusion for multimodal emotion recognition. In: Proc. ICASSP (2020)

[238] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proc. ICML (2021)

[239] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proc. ICML. PMLR (2021)

[240] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. Proc. ICML (2021)

[241] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356 (2022)

[242] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. In: ArXiv (2018)

[243] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. In: ArXiv (2019)

[244] Rao, A., Xu, L., Xiong, Y., Xu, G., Huang, Q., Zhou, B., Lin, D.: A local-to-global approach to multimodal movie scene segmentation. In: CVPR (2020)

[245] Rawal, R., Saifullah, K., Basri, R., Jacobs, D., Somepalli, G., Goldstein, T.: Cinepile: A long video question answering dataset and benchmark. In: arXiv (2024)

[246] Recasens, A., Khosla, A., Vondrick, C., Torralba, A.: Where are they looking? In: NeurIPS (2015)

[247] Recasens, A., Lin, J., Carreira, J., Jaegle, D., Wang, L., Alayrac, J.b., Luc, P., Miech, A., Smaira, L., Hemsley, R., et al.: Zorro: the masked multimodal transformer. arXiv preprint arXiv:2301.09595 (2023)

[248] Recasens, A., Vondrick, C., Khosla, A., Torralba, A.: Following gaze in video. In: ICCV (2017)

[249] Rehg, J.M.: Behavior imaging: Using computer vision to study autism. Machine Vision and Applications (2011)

[250] Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. In: arXiv (2024)

[251] Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning. Proc. ICML (2018)

[252] Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: CVPR (2023)

[253] Ricci, E., Varadarajan, J., Subramanian, R., Rota Bulo, S., Ahuja, N., Lanz, O.: Uncovering interactions and interactors: Joint estimation of head, body orientation and F-formations from surveillance videos. In: ICCV (2015)

[254] R.J. Lehman, H.H.: idesigner 2019. In: FGVC6 (2019)

[255] Rockwell, P.: Lower, slower, louder: Vocal cues of sarcasm. Journal of Psycholinguistic research (2000)

[256] Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., Schiele, B.: Movie description. In: IJCV (2016)

[257] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. CVPR (2022)

[258] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)

[259] Roy, A., Guinaudeau, C., Bredin, H., Barras, C.: Tvd: a reproducible and multiply aligned tv series dataset. In: LREC 2014, 9th Language Resources and Evaluation Conference (2014)

[260] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023)

[261] Ryokai, K., Durán López, E., Howell, N., Gillick, J., Bamman, D.: Capturing, representing, and interacting with laughter. In: CHI Conference on Human Factors in Computing Systems (2018)

[262] Sadhu, A., Gupta, T., Yatskar, M., Nevatia, R., Kembhavi, A.: Visual semantic role labeling for video understanding. In: CVPR (2021)

[263] Saeed, A., Grangier, D., Zeghidour, N.: Contrastive learning of general-purpose audio representations. In: Proc. ICASSP (2021)

[264] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS (2022)

[265] Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., Metze, F.: How2: A large-scale dataset for multimodal language understanding. In: NeurIPS (2018)

[266] Sarfraz, S., Sharma, V., Stiefelhagen, R.: Efficient parameter-free clustering using first neighbor relations. In: CVPR (2019)

[267] Schneiderman, N., Ironson, G., Siegel, S.D.: Stress and health: psychological, behavioral, and biological determinants. Annual Review of Clinical Psychology (2005)

[268] Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D., Carvalho, S.: Chimpanzee face recognition from videos in the wild using deep learning. Science advances **5**(9) (2019)

[269] Schönfeld, E., Sushko, V., Zhang, D., Gall, J., Schiele, B., Khoreva, A.: You only need adversarial supervision for semantic image synthesis. In: ICLR (2021)

[270] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. NeurIPS (2022)

[271] Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: LAION-400M: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)

[272] Senocak, A., Oh, T.H., Kim, J., Yang, M.H., Kweon, I.S.: Learning to localize sound source in visual scenes. In: CVPR (2018)

[273] Sharghi, A., Laurel, J.S., Gong, B.: Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In: CVPR (2017)

[274] Sharma, A., Hamarneh, G.: Missing mri pulse sequence synthesis using multi-modal generative adversarial network. IEEE Transactions on Medical Imaging (2019)

[275] Sharma, V., Sarfraz, M.S., Stiefelhagen, R.: A simple and effective technique for face clustering in tv series. In: CVPR-W (2017)

[276] Sharma, V., Tapaswi, M., Sarfraz, M.S., Stiefelhagen, R.: Self-supervised learning of face representations for video face clustering. In: Proc. Int. Conf. Autom. Face and Gesture Recog. (2019)

[277] Sharma, V., Tapaswi, M., Sarfraz, M.S., Stiefelhagen, R.: Video face clustering with self-supervised representation learning. IEEE Transactions on Biometrics, Behavior, and Identity Science (2019)

[278] Sharma, V., Tapaswi, M., Sarfraz, M.S., Stiefelhagen, R.: Clustering based contrastive learning for improving face representations. In: Proc. Int. Conf. Autom. Face and Gesture Recog. (2020)

[279] Shehata, M., Ghazal, M., Khalifeh, H.A., Khalil, A., Shalaby, A., Dwyer, A.C., Bakr, A.M., Keynton, R., El-Baz, A.: A deep learning-based cad system for renal allograft assessment: Diffusion, bold, and clinical biomarkers. In: ICIP. IEEE (2020)

[280] Shi, B., Hsu, W.N., Lakhotia, K., Mohamed, A.: Learning audio-visual speech representation by masked multimodal cluster prediction. In: ICLR (2022)

[281] Shimasaki, A., Ueoka, R.: Laugh log: E-textile bellyband interface for laugh logging. In: CHI Conference Extended Abstracts on Human Factors in Computing Systems (2017)

[282] Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Charades-ego: A large-scale dataset of paired third and first person videos. In: arXiv (2018)

[283] Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)

[284] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. Proc. ICML (2015)

[285] Soldan, M., Pardo, A., Alcázar, J.L., Heilbron, F.C., Zhao, C., Giancola, S., Ghanem, B.: Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In: CVPR (2022)

[286] Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Guo, X., Ye, T., Lu, Y., Hwang, J.N., et al.: Moviechat: From dense token to sparse memory for long video understanding. In: CVPR (2023)

[287] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2020)

[288] Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. NeurIPS **32** (2019)

[289] Song, Y., Ermon, S.: Improved techniques for training score-based generative models. NeurIPS (2020)

[290] Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. In: arXiv (2012)

[291] Sowrirajan, H., Yang, J., Ng, A.Y., Rajpurkar, P.: MoCo-CXR: MoCo pretraining improves representation and transferability of chest X-ray models. In: Medical Imaging with Deep Learning, 7-9 July 2021, Lübeck, Germany. Proceedings of Machine Learning Research, vol. 143, pp. 728–744. PMLR (2021)

[292] Srivastava, S., Sharma, G.: Omnivec: Learning robust representations with cross modal sharing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1236–1248 (2024)

[293] Stewart, T.: Overview of motor vehicle traffic crashes in 2021. Tech. rep., National Highway Traffic Safety Administration (2023)

[294] Suthanthiran, M., Strom, T.B.: Renal transplantation. The New England Journal of Medicine **331**(6), 365–376 (1994)

[295] Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C.: 3D self-supervised methods for medical imaging. In: NeurIPS (2020)

[296] Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C.: 3D self-supervised methods for medical imaging. In: NeurIPS (2020)

[297] Tan, R., Plummer, B.A., Saenko, K., Jin, H., Russell, B.: Look at what i'm doing: Self-supervised spatial grounding of narrations in instructional videos. In: NeurIPS (2021)

[298] Tan, Z., Chai, M., Chen, D., Liao, J., Chu, Q., Liu, B., Hua, G., Yu, N.: Diverse semantic image synthesis via probability distribution modeling. In: CVPR (2021)

[299] Tan, Z., Chen, D., Chu, Q., Chai, M., Liao, J., He, M., Yuan, L., Hua, G., Yu, N.: Efficient semantic image synthesis via class-adaptive normalization. In: IEEE TPAMI (2021)

[300] Tang, H., Bai, S., Zhang, L., Torr, P.H., Sebe, N.: Xinggan for person image generation. In: ECCV (2020)

[301] Tang, H., Xu, D., Yan, Y., Torr, P.H., Sebe, N.: Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In: CVPR (2020)

[302] Tapaswi, M., Law, M.T., Fidler, S.: Video face clustering with unknown number of clusters. In: ICCV (2019)

[303] Tapaswi, M., Parkhi, O.M., Rahtu, E., Sommerlade, E., Stiefelhagen, R., Zisserman, A.: Total cluster: A person agnostic clustering method for broadcast videos. In: Proc. ICVGIP (2014)

[304] Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: CVPR (2016)

[305] Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. In: arXiv (2023)

[306] Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., et al.: Gemma: Open models based on gemini research and technology. In: arXiv (2024)

[307] Tepperman, J., Traum, D., Narayanan, S.S.: 'yeah right': Sarcasm recognition for spoken dialogue systems. In: Proc. INTERSPEECH (2006)

[308] Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: ICLR (2023)

[309] Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: ECCV (2018)

[310] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. Proc. ICML (2021)

[311] Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. ICCV (2021)

[312] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

[313] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)

[314] Truffaut, F., Scott, H.: Hitchcock/truffaut. revised edition. Simon and Schuster (1985)

[315] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS (2017)

[316] Vicol, P., Tapaswi, M., Castrejon, L., Fidler, S.: Moviegraphs: Towards understanding human-centric situations from videos. In: CVPR (2018)

[317] Vondrick, C., Pirsiavash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: CVPR (2016)

[318] Wang, A.J., Ge, Y., Yan, R., Ge, Y., Lin, X., Cai, G., Wu, J., Shan, Y., Qie, X., Shou, M.Z.: All in one: Exploring unified video-language pre-training. In: CVPR (2022)

[319] Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G.: Multi-modal learning with missing modality via shared-specific feature modelling. In: CVPR. pp. 15878–15887 (2023)

[320] Wang, H., Ma, C., Zhang, J., Zhang, Y., Avery, J., Hull, L., Carneiro, G.: Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In: MICCAI (2023)

[321] Wang, L., Luc, P., Recasens, A., Alayrac, J.B., Oord, A.v.d.: Multimodal self-supervised learning of general audio representations. arXiv preprint arXiv:2104.12807 (2021)

[322] Wang, T., Zheng, H., Yu, M., Tian, Q., Hu, H.: Event-centric hierarchical representation for dense video captioning. IEEE Transactions on Circuits and Systems for Video Technology (2020)

[323] Wang, T.C., Liu, M.Y., Tao, A., Liu, G., Kautz, J., Catanzaro, B.: Few-shot video-to-video synthesis. In: NeurIPS (2019)

[324] Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: CVPR (2018)

[325] Wang, X., Courant, R., Shi, J., Marchand, E., Christie, M.: JAWS: Just A Wild Shot for cinematic transfer in neural radiance fields. In: CVPR (2023)

[326] Wang, X., Dufour, N., Andreou, N., Cani, M.P., Abrevaya, V.F., Picard, D., Kalogeiton, V.: Analysis of classifier-free guidance weight schedulers. Transactions on Machine Learning Research (TMLR) (2024)

[327] Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., Xing, S., Chen, G., Pan, J., Yu, J., Wang, Y., Wang, L., Qiao, Y.: Internvideo: General video foundation models via generative and discriminative learning. In: arXiv (2022)

[328] Wang, Y., Qi, L., Chen, Y.C., Zhang, X., Jia, J.: Image synthesis via semantic composition. In: ICCV (2021)

[329] Wang, Z., Yuan, Z., Wang, X., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. arXiv preprint arXiv:2312.03641 (2023)

[330] Wei, X., Zhang, T., Li, Y., Zhang, Y., Wu, F.: Multi-modality cross attention network for image and sentence matching. In: CVPR (2020)

[331] Weller, O., Seppi, K.: The rjokes dataset: a large scale humor collection. In: LREC (2020)

[332] Wu, B., Lyu, S., Hu, B.G., Ji, Q.: Simultaneous clustering and tracklet linking for multi-face tracking in videos. In: ICCV (2013)

[333] Wu, B., Zhang, Y., Hu, B.G., Ji, Q.: Constrained clustering and its application to face clustering in videos. In: CVPR (2013)

[334] Wu, C.Y., Krähenbühl, P.: Towards long-form video understanding. In: CVPR (2021)

[335] Wu, R., Zhang, A., Ilyas, I., Rekatsinas, T.: Attention-based learning for missing data imputation in holoclean. In: Conference on Machine Learning and Systems (MLsys). vol. 2, pp. 307–325 (2020)

[336] Xia, Y., Zhang, L., Ravikumar, N., Attar, R., Piechnik, S.K., Neubauer, S., Petersen, S.E., Frangi, A.F.: Recovering from missing data in population imaging – cardiac mr image imputation via conditional generative adversarial nets. Med. Image Anal. **67**, 101812 (2021)

[337] Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa:next phase of question-answering to explaining temporal actions. In: CVPR (2021)

[338] Xiao, S., Tan, M., Xu, D.: Weighted block-sparse low rank representation for face clustering in videos. In: ECCV (2014)

[339] Xiao, Z., Kreis, K., Vahdat, A.: Tackling the generative learning trilemma with denoising diffusion GANs. In: ICLR (2021)

[340] Xie, D., Hu, P., Sun, X., Pirk, S., Zhang, J., Mech, R., Kaufman, A.E.: GAIT: Generating aesthetic indoor tours with deep reinforcement learning. In: ICCV (2023)

[341] Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T.: On layer normalization in the transformer architecture. In: Proc. ICML (2020)

[342] Xiong, Y., Huang, Q., Guo, L., Zhou, H., Zhou, B., Lin, D.: A graph-based framework to bridge movies and synopses. In: ICCV (2019)

[343] Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: ACM Multimedia (2017)

[344] Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., Xu, G., Zhang, J., Huang, S., Huang, F., Zhou, J.: mplug-2: A modularized multi-modal foundation model across text, image and video. In: Proc. ICML (2023)

[345] Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: CVPR (2016)

[346] Xue, H., Sun, Y., Liu, B., Fu, J., Song, R., Li, H., Luo, J.: Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. ICLR (2023)

[347] Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Just ask: Learning to answer questions from millions of narrated videos. In: ICCV (2021)

[348] Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Learning to answer visual questions from web videos. In: ICCV (2022)

[349] Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Zero-shot video question answering via frozen bidirectional language models. In: NeurIPS (2022)

[350] Yang, A., Nagrani, A., Seo, P.H., Miech, A., Pont-Tuset, J., Laptev, I., Sivic, J., Schmid, C.: Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In: CVPR (2023)

[351] Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Li, C., Xu, Y., Chen, H., Tian, J., Qi, Q., Zhang, J., Huang, F.: mplug-owl: Modularization empowers large language models with multimodality. In: arXiv (2023)

[352] Ye, V., Pavlakos, G., Malik, J., Kanazawa, A.: Decoupling human and camera motion from videos in the wild. In: CVPR (2023)

[353] Yoon, J., Jordon, J., van der Schaar, M.: GAIN: Missing data imputation using generative adversarial nets. In: 35th International Conference on Machine Learning (ICML). vol. 80, pp. 5689–5698. PMLR (2018)

[354] Yoon, J., Jordon, J., Schaar, M.: Gain: Missing data imputation using generative adversarial nets. In: Proc. ICML (2018)

[355] Yoon, S., Byun, S., Jung, K.: Multimodal speech emotion recognition using audio and text. In: EEE Spoken Language Technology Workshop (SLT) (2018)

[356] Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: Proc. AAAI (2019)

[357] Zadeh, A., Chan, M., Liang, P.P., Tong, E., Morency, L.P.: Social-iq: A question answering benchmark for artificial social intelligence. In: CVPR (2019)

[358] Zadeh, A., Liang, P.P., Mazumder, N., Poria, S., Cambria, E., Morency, L.P.: Memory fusion network for multi-view sequential learning. In: Proc. AAAI (2018)

[359] Zellers, R., Lu, J., Lu, X., Yu, Y., Zhao, Y., Salehi, M., Kusupati, A., Hessel, J., Farhadi, A., Choi, Y.: Merlot reserve: Neural script knowledge through vision and language and sound. In: CVPR (2022)

[360] Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J.S., Cao, J., Farhadi, A., Choi, Y.: Merlot: Multimodal neural script knowledge models. In: NeurIPS (2021)

[361] Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training (2023)

[362] Zhang, B., Gu, S., Zhang, B., Bao, J., Chen, D., Wen, F., Wang, Y., Guo, B.: Styleswin: Transformer-based gan for high-resolution image generation. ArXiv (2021)

[363] Zhang, C., Lu, T., Islam, M.M., Wang, Z., Yu, S., Bansal, M., Bertasius, G.: A simple llm framework for long-range video question-answering. In: arXiv (2023)

[364] Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: Proc. ICML (2019)

[365] Zhang, J., Li, K., Lai, Y.K., Yang, J.: Pise: Person image synthesis and editing with decoupled gan. In: CVPR (2021)

[366] Zhang, J., Liu, X., Li, K.: Human pose transfer by adaptive hierarchical deformation. In: CGF (2020)

[367] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. ICCV (2023)

[368] Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: MotionDiffuse: Text-driven human motion generation with diffusion model. IEEE TPAMI (2024)

[369] Zhang, S., Gong, Y., Wang, J.: Deep metric learning with improved triplet loss for face clustering in videos. In: Pacific Rim Conference on Multimedia. Springer (2016)

[370] Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. Eprint arXiv:2010.00747 (2020)

[371] Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference (2022)

[372] Zhao, L., Zhang, Z., Chen, T., Metaxas, D., Zhang, H.: Improved transformer for high-resolution gans. In: NeurIPS. vol. 34 (2021)

[373] Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J., Wu, W., Keppo, J., Shou, M.Z.: Motiondirector: Motion customization of text-to-video diffusion models. arXiv preprint arXiv:2310.08465 (2023)

[374] Zheng, G., Li, S., Wang, H., Yao, T., Chen, Y., Ding, S., Li, X.: Entropy-driven sampling and training scheme for conditional diffusion generation. In: ECCV. Springer (2022)

[375] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proc. ICCV (2015)

[376] Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proc. ICCV (2017)

[377] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR (2017)

[378] Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. ACM TOG (2018)

[379] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)

[380] Zhu, P., Abdal, R., Qin, Y., Wonka, P.: Sean: Image synthesis with semantic region-adaptive normalization. In: CVPR (2020)

[381] Zhu, W., Pang, B., Thapliyal, A.V., Wang, W.Y., Soricut, R.: End-to-end dense video captioning as sequence generation. In: ACL (2022)

[382] Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. In: CVPR (2019)

[383] Zhu, Z., Xu, Z., You, A., Bai, X.: Semantically multi-modal image synthesis. In: CVPR (2020)