VMChill: A Dataset for Fine-Grained Visual-Musical Synergy

Xiaowei Chi[1*], **Zeyue Tian**[1*], **Jialiang Chen**[1], **Wei Xue**[1]

[1] Hong Kong University of Science and Technology
xchiaa@connect.ust.hk, ztianad@connect.ust.hk, fjchenjl@ust.hk, weixue@ust.hk

Abstract

Massive multi-modality datasets are fundamental to the success of large video-language models. However, existing datasets often focus on providing textual descriptions for visual content, treating audio, particularly music, as weakly related information. This overlooks the inherent semantic correlation between visual narratives and musical scores, limiting the development of models for fine-grained cross-modal understanding and generation. To address this gap, we introduce VMChill, a large-scale, fine-grained multimodal video dataset. We leverage trailers as our data source, as they are professionally edited to create a strong synergy between visual pacing, scene transitions, and background music for narrative and emotional impact. Our dataset comprises over 20 million video clips derived from more than 27.1k hours of high-resolution trailer videos. To annotate this data, we propose a systematic multimodal captioning framework. This framework first employs specialized unimodal models to extract descriptive features from multiple perspectives, including visual content, motion dynamics, and musical attributes (e.g., genre, instruments, mood). Subsequently, a large language model (LLM) is utilized to adaptively fuse these diverse descriptions into a single, coherent, and rich multimodal caption. This process yields VMChill-2M, a high-quality subset of 2 million clips with detailed multimodal annotations, and VMChill-Test, a manually refined test set for evaluation. We conduct extensive experiments on downstream tasks, including video understanding and generation, to establish benchmarks and demonstrate the dataset's quality. The results validate that VMChill effectively enhances model performance, highlighting its potential to facilitate future research in fine-grained multimodal learning. We will release the dataset, annotation codebase, and processing pipelines to support community research.

Introduction

AI-driven movies and short video production have a wide range of applications in people's daily lives. Clearly, creating vivid videos requires more than just visual frame generation or individual modality-based ones. Thanks to various large-scale video-language datasets, numerous generative multimodal large language models have been developed to achieve this goal (Blattmann et al. 2023; Long et al. 2024; Chen et al. 2023a; He et al. 2023; Lin et al. 2023; Kondratyuk et al. 2023; Wang et al. 2023a; Henschel et al. 2024). However, the existing video-language datasets (Chen et al. 2024; Miech et al. 2019; Wang et al. 2023c) typically focus on visual-based text descriptions, and they overlook the significance of the inherent visual-audio dependencies, especially the semantic alignment between music and visual narratives. It presents a complex challenge that demands cohesive integration of multiple modalities, yet remains largely unexplored.

Collecting high-quality multi-modality source data that preserves consistency between different modalities is challenging. Unlike previous datasets that only provide visual frame-based caption (Bain et al. 2021), multimodal datasets contain complex data formats (*e.g.*music), resulting in more labor-intensive and time-consuming costs in data processing and annotation. Music, as a critical component of trailers, requires specialized annotation frameworks to capture its stylistic and instrumental attributes while maintaining sync with visual content. Moreover, achieving a high correlation between audio and visual content presents challenges.

Targeting to fill the dataset gap by creating a comprehensive and accurate multimodal visual-audio dataset, we first notice trailers. As a precursor to a full-length work, the video trailer has emerged as a vital tool for artists to showcase and disseminate their creations. These short videos typically combine the most compelling visual shots with carefully selected music, have high cross-modality consistency, and hold significant potential in broader multimodal research. The topics are diverse, and the content characters are of various types, *e.g.*film, comedy, and gaming, as shown in Fig. 3. Significantly, the trailer format represents a unique, high-quality, video-centric multimodal data source that benefits further multimodal research exploration and analysis.

In this work, we propose **VMChill**, which aims to unlock the potential of multimodal content understanding and generation for innovative applications in video content generation. We first recognize the immense value of trailers as a video-centric dataset, especially considering the music alongside the videos. **VMChill** contains 20M+ video clips from 290k trailer videos encompassing various source categories as shown in Fig. 3. To ensure the quality of our dataset, we have carefully designed a robust data filtering and cleaning methodology. We also extract a music-enriched subset with our specifically designed music extraction pipeline. Extensive

[*]These authors contributed equally.

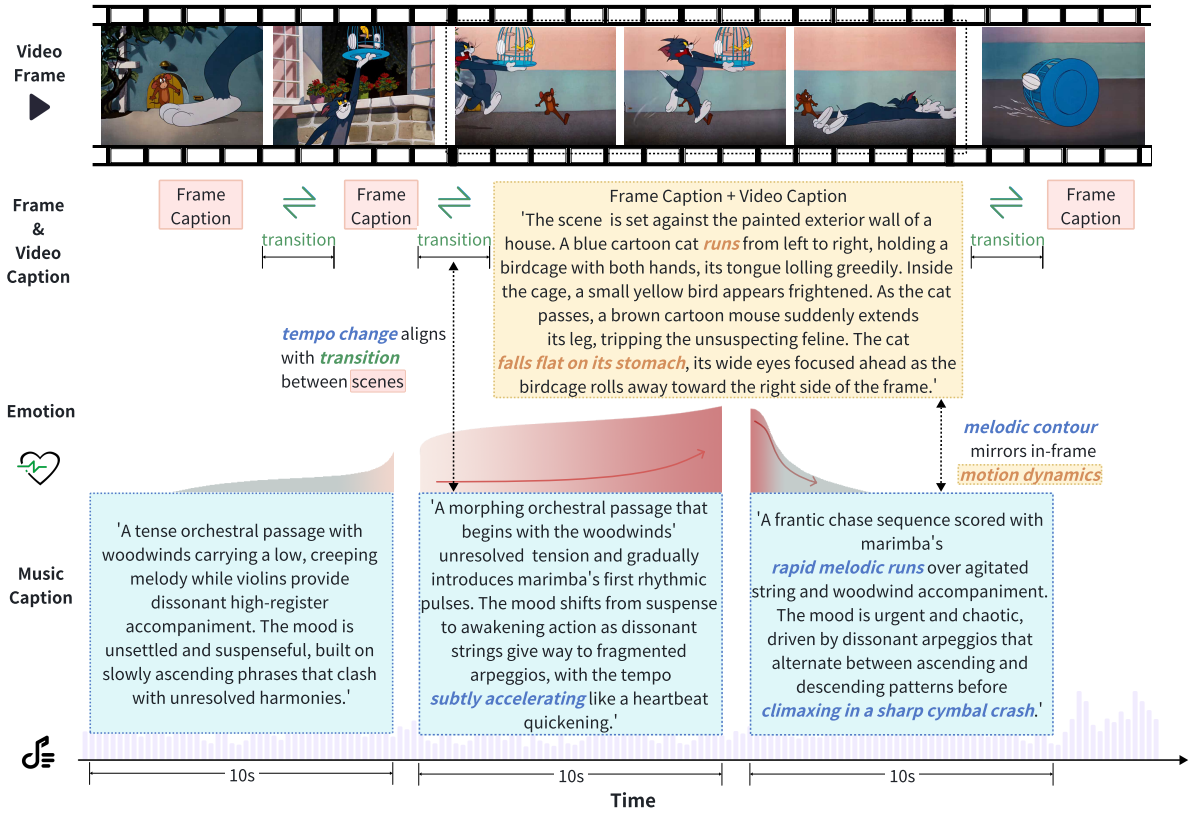

Figure 1: We present VMChill, a large dataset with multimodal captions. Visual inputs are segmented by scene transitions and annotated with both keyframe content and dynamic motion features, while audio segments are analyzed across musical facets to produce structured music descriptors. High-quality clips then fuse aligned visual and musical annotations into comprehensive multimodal captions.

statistical results are provided to demonstrate the diversity and complexity of our dataset.

To address the multimodal-to-language annotation challenge, we have designed a multimodal captioning pipeline incorporating diverse state-of-the-art (SOTA) captioning models (Doh et al. 2023; Yu et al. 2022; Liu et al. 2024). Furthermore, we propose a language model fusion strategy to generate fine-grained multimodal captions. We have performed small-scale annotations on the entire dataset, created a multimodal annotation subset of 3 million samples **VMChill-2M**, and provided a testing set **VMChill-Test** with manually-refined multimodal caption. For the music-enriched subset **VMChill-Music**, we perform detailed music annotating on music clips of fixed duration.

We present evaluation metrics and benchmark results on our dataset, demonstrating the high quality of our annotations and their effectiveness for model training. Through extensive experiments and benchmarking, we showcase the difficulty and diversity of our dataset. We also conduct human evaluations to validate the quality of our multimodal captioning pipeline. Furthermore, we fine-tune understanding models (Zhang, Li, and Bing 2023) and generative models (Chen et al. 2023a) on a subset of our dataset, providing evidence of its high quality and efficacy. Additionally, we evaluate video understanding models on the VMChill-Test to highlight the challenges posed by our dataset, and video-music-based models on the VMChill-Music to demonstrate the effectiveness of cross-modality tasks.

Related Work

Video Understanding and Generation

Video understanding and text-to-video generation are inherently connected tasks. In recent years, there has been remarkable progress in understanding models (Bai et al. 2025; Zhang, Li, and Bing 2023; Chen et al. 2023b), which have greatly contributed to the advancement of text-based video generation techniques. The availability of large-scale datasets and diffusion models has revolutionized video generation, from early pixel-level methods (Hong et al. 2022) to recent latent diffusion frameworks (Wan et al. 2025; ai et al. 2025; Yang et al. 2024). Concurrently, understanding models have also witnessed significant improvements. A series of MLLM-based understanding models (Wang et al. 2025; Liu et al. 2024; Yang et al. 2025; Jin et al. 2023) has reach a satisfied understanding ability. The iterative interaction between video understanding and generation has led to the development of excellent large-scale datasets and models encompassing

Dataset	Year	Size	Caption	Modality	Clips	E(V)	E(T)	Resolution
WebVid (Bain et al. 2021)	2021	52khr	Alt-text	Video	10M	10.0s	-	360p
Panda (Chen et al. 2024)	2024	167khr	Auto	Video	70M	8.5s	13.2	720p
HD-VILA (Xue et al. 2021)	2022	371.5khr	ASR	Video	100M	3.6s	32.5	720p
MSR-VTT (Xu et al. 2016)	2016	40hr	Manual	Video	10K	15.0s	9.3	240p
InternVid (Wang et al. 2023c)	2023	760.3khr	Auto	MM	100M	11.7s	11.6	720p
HowTo100M (Miech et al. 2019)	2023	134.5khr	ASR	MM	136M	3.6s	4.0	720p
HarmonySet (Zhou et al. 2025)	2024	458.8hr	Auto	MM	48,328	-	-	720p
VMChill-20M	2025	27.1khr	Auto	Video	20M	4.6s	10.7	720p
VMChill-Music	2025	15.3khr	Auto	MM	5M	10.0s	25.0	720p
VMChill-2M	2025	8.2khr	Auto	MM	2M	13.8s	39.4	720p
VMChill-Test	2025	3.2hr	Manual	MM	1k	11.6s	98.2	720p

Table 1: Comparison of VMChill-X and other Video to language datasets. VMChill-X contains four sets(20M, Music, 2M, Test) with 720p resolution.

diverse approaches. Panda (Chen et al. 2024) introduced an auto-caption model distilled from video understanding models like (Li et al. 2024b; Yang et al. 2025; Zhang, Li, and Bing 2023; Zhu et al. 2023a), as well as multimodal generation models (Chen et al. 2025; Wu et al. 2024; Xiang et al. 2024; Chi et al. 2023; Ma et al. 2024).

Video-Language Datasets

Captioned video datasets are essential for text-to-video understanding and generation tasks. MSR-VTT (Xu et al. 2016), UCF-101 (Soomro, Zamir, and Shah 2012) are commonly used as evaluation sets. WebVid (Bain et al. 2021), VideoFactory (Wang et al. 2023b), and other works (Sanabria et al. 2018; Wang et al. 2019; Stroud et al. 2020; Nagrani et al. 2022) contain multilingual video descriptions, video clips, and basic metadata, and are used for tasks like video understanding, text-to-video retrieval, and audio-video captioning with weak annotations. Several datasets and approaches have utilized audio to enhance video captioning (Miech et al. 2019; Rohrbach et al. 2016; Zellers et al. 2021; Wang et al. 2023c; Chen et al. 2024; Xue et al. 2021), providing either transcribed audio descriptions, narrations, ASR transcriptions, or multiple captions generated through an auto caption model. Recently, (Zhou et al. 2025) noticed the alignment between audio and video.

Large-scale video-audio collections like AudioSet (Gemmeke et al. 2017) and VGGSound (Chen et al. 2020) provide even broader audio-visual pairings. However, these audio-focused works treat music as a generic category without temporal or instrumental annotations. While pioneering in scale, such datasets fail to capture the fine-grained alignment between music and visual dynamics. Music-specific efforts like SymMV (Zhuo et al. 2023) and BGM909 (Li et al. 2024a) address video-music alignment along with structured descriptors (*e.g.*chords, genres) and full-track consistency annotation. However, their whole-video-level annotation scheme with unsegmented, lengthy clips combined with a limited dataset scales makes them impractical for training frame-accurate tasks. This leaves video-language models unable to leverage music as a semantically rich modality—a critical shortcoming for content like trailers, where music drives narrative impact.

Cross-Modal Music-Video Systems

Current research on music-aware video systems reveals critical gaps in both understanding and generation tasks. For video understanding, Music-AVQA (Li et al. 2022) establishes instrument-centric question answering, while Sound2Synth (Chen et al. 2022) specializes in timbre recognition, yet both remain limited to narrow musical domains. AVFormer (Seo, Nagrani, and Schmid 2023) demonstrates improved joint audio-visual learning but suffers from coarse musical annotations. In the generation domain, while text-to-music models like MusicGen (Copet et al. 2024) and MusicLM (Agostinelli et al. 2023) produce high-quality audio, their lack of visual conditioning limits video applications. A series of work including V2Meow (Su et al. 2024), VidMuse (Tian et al. 2025b), and AudioX (Tian et al. 2025a) features direct video-to-music synthesis through cross-modal attention, yet remain constrained by the data scarcity issues and struggle with long-term musical structure preservation.

Collectively, these approaches demonstrate the potential of cross-modal music-video systems, yet their effectiveness is fundamentally hindered by the absence of large-scale datasets with fine-grained, temporally aligned music-video annotations - a gap our work aims to address.

VMChill Dataset

To boost the performance of multimodal generation tasks, we construct a high-quality multimodal dataset based on trailers, which offers a wealth of multimodal information and diverse categories that distinguish it from existing large-scale video-language datasets.

Data Collection Pipeline

Trailers are uniquely crafted videos with distinct data distributions and high quality. They showcase diverse themes (*e.g.*movies, TV shows, games) through compelling clips, often paired with background music and narration. This dense, rich content poses a significant challenge for AI-generated content (AIGC) tasks and complicates data cleaning. To solve

Figure 2: Word cloud of the (left) objects and (right) background in VMChill. Most of the objects are human, and most of the backgrounds are indoor scenes like office, kitchen, etc.

this challenging problem, we design a comprehensive data collection and cleaning process, as shown in Fig. 4, to deal with such complex videos, as described in this section.

Collection Strategies To enrich the diversity of our VMChill, we first employ the keyword "trailer" to retrieve various in-the-wild trailer videos that encapsulate a wide range of artistic works and genres. Then, we tailor the keywords (*e.g.*"Movie Trailers", "Video Game Trailers", "TV Show Trailers", "Documentary Trailers") more specifically to explicitly collect trailer videos with divergent sources. Through these two methods, we collect 285,518 comprehensive trailer videos with a total duration of 94,911,802.8 seconds.

Trimming To facilitate the extraction of various video information in subsequent analyses, we cut the original videos into clips based on the scenes. As the most mature and practical tool currently available, PySceneDetect offers robust functionalities for this purpose. Thus, we use the ContentDetector of PySceneDetect to compare the difference in content between adjacent frames and then cut the videos according to the predetermined threshold of 30. We finally get 21,588,792 clips with an average duration of 4.6 seconds.

Motion Filtering Evaluating motion quality is a crucial step in selecting high-quality video content. Motion vectors and optical flow are both mainstream motion quality evaluation methods. In our case, trailer videos often have rapid cuts and transitions between scenes, posing challenges for optical flow-based analysis. Other than optical flow, motion vectors are more robust to these quick changes, as they rely on larger block units' displacement rather than individual pixels' continuous flow. On the other hand, motion vectors are more lightweight and can be calculated more efficiently. Thus, we leverage motion vectors to filter out clips with problems like static frames, title sequences, and slideshow-like playback.

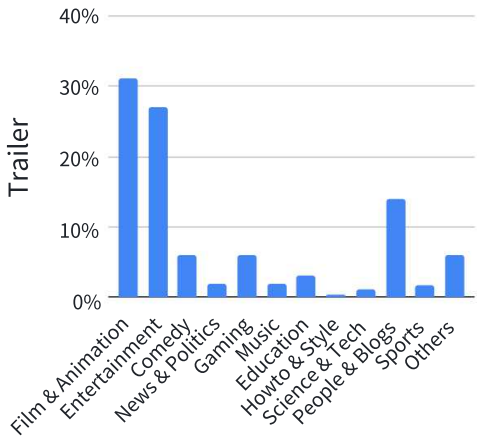

Figure 3: Distribution of video categories of the VMChill dataset.

Diversity We evaluate the diversity and richness of our dataset from three aspects: theme, objects, and backgrounds. While collecting, we first assess the categories from the Yotoube metadata provided by the video provider, as shown in Fig. 3. Furthermore, we generate an object-level caption list and background by Llava-NexT (Liu et al. 2024) for a more accurate category-based generation. The word cloud of objects and backgrounds is shown in Fig. 2.

OCR Trailer videos often have text-heavy sections with high-quality text animations, like opening and ending credits. To identify these text-rich segments, we utilize OCR to detect the text content in the video frames and calculate the bounding box area of the text. This measurement reflects the amount of text in the clips.

Quality Statistics In addition to text detection, we consider image quality (Huang et al. 2024) and aesthetic scores (Schuhmann et al. 2021) to enhance our analysis of trailer videos. These measures allow us to evaluate frames' visual fidelity, clarity, and aesthetic appeal, providing more comprehensive insights for trailer analysis and editing.

Audio and Music Extraction We extract the audio from the video source with a sampling rate of 44.1 kHz. We use PANNs (Kong et al. 2020) to perform music event detection, and over 70% of the audio segments contain music. To isolate musical components, we employ Demucs for source separation, extracting discrete stems. This separation enables precise music analysis by removing speech interference (*e.g.*, dialogue/narration), with 89% of processed clips achieving <3dB vocal suppression. The purified music stems are stored alongside original audio for multimodal comparison.

Video Captioning Pipeline

The VMChill contains many complex themes, like subtitles and character animations, as shown in Fig. 3, which brings extraordinary complex work for video captioning. At the same time, smooth transition shots also make it impossible for traditional single-frame annotation methods to convey

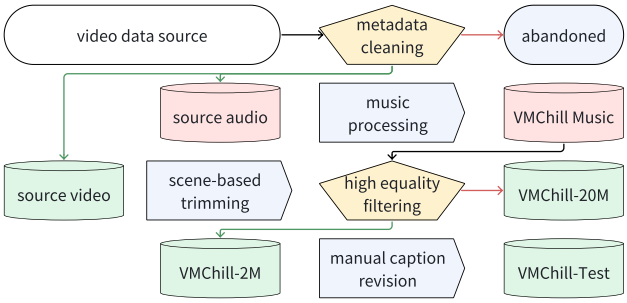

Figure 4: Data cleaning and separation of the VMChill. We first clean metadata and extract synchronized audio-visual streams. Scene-cut segmentation generates the full VMChill-20M collection, which undergoes quality control to yield the refined VMChill-2M subset. Parallelly, audio tracks are processed through our music pipeline to produce VMChill-Music.

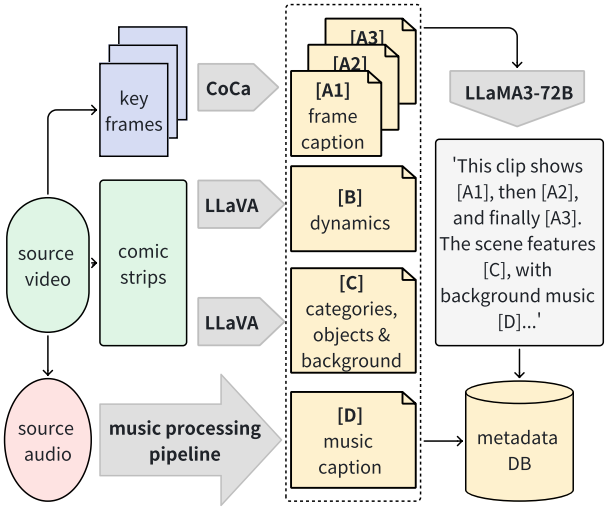

Figure 5: Data captioning pipeline. Starting from video clips, we extract frames and audio, and then perform multiple rounds of captioning. A predefined instruction format combines multimodal captions, which serve as prompts for the language model and generate the final merged prompts.

semantics coherently. Therefore, this section introduces a multi-temporal and multimodal caption pipeline containing a detailed video description from frame, motion, and music levels.

Frame Caption The auto-captioning pipeline has proven efficient in cutting-edge video generation foundation models. SVD (Blattmann et al. 2023) and Pandas (Chen et al. 2024) have given promising results and demonstrated the importance of high-quality frame captions for the generation model. We initially perform image-level captioning on the individual frames of the data. We employ CoCa (Yu et al. 2022) for each video clip to generate separate captions for three frames(first, middle, and last), resulting in relevant captions.

Clip Caption We use concise captions for three frames that capture the essential information. We aim to obtain fine-grained captions and variations between frames in the video. We concatenate multiple frames into a comic strip format and employ the LLaVA (Liu et al. 2024) image model to guide the description of the dynamic differences between frames. Additionally, leveraging a powerful multimodal language model, we incorporate OCR and more detailed summary descriptions to expand the information within the frame captions.

Categories and Background Noticing the LLM-based caption has hallucinations when describing the frame, we further generate word-level labels to enhance the annotation of the main objects and background. Initially, we utilize LLaVA's QA capabilities to have the model answer questions about the background. Subsequently, through QA, we prompt the model to provide relevant category information. We perform the word cloud in Fig. 2, and certify the quality of the captions by subjective experience.

Music-Aware Filtering and Detailed Annotation To construct a music-enriched subset with high-quality annotations, we implement a two-stage pipeline. First, we process music quality filtering: For all clips, we employ Qwen2-Audio (Chu et al. 2024) to evaluate two metrics: sound quality (*e.g.*clarity, noise level) and musicality (*e.g.*harmonic richness, rhythmic consistency). Then for the clips with higher quality and scores above the threshold in both metrics, Qwen2-Audio generates structured captions covering genre (*e.g.*electronic rock, ambient), instruments (*e.g.*distorted guitar, piano), mood (*e.g.*energetic, melancholic), tempo (*e.g.*120 BPM, moderate), and music dynamics (*e.g.*ascending, crescendo).

OCR The trailer contains many text animations. Captioning the context inside the movie can also be a challenging task. In this task, we utilize the OCR ability by LLaVA (Liu et al. 2024) and caption the context for 5 frames of each video. We merge reliable text animation captions based on the previous method.

General caption Combining all the captions mentioned above, we use the language model LLaMA2-13B (Touvron et al. 2023) to merge all the captions and generate complete and high-quality multimodal captions. We evaluate the caption accuracy and quality by human preference.

SubSet Separation

We annotate all video clips in VMChill-20M with frame captions, comprising over 20 million clips. As shown in Tab.1, our dataset is comparable in scale to existing benchmarks. All videos maintain 720p or higher resolution, with an average clip duration of 4.6 seconds.

Music-Enriched Subset contains samples selected from VMChill-20M based on sound quality and musicality scores. These segments are annotated with the aforementioned four musical dimensions (genre, instruments, mood, tempo), ensuring to retain both high acoustic fidelity and rich musical semantics critical for tasks like video-to-music generation. Examples include film trailers with orchestral scores (high musicality) and game trailers with electronic beats (high tempo diversity).

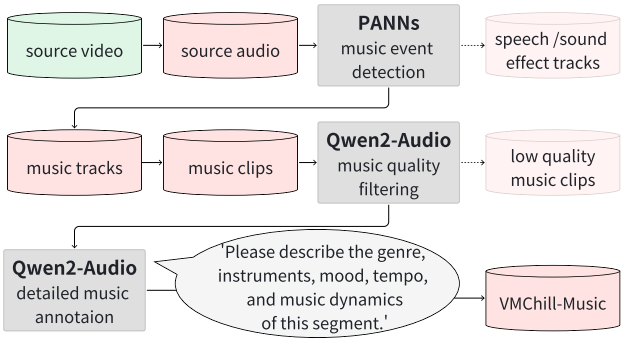

Figure 6: Music processing pipeline of VMChill. Music segments are extracted from raw audio through vocal separation and segmented into fixed-duration clips. Each clip undergoes assessment, where only segments meeting sound quality and musicality thresholds proceed to multi-dimensional annotation. High-quality musical descriptors are then temporally aligned with visual content for multimodal caption generation.

Dataset	Year	Dur./Clip	#Clips	#Hours
Audioset	2017	10s	2M	5.8khr
Vggsound	2020	10s	210K	550hr
VMChill-Music	2025	10.0s	5M	15.3khr
VMChill-2M	2025	13.8s	2M	8.2khr
VMChill-Test	2025	11.6s	1k	3.2hr

Table 2: Comparison of VMChill-Music, VMChill-2M, VMChill-Test, and other Video-Audio Generation Datasets. For each dataset, we list the following information in each column: dataset name (Dataset), public year (Year), average duration per clip (Dur./Clip), total number of clips (#Clips), total number of hours (#Hours).

High-Quality Subset , named VMChill-2M, contains a detailed multimodal caption. From the original set, we sample VMChill-2M using the following criteria: 1. We filter out clips with motion scores below 0.45 or above 50. 2. We only retain the clips within the top 85% of image quality scores. 3. We only keep the clips within the top 85% of aesthetic scores. All clips in VMChill-2M are longer than 4s and are provided with all captions, as shown in Fig. 5.

Manually Refined Test Set is extracted from the VMChill-2M, we extract a fine-branded testing set that contains 1k video clips and multiple multimodal captions. Then, we revise the merged caption manually to build a testing subset with trust-wise multimodal prompts. We test several tasks and models on the test set to show the complexity and difficulties of VMChill. The test set has 98.2 words of caption on average and includes 3.2hr video clips.

Experiments

This section presents comprehensive experiments on multiple tasks to demonstrate the effectiveness, diversity, complexity, and difficulty of VMChill.

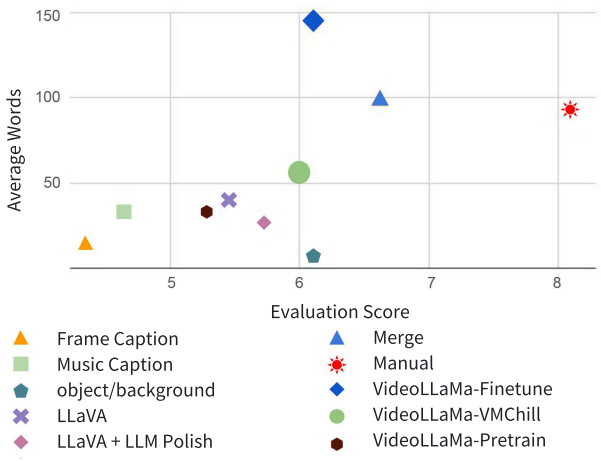

Figure 7: Human evaluation results of the captioning models on the VMChill-Test. X-axis is the average evaluation score from 0-10, and the Y-axis is the average number of words.

Multimodal Captioning

We present the results of our human evaluation of the quality of video caption in Fig. 7. Ten videos are randomly selected from the VMChill-Test dataset. They are rated on a scale of 0 to 10 based on general impressions, including aspects such as correctness, level of detail, richness, and fluency. The results of more than 100 sets of samples indicate that the manually adjusted prompts rating of 8.12 outperforms the auto-caption pipeline, while our merged captions achieve the second-best performance of 6.62. Despite being short and straightforward, object/background labels achieve a 6.02 evaluation score, demonstrating more correctness than other captions. The frame caption, music caption, and LLaVA caption obtain 4.3, 4.6, and 5.4, respectively, and these findings demonstrate the effectiveness of our captions and highlight the quality of our labeled captions by human annotators.

Video Generation

The model is fine-tuned on the VideoCrafter-2.0 (Long et al. 2024) dataset using 8 Tesla-H800 GPUs with a batch size of 3 for 10,000 steps at a learning rate of 6e-6. The training data is randomly sampled from the VMChill-2M dataset, using the video captions as input. The evaluation results, shown in Tab.3, include 9 metrics on VBench (Huang et al. 2024), indicating that fine-tuning the model on the VMChill-2M dataset leads to improvements of 0.6 in motion smoothness and 1.77 in subject consistency, with a slight overall performance boost(0.12 higher) compared with the official VideoCrafter-2.0 checkpoint. Visual examples of the generated content are provided in Appendix, and additional demonstrations and experiment details will be included in the supplementary material. This thorough evaluation and comparison of the tuned model's performance on critical metrics provides valuable insights into the effectiveness of the fine-tuning process and

Dimentions(↑)	VC-2.0	**VC-2.0(VMChill)**
temporal style	25.84	24.61
appearance style	25.13	24.10
image quality	67.22	69.78
dynamic degree	42.50	43.50
motion smoothness	97.73	98.33
temporal flickering	98.41	98.50
Subject consistency	96.85	98.62
background consistency	98.22	98.40
Overall consistency	28.23	25.33
Sum	64.45	**64.57**

Table 3: Comparison of VideoCrafter-2.0 and VideoCrafter-2.0(VMChill) on 9 different dimensions. For every dimension, a higher score is better.

Task	Method	KL↓	ISc↑	FD↓	FAD↓
Text-to-Music	Stable-Audio-Open	1.51	2.04	36.33	3.23
	Stable-Audio-Open (tuned)	1.20	3.19	24.82	2.40
	T2M-DiT	1.50	2.98	25.74	3.46
Video-to-Music	VidMuse	0.73	1.32	29.95	2.46
	V2M-DiT	0.77	1.25	30.52	2.51

Table 4: Comparative results of music generation performance across our DiT model trained on VMChill-2M and baseline methods on text-to-music and video-to-music generation tasks.

the potential benefits of leveraging the VMChill-2M dataset for video generation tasks.

Video Understanding

Experiment Setting To evaluate the capability of our dataset in multimodal video understanding, we choose Video-LLaMA (Zhang, Li, and Bing 2023) as the baseline. We use the same model and training config as Video-LLaMA, which use Vicuna-v0-7B as LLaMA model (Zheng et al. 2023), ViT (Dosovitskiy et al. 2021), and Q-Former (Zhang et al. 2023) as the video encoder and the linear projection layer from MiniGPT-4 (Zhu et al. 2023b). We train 4 epochs on VMChill-2M, each containing 2500 iters with batch size 32. We compare it with two official model weights: the pre-train Video-LLaMA weight on WebVid (2.5M video-caption pairs) and the fine-tuned Video-LLaMA.

Evaluation Metric We evaluate video understanding models on VMChill-Test. We choose the commonly used metrics in text generation tasks-BLEU-4 (Papineni et al. 2002), ROGUE-L (Lin and Och 2004), METEOR (Banerjee and Lavie 2005), and CIDEr (Vedantam, Zitnick, and Parikh 2015) to evaluate our result. All the metrics are computed using the pycocoevalcap (Lin et al. 2015) package. We also use BERTScore (Zhang et al. 2020) to evaluate the contextual similarity for each token in the ground truth and the predicted captions. The results are reported in Tab.5. The official weights' underperformance highlights VMChill's challenges and its distribution shift from the original training data.

Model	BLEU-4↑	M↑	ROGUE-L↑	CIDEr↑	BERT↑
Video-LLaMA(Pretrain)	0.52	4.57	11.57	0.09	84.42
Video-LLaMA(Finetuned)	3.94	14.05	22.67	2.45	85.48
Video-LLaMA(VMChill)	5.59	13.83	24.97	24.79	87.21

Table 5: Comparison of Video-LLaMA model performance on the Trailer-Test dataset. The figure shows the results of three different versions of the Video-LLaMA model across five evaluation metrics, and the Video-LLaMA(VMChill) version performs better on most evaluation indicators.

In addition, we also evaluate three checkpoints from Video-LLaMA (Zhang, Li, and Bing 2023) by human evaluation in Fig. 7 and found that the Video-LLaMA-VMChill evaluation result slightly lags behind Video-LLaMA-Finetune but performs significantly better than Video-LLaMA-Pretrain.

Music Generation

To validate the quality and effectiveness of our dataset's video and music captions and video-music pairs, we train a DiT model capable of generating high-quality music from either text (T2M-DiT) or video (V2M-DiT) inputs. Comparative evaluations against baseline text-to-music model Stable-Audio-Open (Evans et al. 2024) and video-to-music model VidMuse on corresponding benchmarks MusicCaps and V2M-Bench are conducted using four key metrics: Kullback-Leibler Divergence (KL), Inception score (ISc), Frechet distance (FD), and Frechet Audio Distance (FAD) (Kilgour et al. 2018). While our model outperforms Stable-Audio-Open, the latter achieves superior results when fine-tuned on our dataset. For video-to-music tasks, our model show competitive performance against the in-domain strong baseline VidMuse, demonstrating the effectiveness of our dataset.

Conclusion

We introduce VMChill, a comprehensive and accurate multimodality visual-audio dataset to address the dataset gap. By utilizing the inherent value of trailers, which integrate visual, audio, and contextual elements, VMChill offers detailed and precise multi-modality annotations. Our systematic captioning framework adaptively merges visual and musical perspectives, ensuring that the annotations capture the richness of multimodal content. Experimental results demonstrate the high quality of the VMChill dataset, its effectiveness for fine-grained multimodal-language model training, and a variety of downstream applications. We believe this innovative dataset will unlock new possibilities in video content generation and significantly advance research in visual-audio understanding. The comprehensive and diverse nature of VMChill makes it a valuable asset for the research community, paving the way for novel applications that leverage the power of multimodal learning.

Acknowledgements

The research was supported by Theme-based Research Scheme(T45- 205/21-N) from Hong Kong RGC, NSFC (No. 62206234), and Generative AI Research and Development Centre from InnoHK. Correspondence to: Wei Xue <weixue@ust.hk>.

References

Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; Sharifi, M.; Zeghidour, N.; and Frank, C. 2023. MusicLM: Generating Music From Text. arXiv:2301.11325.

ai, S.; Teng, H.; Jia, H.; Sun, L.; Li, L.; Li, M.; Tang, M.; Han, S.; Zhang, T.; Zhang, W. Q.; Luo, W.; Kang, X.; Sun, Y.; Cao, Y.; Huang, Y.; Lin, Y.; Fang, Y.; Tao, Z.; Zhang, Z.; Wang, Z.; Liu, Z.; Shi, D.; Su, G.; Sun, H.; Pan, H.; Wang, J.; Sheng, J.; Cui, M.; Hu, M.; Yan, M.; Yin, S.; Zhang, S.; Liu, T.; Yin, X.; Yang, X.; Song, X.; Hu, X.; Zhang, Y.; and Li, Y. 2025. MAGI-1: Autoregressive Video Generation at Scale. arXiv:2505.13211.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1708–1718.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Goldstein, J.; Lavie, A.; Lin, C.-Y.; and Voss, C., eds., *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics.

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

Chen, H.; Xia, M.; He, Y.; Zhang, Y.; Cun, X.; Yang, S.; Xing, J.; Liu, Y.; Chen, Q.; Wang, X.; et al. 2023a. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 721–725. IEEE.

Chen, J.; Zhu, D.; Haydarov, K.; Li, X.; and Elhoseiny, M. 2023b. Video chatcaptioner: Towards the enriched spatiotemporal descriptions. *arXiv preprint arXiv:2304.04227*.

Chen, T.-S.; Siarohin, A.; Menapace, W.; Deyneka, E.; Chao, H.-w.; Jeon, B. E.; Fang, Y.; Lee, H.-Y.; Ren, J.; Yang, M.-H.; et al. 2024. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. *arXiv preprint arXiv:2402.19479*.

Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. *arXiv preprint arXiv:2501.17811*.

Chen, Z.; Jing, Y.; Yuan, S.; Xu, Y.; Wu, J.; and Zhao, H. 2022. Sound2Synth: Interpreting Sound via FM Synthesizer Parameters Estimation. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 4921–4928. International Joint Conferences on Artificial Intelligence Organization. AI and Arts.

Chi, X.; Liu, Y.; Jiang, Z.; Zhang, R.; Lin, Z.; Zhang, R.; Gao, P.; Fu, C.; Zhang, S.; Liu, Q.; et al. 2023. Chatillusion: Efficient-aligning interleaved generation ability with visual instruction model. *arXiv preprint arXiv:2311.17963*.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; Zhou, C.; and Zhou, J. 2024. Qwen2-Audio Technical Report. arXiv:2407.10759.

Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2024. Simple and Controllable Music Generation. arXiv:2306.05284.

Doh, S.; Choi, K.; Lee, J.; and Nam, J. 2023. LP-MusicCaps: LLM-Based Pseudo Music Captioning. *arXiv preprint arXiv:2307.16372*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Evans, Z.; Parker, J. D.; Carr, C.; Zukowski, Z.; Taylor, J.; and Pons, J. 2024. Stable Audio Open. arXiv:2407.14358.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 776–780. IEEE.

He, Y.; Xia, M.; Chen, H.; Cun, X.; Gong, Y.; Xing, J.; Zhang, Y.; Wang, X.; Weng, C.; Shan, Y.; et al. 2023. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*.

Henschel, R.; Khachatryan, L.; Hayrapetyan, D.; Poghosyan, H.; Tadevosyan, V.; Wang, Z.; Navasardyan, S.; and Shi, H. 2024. StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Text. *arXiv preprint arXiv:2403.14773*.

Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2022. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. *arXiv preprint arXiv:2205.15868*.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; Wang, Y.; Chen, X.; Wang, L.; Lin, D.; Qiao, Y.; and Liu, Z. 2024. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jin, P.; Takanobu, R.; Zhang, C.; Cao, X.; and Yuan, L. 2023. Chat-UniVi: Unified Visual Representation Empowers Large

Language Models with Image and Video Understanding. *arXiv preprint arXiv:2311.08046*.

Kilgour, K.; Zuluaga, M.; Roblek, D.; and Sharifi, M. 2018. Fr\'echet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms. *arXiv preprint arXiv:1812.08466*.

Kondratyuk, D.; Yu, L.; Gu, X.; Lezama, J.; Huang, J.; Hornung, R.; Adam, H.; Akbari, H.; Alon, Y.; Birodkar, V.; et al. 2023. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*.

Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; and Plumbley, M. D. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2880–2894.

Li, G.; Wei, Y.; Tian, Y.; Xu, C.; Wen, J.-R.; and Hu, D. 2022. Learning to Answer Questions in Dynamic Audio-Visual Scenarios. arXiv:2203.14072.

Li, S.; Qin, Y.; Zheng, M.; Jin, X.; and Liu, Y. 2024a. Diff-BGM: A Diffusion Model for Video Background Music Generation. arXiv:2405.11913.

Li, X.; Wang, Y.; Yu, J.; Zeng, X.; Zhu, Y.; Huang, H.; Gao, J.; Li, K.; He, Y.; Wang, C.; et al. 2024b. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*.

Lin, C.-Y.; and Och, F. J. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 605–612. Barcelona, Spain.

Lin, H.; Zala, A.; Cho, J.; and Bansal, M. 2023. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Long, F.; Qiu, Z.; Yao, T.; and Mei, T. 2024. VideoDrafter: Content-Consistent Multi-Scene Video Generation with LLM. *arXiv preprint arXiv:2401.01256*.

Ma, Y.; Liu, X.; Chen, X.; Liu, W.; Wu, C.; Wu, Z.; Pan, Z.; Xie, Z.; Zhang, H.; yu, X.; Zhao, L.; Wang, Y.; Liu, J.; and Ruan, C. 2024. JanusFlow: Harmonizing Autoregression and Rectified Flow for Unified Multimodal Understanding and Generation.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2630–2640.

Nagrani, A.; Seo, P. H.; Seybold, B.; Hauth, A.; Manén, S.; Sun, C.; and Schmid, C. 2022. Learning Audio-Video Modalities from Image Captions. In *European Conference on Computer Vision*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Rohrbach, A.; Torabi, A.; Rohrbach, M.; Tandon, N.; Pal, C. J.; Larochelle, H.; Courville, A. C.; and Schiele, B. 2016. Movie Description. *International Journal of Computer Vision*, 123: 94 – 120.

Sanabria, R.; Caglayan, O.; Palaskar, S.; Elliott, D.; Barrault, L.; Specia, L.; and Metze, F. 2018. How2: A Large-scale Dataset for Multimodal Language Understanding. *ArXiv*, abs/1811.00347.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. arXiv:2111.02114.

Seo, P. H.; Nagrani, A.; and Schmid, C. 2023. AVFormer: Injecting Vision into Frozen Speech Models for Zero-Shot AV-ASR. arXiv:2303.16501.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Stroud, J. C.; Ross, D. A.; Sun, C.; Deng, J.; Sukthankar, R.; and Schmid, C. 2020. Learning Video Representations from Textual Web Supervision. *ArXiv*, abs/2007.14937.

Su, K.; Li, J. Y.; Huang, Q.; Kuzmin, D.; Lee, J.; Donahue, C.; Sha, F.; Jansen, A.; Wang, Y.; Verzetti, M.; and Denk, T. I. 2024. V2Meow: Meowing to the Visual Beat via Video-to-Music Generation. arXiv:2305.06594.

Tian, Z.; Jin, Y.; Liu, Z.; Yuan, R.; Tan, X.; Chen, Q.; Xue, W.; and Guo, Y. 2025a. Audiox: Diffusion transformer for anything-to-audio generation. *arXiv preprint arXiv:2503.10522*.

Tian, Z.; Liu, Z.; Yuan, R.; Pan, J.; Liu, Q.; Tan, X.; Chen, Q.; Xue, W.; and Guo, Y. 2025b. VidMuse: A Simple Video-to-Music Generation Framework with Long-Short-Term Modeling. arXiv:2406.04321.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based Image Description Evaluation. arXiv:1411.5726.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; Zeng, J.; Wang, J.; Zhang, J.; Zhou, J.; Wang, J.; Chen, J.; Zhu, K.; Zhao, K.; Yan, K.; Huang, L.; Feng, M.; Zhang, N.; Li, P.; Wu, P.; Chu, R.; Feng, R.; Zhang, S.; Sun, S.; Fang, T.; Wang, T.; Gui, T.; Weng, T.; Shen, T.; Lin, W.; Wang, W.; Wang, W.; Zhou, W.; Wang, W.; Shen, W.; Yu, W.; Shi, X.; Huang, X.; Xu, X.; Kou, Y.; Lv, Y.; Li, Y.; Liu, Y.; Wang, Y.; Zhang, Y.; Huang, Y.; Li, Y.; Wu, Y.; Liu, Y.; Pan, Y.; Zheng, Y.; Hong, Y.; Shi, Y.; Feng, Y.; Jiang, Z.; Han, Z.; Wu, Z.-F.; and Liu, Z. 2025. Wan:

Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*.

Wang, F.-Y.; Chen, W.; Song, G.; Ye, H.-J.; Liu, Y.; and Li, H. 2023a. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*.

Wang, W.; Yang, H.; Tuo, Z.; He, H.; Zhu, J.; Fu, J.; and Liu, J. 2023b. VideoFactory: Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. *ArXiv*, abs/2305.10874.

Wang, X. E.; Wu, J.; Chen, J.; Li, L.; fang Wang, Y.; and Wang, W. Y. 2019. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 4580–4590.

Wang, Y.; He, Y.; Li, Y.; Li, K.; Yu, J.; Ma, X. J.; Chen, X.; Wang, Y.; Luo, P.; Liu, Z.; Wang, Y.; Wang, L.; and Qiao, Y. 2023c. InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation. *ArXiv*, abs/2307.06942.

Wang, Y.; Li, X.; Yan, Z.; He, Y.; Yu, J.; Zeng, X.; Wang, C.; Ma, C.; Huang, H.; Gao, J.; et al. 2025. Internvideo2.5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*.

Wu, C.; Chen, X.; Wu, Z.; Ma, Y.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; Ruan, C.; et al. 2024. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*.

Xiang, J.; Liu, G.; Gu, Y.; Gao, Q.; Ning, Y.; Zha, Y.; Feng, Z.; Tao, T.; Hao, S.; Shi, Y.; Liu, Z.; Xing, E. P.; and Hu, Z. 2024. Pandora: Towards General World Model with Natural Language Actions and Video States. *arXiv preprint arXiv:2406.09455*.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5288–5296.

Xue, H.; Hang, T.; Zeng, Y.; Sun, Y.; Liu, B.; Yang, H.; Fu, J.; and Guo, B. 2021. Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5026–5035.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072*.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. *Trans. Mach. Learn. Res.*, 2022.

Zellers, R.; Lu, X.; Hessel, J.; Yu, Y.; Park, J. S.; Cao, J.; Farhadi, A.; and Choi, Y. 2021. MERLOT: Multimodal Neural Script Knowledge Models. In *Neural Information Processing Systems*.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Zhang, Q.; Zhang, J.; Xu, Y.; and Tao, D. 2023. Vision Transformer with Quadrangle Attention. *arXiv preprint arXiv:2303.15105*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Zhou, Z.; Mei, K.; Lu, Y.; Wang, T.; and Rao, F. 2025. Harmonyset: A comprehensive dataset for understanding video-music semantic alignment and temporal synchronization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3152–3162.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023b. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592*.

Zhuo, L.; Wang, Z.; Wang, B.; Liao, Y.; Bao, C.; Peng, S.; Han, S.; Zhang, A.; Fang, F.; and Liu, S. 2023. Video background music generation: Dataset, method and evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15637–15647.