VIMIX-14M: A CURATED VIDEO-TEXT DATASET WITH HIGH-QUALITY CAPTIONS AND CRAWL-FREE ACCESS

Anonymous authors
Paper under double-blind review

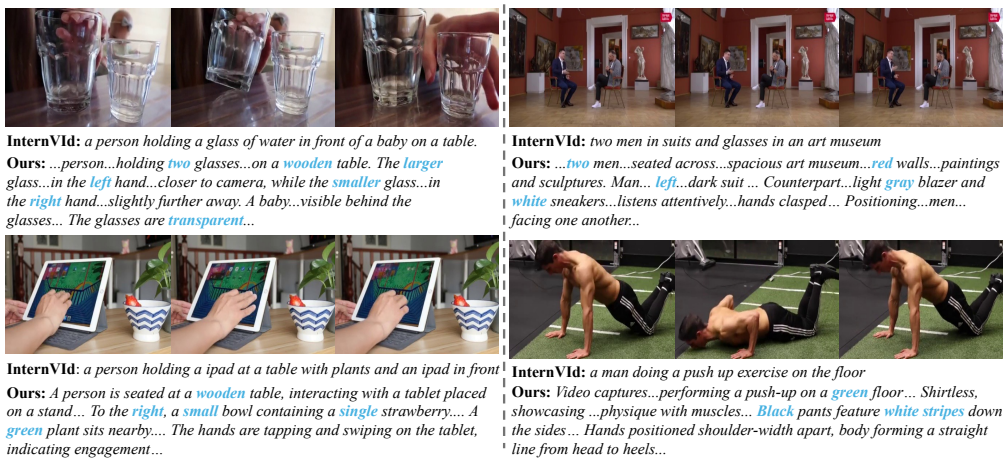


Figure 1: ***Qualitative comparison of video captions.*** In contrast to existing datasets (e.g., InternVid (Wang et al., 2023c)) that use generic descriptions, our ViMix-14M contains rich visual details in captions. As shown, the text includes precise spatial relationship, object attributes, action states, and motion details, demonstrating superior scene understanding. Key visual details are marked in blue.

ABSTRACT

Text-to-video generation has surged in interest since Sora (OpenAI, 2024), yet open-source models still face a data bottleneck: there is no large, high-quality, easily obtainable video–text corpus. Existing public datasets typically require manual YouTube crawling, which yields low usable volume due to link rot and access limits, and raises licensing uncertainty. This work addresses this challenge by introducing ViMix-14M, a curated multi-source video–text dataset of around 14 million pairs that enables end-user crawl-free, download-ready access and provides long-form, high-quality captions tightly aligned to video. ViMix-14M is built by drawing from diverse openly licensed video sources, followed by unified de-duplication and quality filtering, and a multi-granularity, ground-truth-guided re-captioning pipeline that refines descriptions to better match actions, scenes, and temporal structure. We evaluate the dataset by multimodal retrieval, text-to-video generation, and video question answering tasks, observing consistent improvements over counterpart datasets. We hope this work can help remove the key barrier to training and fine-tuning open-source video foundation models, and provide insights into building high-quality and generalizable video-text datasets.

1 INTRODUCTION

Recent advances in text-to-video generation (OpenAI, 2024; Wan et al., 2025; Blattmann et al., 2023; Yang et al., 2024; Bar-Tal et al., 2024) have demonstrated remarkable capabilities in producing photorealistic videos from natural language descriptions. These breakthroughs share a common foundation that models are trained on massive-scale, high-quality video-text paired datasets with rich semantic annotations. The success mirrors earlier progress in text-to-image generation (Balaji et al., 2022; Ramesh et al., 2021; Rombach et al., 2022; Saharia et al., 2022; Yu et al., 2022), which

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

transforms visual content creation through large-scale pre-training. However, while text-to-image generation has been democratized through large public datasets (Schuhmann et al., 2022; Sharma et al., 2018; Changpinyo et al., 2021; Byeon et al., 2022; Gadre et al., 2023), video-language research faces a critical asymmetry. Closed-source commercial models leverage proprietary datasets with curated video-text pairs, while open-source alternatives lack accessible, high-quality datasets. This accessibility barrier constrains the research community to develop competitive video understanding and generation systems, creating a persistent performance gap that has yet to be addressed through publicly available resources.

Current large-scale video datasets face three critical limitations that constrain model development. First, caption quality is limited by ASR-based annotations. Most existing large-scale datasets rely on automatic speech recognition (ASR) to extract text from video audio (Miech et al., 2019; Xue et al., 2022; Zellers et al., 2021; Lee et al., 2021). For example, HowTo100M (Miech et al., 2019) contains 136M video clips with ASR-derived transcripts, while HD-VILA-100M (Xue et al., 2022) provides 103M clips with audio-based captions. These ASR transcripts describe the spoken narration rather than the visual content—capturing what is *said* about the video rather than what is *shown*. This audio-visual misalignment limits visual semantic knowledge that models can acquire during training.

Second, limited scale and domain coverage restrict model generalization. Many existing datasets (Soomro et al., 2012; Carreira et al., 2019; Kay et al., 2017; Carreira et al., 2018; Goyal et al., 2017; Heilbron et al., 2015; Zhou et al., 2018; Rohrbach et al., 2015) concentrate on narrow domains with insufficient scale or diversity. For example, UCF101 (Soomro et al., 2012) contains only 13k videos across 101 action categories, while Kinetics-700 (Carreira et al., 2019) focuses on human actions with approximately 650k clips. Similarly, Something-Something V2 (Goyal et al., 2017) emphasizes temporal reasoning through human-object interactions but remains constrained to 220k videos within a narrow domain of basic physical actions. Although these datasets provide clean categorical labels, their limited scale and narrow domain focus are insufficient for training large-scale video-language models that require diverse visual concepts and rich semantic supervision.

Third, dataset accessibility poses significant practical barriers. Many existing datasets source videos from YouTube (Bain et al., 2021; Chen et al., 2024; Xue et al., 2022; Wang et al., 2023c), where researchers face persistent download failures due to geographic access restrictions, content removal by uploaders, API rate limits, anti-scraping measures, and copyright takedowns. ***As a result, the truly accessible videos of these datasets are often less than 1% of the originally released or claimed volume.*** For example, WebVid-10M (Bain et al., 2021) has been permanently withdrawn following copyright holder requests. Panda-70M (Chen et al., 2024) has been discontinued due to licensing restrictions, leaving only caption metadata accessible without corresponding videos. InternVid (Wang et al., 2023c), in contrast, offers a partially downloadable subset, yet the majority of videos still require YouTube access and remain subject to link rot, while its captions provide only brief descriptions that may miss temporal dynamics. This situation undermines reproducibility and democratization of video-language research, widening the performance gap between closed-source models with proprietary data and open-source alternatives.

We present ViMix-14M, a curated multi-source video-text dataset that addresses these limitations through three principles. First, we ensure reliable accessibility by sourcing from datasets with stable hosting and open licenses, enabling large-scale ***crawl-free access*** by aggregating the majority of publicly available, readily downloadable video datasets. Second, we prioritize caption quality through comprehensive re-captioning using Qwen2.5-VL-Instruct (Bai et al., 2025). Leveraging original ground-truth labels as prompts when available, we guide the model to generate rich, visually-grounded descriptions that capture objects, actions and scene dynamics. Beyond descriptive captions, we employ VBench (Huang et al., 2024) to systematically assess video quality through multi-dimensional evaluation, ensuring high standards and facilitating quality-based filtering for downstream applications. Third, we maximize content diversity by drawing from seven complementary sources spanning action recognition benchmarks, large-scale video collections, professional stock footage, and user-focused content.

Our dataset comprises 13.7 million videos totaling 22.8K hours, drawn from seven major sources: InternVid-10M-FLT(Wang et al., 2023c), VideoUFO(Wang & Yang, 2025), VidGen-1M(Tan et al., 2024), Kinetics-700(Carreira et al., 2019), Something-Something V2(Goyal et al., 2017), OpenVideo(UmiMarch, 2025), and UCF-101(Soomro et al., 2012). To accommodate diverse application needs, we generate captions at three granularity levels: short captions for quick summarization,

middle captions describing objects, colors, backgrounds, styles, and actions, and long captions that capture detailed spatial relationships, temporal reasoning, and comparative analysis between objects. Unlike existing datasets that provide only brief, generic descriptions lacking fine-grained visual information, our captions encode rich semantic details. As Figure 1 demonstrates, our captions capture precise spatial relationships (“larger glass in the left hand closer to camera"), specific color and material attributes (“wooden table"), exact quantities, and contextual scene elements. To ensure dataset video quality, we employ systematic VBench (Huang et al., 2024) evaluation across five dimensions: temporal flickering, subject consistency, background consistency, imaging quality, and aesthetic quality, guaranteeing high video standards across all samples. In summary, our work makes three principal contributions to advancing video-language research:

- We provide a unified re-captioning of 13.7M videos drawn from stably-downloadable, openly licensed source datasets, delivering end-user crawl-free access without re-hosting source videos or introducing new license constraints, enabling the community to train on a single unified corpus that consolidates widely-used public video datasets.
- We provide ground-truth-guided multi-granularity captions that leverage verified source annotations to mitigate hallucination while capturing fine-grained visual details including precise spatial relationships, object attributes, color and material specifications, exact quantities, and semantic reasoning—significantly surpassing existing brief, generic descriptions.
- We ensure content diversity through multi-source aggregation from seven major datasets and domain collections, with quality evaluation on temporal flickering, subject consistency, background consistency, imaging quality, and aesthetic quality to guarantee high video standards.

2 Related Work

Video-Language Datasets. The success of image-language models (Radford et al., 2021; Jia et al., 2021; Li et al., 2022; Liu et al., 2023) is driven by large-scale image-text datasets (Schuhmann et al., 2022; Sharma et al., 2018; Changpinyo et al., 2021; Byeon et al., 2022; Gadre et al., 2023) . However, scaling to video-language datasets presents challenges. Early video datasets such as UCF101 (Soomro et al., 2012) and Kinetics (Carreira et al., 2019) provide high-quality annotations but are limited in scale and focus on action recognition with categorical labels rather than rich textual descriptions. To address scale limitations, several works (Miech et al., 2019; Xue et al., 2022; Zellers et al., 2021; Bain et al., 2021) leverage automatic speech recognition (ASR) to extract subtitles from videos, resulting in datasets like HowTo100M (Miech et al., 2019) and HD-VILA-100M (Xue et al., 2022). While achieving scale, these ASR-based captions describe audio narration rather than visual content, creating an audio-visual misalignment that limits visual semantic learning. InternVid (Wang et al., 2023c) provides 10M downloadable clips with generated captions averaging 17.6 words, though these descriptions may inadequately capture temporal dynamics due to frame-level captioning followed by summarization. Panda-70M (Chen et al., 2024) employs multiple cross-modality teacher models to improve caption quality but faces video sources accessibility issues. In contrast, our dataset provides 13.7M stable, accessible videos with multi-granularity captions that capture rich visual semantics including spatial relationships, temporal dynamics, and object attributes.

Vision-Language Models. Vision-language models learn correlations between visual inputs and textual representations, supporting downstream tasks such as text-guided image/video generation (Balaji et al., 2022; Blattmann et al., 2023; Ho et al., 2022; Ramesh et al., 2021; Rombach et al., 2022; Saharia et al., 2022; Singer et al., 2022; Yu et al., 2022), captioning (Alayrac et al., 2022; Li et al., 2022; 2023a; Liu et al., 2023), VQA (Chen et al., 2023; Li et al., 2023b; Zhu et al., 2023), and retrieval (Luo et al., 2021; Ma et al., 2022; Li et al., 2023c; Wang et al., 2023a; 2024a). Recent advances in large vision-language models (LVLMs) such as GPT-4V (Achiam et al., 2023) and Gemini (Team et al., 2023) have demonstrated impressive multimodal understanding capabilities. For video-specific models, Video-LLaMA (Zhang et al., 2023) and VideoChat (Li et al., 2023b) extend LLMs to video understanding through learned visual projections. The Qwen series of models (Bai et al., 2025; Team et al., 2024; Yang et al., 2025; Wang et al., 2024b), including Qwen2.5-VL (Bai et al., 2025), have demonstrated strong performance on video understanding benchmarks. We leverage Qwen2.5-VL-Instruct-7B as our captioning model, which generates visually-grounded descriptions capturing objects, actions, and scene dynamics more effectively than frame-level or ASR-based approaches.

Dataset	Statistics			Duration (words)			Quality Metrics				
	#Videos	Total hrs.	Avg. len.(s)	Short	Middle	Long	Subject consist.	Background consist.	Temporal flickering	Aesthetic quality	Imaging quality
InternVid	10.6M	16.3K	5.5	14.4	50.5	109.2	0.87	0.91	0.93	0.41	0.57
VideoUFO	1.09M	2.12K	7.0	15.9	53.2	113.4	0.80	0.86	0.91	0.26	0.52
VidGen-1M	1.00M	2.26K	8.1	13.7	47.2	110.1	0.83	0.90	0.87	0.47	0.63
Kinetics-700	0.63M	1.58K	9.0	13.6	52.2	111.4	0.79	0.89	0.87	0.43	0.52
Sthv2	0.22M	234	3.8	13.9	46.1	107.8	0.85	0.90	0.88	0.33	0.54
OpenVideo	0.11M	331	11.2	17.8	57.4	117.8	0.90	0.92	0.91	0.53	0.68
UCF-101	0.01M	26.6	7.2	13.4	50.3	109.3	0.88	0.92	0.92	0.40	0.45
ViMix-14M	**13.7M**	**22.8K**	**6.0**	**14.4**	**50.6**	**109.8**	**0.86**	**0.90**	**0.92**	**0.40**	**0.57**

Table 1: ***ViMix-14M composition*** with video counts, durations, caption statistics, and quality metrics.

Video Annotation and Quality Evaluation. Several concurrent works (Wang et al., 2023c; Chen et al., 2024; Wang et al., 2023b) explore automatic video annotation using vision-language models. InternVid (Wang et al., 2023c) uses multi-modal inputs with LLM summarization, though LLMs may propagate errors from noisy captioning models. Panda-70M (Chen et al., 2024) employs multiple teacher models including BLIP-2(Li et al., 2023a), MiniGPT-4 (Zhu et al., 2023), and video-specific models, demonstrating the benefits of multi-modal ensembling. Beyond annotation, ensuring video quality is critical for training robust models. We employ VBench (Huang et al., 2024) for systematic quality evaluation across temporal flickering, subject consistency, background consistency, imaging quality, and aesthetic quality. This comprehensive evaluation framework, combined with multi-granularity captioning at short, middle, and long levels, ensures our dataset provides both high-quality videos and rich semantic annotations suitable for training large-scale video-language models.

3 Dataset ViMix-14M

Constructing ViMix-14M includes three primary stages. First, we select and curate videos from seven source domains under clear principles that all data should be end-user crawl-free, directly downloadable, and diverse across domains (detailed in Section 3.1). This process yields a total of 13.7M curated videos, which are then subjected to captioning (Section 3.2) and filtering (Section 3.3) stages. To best support diverse video–language tasks and broad applications, our data processing follows a customization-centric strategy. Specifically, for each video, we produce three caption granularities including a brief summary, a finer description, and a detailed, ground-truth-guided caption for reasoning-related tasks, which broadly cover diverse potential multimodal applications. For filtering, rather than discarding videos by fixed criteria, we provide rich per-video metadata, including duration, spatial size, and multiple video-quality indicators. This label-instead-of-filter strategy allows practitioners to flexibly customize their own filtering rules by thresholding along specific dimensions when using ViMix-14M.

3.1 Data Collection and Curation

To ensure comprehensive coverage of visual concepts, temporal dynamics, and annotation granularities, we curate a diverse video corpus spanning seven complementary datasets. Our selection combines four categories: (1) *Action recognition benchmarks* (UCF-101 (Soomro et al., 2012), Kinetics-700 (Carreira et al., 2019), Something-Something V2 (Goyal et al., 2017)) providing foundational coverage of human activities and temporal reasoning, (2) *Large-scale video-text datasets* (VidGen-1M (Tan et al., 2024), InternVid-10M-FLT (Wang et al., 2023c)) featuring diverse real-world scenarios from user-generated content with detailed captions and high visual quality, (3) *Professional stock footage* (OpenVideo (UmiMarch, 2025)) sourced from Pexels (Pexels) curated video library, offering 106k+ professionally-produced clips with high production values that provide a complementary perspective to the spontaneous, naturalistic style of user-generated datasets, and (4) *User-focused content* (VideoUFO (Wang & Yang, 2025)) addressing real user interests in text-to-video generation with minimal overlap to existing datasets. This multi-source approach ensures balanced representation across action categories, scene types, motion patterns, and quality levels, professional cinematography, and contemporary user preferences. All selected datasets are publicly available and can be downloaded both through their own official distribution channels and through third-party mirrors such as OpenDataLab (OpenDataLab) and HuggingFace (Hugging Face), which provide high-speed download infrastructure as a convenience layer rather than the sole access point. This facilitates reproducibility adoption by eliminating typical barriers of multi-source data collection.

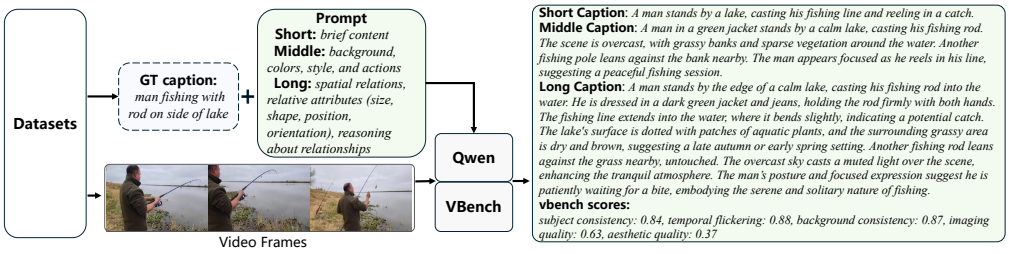


Figure 2: ***Multi-granularity caption generation and filtering pipeline.*** Given video frames and ground truth captions, we employ Qwen2.5-VL (Bai et al., 2025) with three prompt levels to generate captions at varying granularities: short captions for brief summarization, middle captions capturing scene details, and long captions encoding spatial relations and reasoning. Ground truth labels serve as optional contextual guidance to enhance caption quality but are not required for generation. VBench (Huang et al., 2024) additionally scores video quality across five dimensions.

De-duplication. De-duplication is performed at the video-segment level using YouTube video ID and temporal boundaries (`start_time`, `end_time`). Two clips are considered duplicates if they share the same video ID and overlapping temporal boundaries, and we retain one copy. Cross-source overlap accounts for only $\sim$0.35% of the full 13.7M corpus.

3.2 Multi-Granularity Captioning

We present our captioning and filtering pipeline in Figure 2. To obtain high-quality, visually-grounded captions that capture rich semantic information beyond ASR-based descriptions, we employ Qwen2.5-VL (Bai et al., 2025), a state-of-the-art open-source vision-language model with strong video understanding capabilities. The model processes video sequences at 1 FPS, enabling it to capture temporal dynamics and inter-frame relationships. To validate caption quality, we conduct (1) a *human evaluation*: we randomly sample 10,000 videos from our dataset and ask five human annotators to independently compare our generated captions against the original dataset captions. Each annotator chooses one of three options: (a) ViMix's caption is better, (b) the original caption is better, or (c) both are of similar quality, with the majority vote taken as the final preference. Our captions are preferred (88.7%) or tied (6.4%) in over 95% of cases, with only 4.9% favoring the original. The high agreement rate across annotators further validates the clarity and quality of our generated captions; and (2) detailed *downstream evaluations* in Section 4, including direct caption quality analysis, text-video retrieval, video generation, and question answering, all confirming that ViMix-14M's captions significantly outperform existing counterparts.

To leverage existing annotations while generating richer descriptions, we integrate original ground-truth labels as contextual guidance when available. We prepend the original label to our prompts in the format: "This video shows '[GT_LABEL]'. [PROMPT]". This approach provides the model with domain-specific context while allowing it to generate detailed descriptions beyond the concise ground-truth labels. For videos without ground-truth labels, we use base prompts directly. To support diverse downstream applications, we generate captions at three granularity levels:

- Short: "*Briefly describe the main content of the video in no more than 20 words.*"
- Middle: "*Describe objects in the video, including their colors, background, style, and actions, in 40 to 60 words.*"
- Long: "*Provide a detailed caption that not only describes the objects and actions in video but also highlights their relative attributes, such as differences in size, shape, height, body build, orientation, position, or quantity. The caption should also include reasoning that explains relationships between the objects, movements, or events, going beyond surface-level description, in 80 to 130 words*"

3.3 Video Quality Assessment

To ensure high video quality standards across our dataset, we employ VBench (Huang et al., 2024), a comprehensive video quality evaluation framework that assesses videos across multiple dimensions with alignment to human perception. This quality assessment allows us to filter low-quality videos and

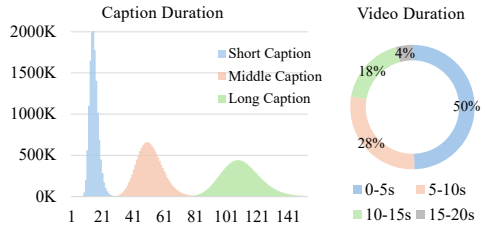


Figure 3: ***Caption and video statistics.*** Left: frequency of captions (y-axis) by word count (x-axis) for short, middle, and long granularities. Right: percentage distribution of video durations.

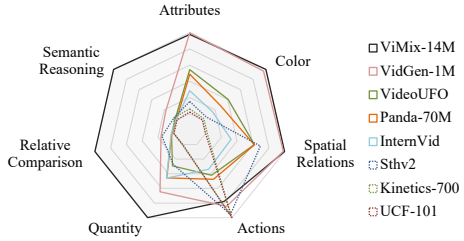


Figure 4: ***Multi-dimensional caption quality comparison.*** Our captions demonstrate superior semantic richness across all evaluated dimensions compared to existing datasets.

provide quality scores that enable researchers to customize their training data according to specific requirements. In detail, VBench provides a multi-dimensional evaluation that decomposes video generation quality into specific, hierarchical dimensions including temporal quality, frame-wise quality, and video-condition consistency. It goes beyond single-metric scoring by offering automated, multi-dimensional evaluation, with its metrics validated against human preference data. We evaluate each video across five key dimensions in VBench: *Temporal Flickering*, *Subject Consistency*, *Background Consistency*, *Aesthetic Quality*, and *Imaging Quality*. Details are provided in Appendix C.

Dataset Statistics. Table 1 presents statistics across all seven source datasets. Our corpus comprises 13.7M video clips totaling 22.8K hours, with an average duration of 6.0 seconds. Figure 3 illustrates the distribution of caption lengths and video durations in our dataset. The left histogram corresponds to the length distributions of short, middle, and long captions, while the right pie chart shows the proportion of videos across different duration ranges.

4 Observations

4.1 Multi-Dimensional Caption Quality Analysis

To assess caption semantic richness, we conduct a multi-dimensional evaluation across seven key dimensions of visual understanding. Specifically, *Object Attributes* examines physical and visual characteristics including shape, size, material, texture, and structure. *Spatial Relations* evaluates spatial organization through position, direction, distance, and relational patterns. *Color* assesses color-related information including specific hues, tones, combinations, and contrasts. *Quantity* measures numerical information through exact counts, quantity words, distribution, and groups. *Relative Comparison* evaluates explicit comparisons between objects through size differences, contrast words, and relative terms. *Actions* examines movements and behaviors through action verbs, manner descriptions, and object interactions. *Semantic Reasoning* assesses high-level interpretation beyond surface description including symbolism, inference, atmosphere, and abstract concepts. Using Qwen3-4B-Instruct (Yang et al., 2025) as an automated evaluator, we analyze captions through dimension-specific binary classification prompts. Each prompt begins with a yes/no question in the form *"Does this caption ...?"*, followed by key cues, example pairs, and an instruction to answer with "True" or "False." Figure 4 presents the comparative analysis across datasets. Our captions demonstrate superior coverage across all dimensions. While action recognition datasets like Kinetics-700 and UCF-101 excel in Actions, our multi-granularity approach achieves balanced coverage spanning low-level perceptual details to high-level semantic reasoning.

4.2 Video Generation with Finetuning

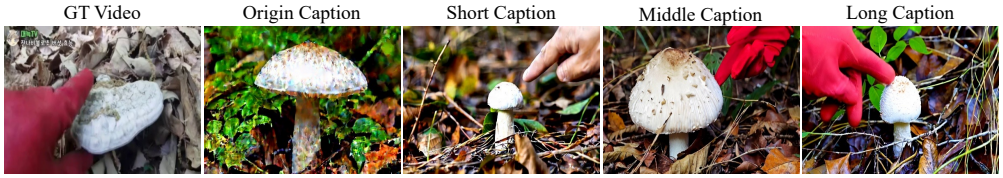
To validate the effectiveness of ViMix-14M for training video generation models, we finetune FastVideo (The FastVideo Team, 2024) on InternVid (Wang et al., 2023c) videos using different caption sources. We perform multiple rounds of proportional randomly sampling, each drawing 15K videos for finetuning and 200 videos for evaluation, with numbers averaged across runs. We compare four settings: (1) InternVid: finetuned and tested with original InternVid captions; (2) InternVid†: finetuned with InternVid captions but tested with ViMix long captions; (3) FastVideo: pretrained

Caption	Text-to-Video							Video-to-Text						
	InternVid	K700	UCF	VidGen	UFO	Open	Sthv2	InternVid	K700	UCF	VidGen	UFO	Open	Sthv2
Recall@1 (%) ↑														
Origin	76.5	31.0	12.4	93.0	65.4	89.8	22.4	73.5	34.0	12.6	93.6	71.2	90.8	24.0
ViMix-14M (results shown below):														
Short	83.6	71.8	39.8	87.2	82.0	88.2	47.8	80.0	75.8	41.4	87.4	81.0	90.4	50.6
Middle	**90.2**	84.2	**63.4**	**95.6**	**91.4**	88.8	61.8	**89.8**	**90.8**	**68.0**	94.4	91.8	90.4	69.0
Long	89.6	**86.6**	59.6	94.2	88.6	**89.8**	**66.0**	87.2	89.6	64.4	**95.8**	**92.6**	**91.2**	**72.4**

Table 2: ***Bidirectional video-text retrieval across different caption lengths.*** We compare four caption configurations on seven datasets using R@1. **Bold** indicates best performance per dataset. Shading from light to medium to dark corresponds to Short, Middle and Long.

Caption	T→GV R@1 ↑	GV→T R@1 ↑	GV→OV R@1 ↑	T-GV Sim. ↑	GV-OV Sim. ↑	Subj. Cons. ↑	Bg. Cons. ↑	Temp. Flicker ↑	Motion Smooth. ↑	Dynamic Degree ↑	Aesth. Qual. ↑	Imaging Qual. ↑
InternVid	73.00	66.50	33.50	0.29	0.59	0.93	0.93	0.97	0.98	0.39	0.45	0.55
InternVid†	88.00	92.50	74.00	0.33	0.67	0.94	0.94	0.97	0.98	**0.61**	0.46	0.57
FastVideo	92.00	94.50	75.00	0.35	0.68	0.95	0.95	0.98	0.98	0.51	**0.50**	**0.69**
Ours	**92.50**	**95.50**	**77.00**	**0.35**	**0.70**	**0.96**	**0.95**	**0.99**	**0.99**	0.55	0.47	0.61

Table 3: ***Video generation quality with finetuning.*** We finetune FastVideo on InternVid to evaluate our caption quality. InternVid†: tested with ViMix long captions. GV: generated video, OV: original video, T: text caption. **Bold** indicates best.



Origin Caption: *a white mushroom in the middle of fallen leaves.* **Short Caption:** *A hand points to a small white mushroom on the forest floor among dry leaves..* **Middle Caption:** *The video features a white mushroom with a textured cap, surrounded by dry, brown leaves on the forest floor. A hand wearing a red glove points at the mushroom, highlighting its details. The scene is natural and earthy, capturing the mushroom's texture and the surrounding environment.* **Long Caption:** *The video captures a close-up view of a white mushroom nestled among dry, brown leaves on the forest floor. A hand wearing a red glove gently points at the mushroom, emphasizing its small size compared to the surrounding foliage. The mushroom's cap is slightly cracked, revealing a textured surface beneath. The scene suggests a natural setting where the mushroom is likely part of a larger ecosystem, possibly being examined for its characteristics or potential use.*

Figure 5: ***Video generation quality comparison across caption granularities.*** Origin Caption (InternVid (Wang et al., 2023c)) provides basic scene; Short Caption adds hand-mushroom interaction; Middle Caption predicts the gloved hand with color; Long Caption achieves best results with improved posture and correct contact direction. Quality hierarchy: Origin < Short < Middle < Long.

model without finetuning; and (4) Ours: finetuned and tested with ViMix long captions. We evaluate text-video alignment via retrieval metrics (Luo et al., 2021) and CLIP similarity, along with video quality metrics from VBench (Huang et al., 2024). We denote T as text caption, GV as generated video, and OV as original video. As shown in Table 3, finetuning with ViMix-14M captions yields the best overall performance, confirming that richer captions provide superior training supervision.

4.3 Text-Video Retrieval Evaluation

To validate caption quality, we perform multiple rounds of retrieval experiments using Clip4Clip (Luo et al., 2021), each on 500 randomly sampled video-text pairs per dataset, with all reported numbers averaged across runs. We evaluate both text-to-video (T2V) and video-to-text (V2T) retrieval, comparing four text representations: original captions, short captions, middle captions, and long captions. As shown in Table 2, our captions substantially outperform origin labels in both directions validating the robust video-text alignment of our multi-granularity design.

4.4 Quality Evaluation via Video Generation

To evaluate caption quality for text-to-video generation, we measure CLIP-based (Radford et al., 2021) similarity scores between videos generated by Wan2.2 (Wan et al., 2025) model from different captions and their corresponding source videos using Clip4Clip (Luo et al., 2021). Table 4



a: InterVid: *A meat product with a sign that says price at the store*
b: Ours(short): *Pork loin back ribs on sale for $5.99 per pound*
Q1: What type of meat is displayed in the image? *A1*: Pork loin back ribs. (*a*|*b*)
Q2: What is the price per pound for the pork ribs on sale? *A2*: $5.99 per pound. (*a*|*b*)
Q3: What does the sign indicate about the meat? *A3*: It says "Quality." (*a*|*b*)

a: Panda-70M: *A group of men in suits and ties sitting next to each other.*
b: Ours(short): *Hosts discuss food on a TV cooking show set with prepared dishes.*
Q1: What type of TV program is shown in the image? *A1*: A cooking show. (*a*|*b*)
Q2: How many people are standing behind the kitchen counter? *A2*: Three people. (*a*|*b*)
Q3: What is on the counter in front of the hosts? *A3*:Fried food, lettuce, and baked pasta. (*a*|*b*)

a: VideoUFO: *A bicycle parked on a sidewalk next to a building*
b: Ours(short): *A bicycle with large tires is parked on a sidewalk near a building, surrounded by greenery and a paved area.*
Q1: What type of vehicle is shown in the image? *A1*: A bicycle. (*a*|*b*)
Q2: What color are the tires of the bicycle? *A2*: The tires are beige or light tan. (*a*|*b*)
Q3: Where is the bicycle parked? *A3*: It is parked near a building. (*a*|*b*)

Figure 6: ***Caption evaluation via video question-answering tasks.*** Comparison between baseline captions (a) and ours (b) across multiple datasets. Color markers show answer coverage. Our captions capture richer visual details, leading to more accurate responses for attribute-specific and content-focused questions.

demonstrates that our generated captions achieve substantially higher video-text alignment scores than original labels, with long captions yielding the highest average similarity, followed by middle captions—both significantly exceeding the origin baseline. Action recognition datasets exhibit the most significant improvements, while high-quality source datasets such as OpenVideo and VidGen-1M show consistently strong alignment. These results validate that our multi-granularity captions establish strong video-text alignment, supporting text-to-video generation and video-language tasks. As shown in Figure 5, captions evolve from basic object descriptions (“a white mushroom in fallen leaves") to richer representations incorporating human actions (“hand points"), fine-grained attributes (“textured cap", “red glove"), comparative spatial relationships (“small size compared to surrounding foliage"), and contextual reasoning (“examined for characteristics"). This progression reflects visual understanding, capturing not only objects but also their properties, spatial arrangements, and context.

4.5 Quality Assessment via VQA

To further validate the semantic richness of our generated captions, we conduct text-only Video Question Answering (VQA) experiments where models answer video-related questions using solely textual descriptions without visual input. We evaluate caption quality by comparing the accuracy of answers derived from original captions *vs.* our captions across diverse question types, including object counting, object identification, attribute recognition (e.g., color, price), spatial localization, and scene understanding. For each video sample, we generate 3 diverse questions covering various semantic aspects using GPT-5 (OpenAI, 2025), which views the original video frames and creates question-answer pairs as ground truth. We then employ GPT-5 to answer these questions based purely on textual captions—both original and ours—without access to visual content. Answer correctness is evaluated using an LLM-as-judge approach, where GPT-5 assesses semantic equivalence between predicted and ground truth answers, accounting for synonyms and paraphrasing. Figure 6 illustrates qualitative comparisons across multiple datasets. For baseline captions (a), sparse descriptions limit answer coverage, particularly for attribute-specific questions (“What color is the object?") and fine-grained content queries (“How many people are visible?"). In contrast, our captions (b) provide richer visual details—capturing object counts, colors, spatial arrangements, and contextual information—enabling accurate responses across diverse question types. The color-coded markers show our captions achieve comprehensive answer coverage, while baseline captions fail to provide sufficient information. Our multi-granularity captions encode richer semantic information, enabling effective video understanding through text alone.

4.6 Video Quality Analysis via VBench

We examine how different VBench quality thresholds affect retrieval performance. We apply quality-based filtering with thresholds of 0.5 for imaging quality, 0.3 for aesthetic quality, 0.7 for background consistency, 0.7 for temporal flickering, and 0.5 for subject consistency, then evaluate using Clip4Clip (Luo et al., 2021). Table 5 shows that imaging quality filtering achieves the best overall performance on T2V retrieval across all caption granularities. Notably, aesthetic quality performs

Caption	Sthv2	VidGen	UFO	Open	InternVid	UCF	K700	Avg
Origin	0.50	0.69	0.55	0.79	0.54	0.42	0.45	0.56
ViMix-14M (results shown below):								
Short	0.58	0.60	0.57	0.69	0.55	0.64	0.64	0.61
Middle	0.65	0.70	0.59	0.77	0.66	**0.74**	**0.71**	0.69
Long	**0.65**	**0.70**	**0.62**	**0.80**	**0.70**	0.72	0.70	**0.70**

Table 4: ***Video-text similarity scores*** across caption granularities (mean CLIP similarity between generated and source videos). Long captions (highlighted) achieve superior alignment.

Threshold	T2V R@1 ↑			V2T R@1 ↑		
	Short	Middle	Long	Short	Middle	Long
Baseline	83.6	90.2	89.6	80.0	89.8	87.2
subject_cons.	83.0	91.0	89.6	82.0	91.0	89.6
temporal_flick.	82.4	91.4	91.0	80.4	91.6	89.6
background_cons.	82.2	89.6	89.8	79.2	91.2	90.4
aesthetic_qual.	83.0	91.4	90.4	**84.0**	92.0	90.6
imaging_qual.	**88.4**	**94.2**	**93.8**	83.2	**93.6**	**92.2**

Table 5: ***Comparison of R@1 scores*** for T2V and V2T retrieval across different VBench quality dimensions. Imaging quality filtering (highlighted) achieves superior performance.

best on short V2T tasks, suggesting that aesthetic features better capture semantic information in brief content. These results demonstrate that quality-based filtering effectively enhances cross-modal retrieval, with imaging quality being the most impactful dimension.

4.7 Captioner Backbone Ablation

To justify our choice of model size over larger backbones, we re-generate captions using Qwen2.5-VL-72B under the same pipeline configuration. Following the evaluation protocols in Sections 4.3 and 4.4, we compare similarity and bidirectional R@1 averaged across all seven benchmarks. As shown in Table 6, all gaps fall within $\pm 3\%$(relative) across every granularity, with no consistent trend favoring either model, confirming random fluctuation rather than a systematic quality gap.

Model	Similarity			T2V R@1			V2T R@1		
	Short	Middle	Long	Short	Middle	Long	Short	Middle	Long
7B	0.61	0.69	0.70	71.5	82.2	82.1	72.4	84.9	84.7
72B	0.60	0.70	0.72	71.2	82.2	82.7	71.5	84.3	85.2

Table 6: ***Captioner backbone scaling.*** Differences in similarity and bidirectional R@1 are within $\pm 3\%$(relative) across all granularities, confirming that 7B is already sufficient.

4.8 Comparison with Human References on MSR-VTT

To verify caption quality against human-written references, we evaluate on the MSR-VTT (Xu et al., 2016) test set using CLAIR (Chan et al., 2023), an LLM-based metric that scores a candidate caption against a human reference on a 0–100 scale. We adopt Claude Sonnet 4.6 (Anthropic, 2026) as the judge, which is independent of the Qwen2.5-VL captioner and thus avoids shared hallucination patterns. We report two conditions: *w/o GT guidance*, where Qwen2.5-VL sees only video frames, and *w/ GT guidance*, where the original label is also given as context—matching our actual pipeline. Human inter-annotator agreement (88.4) serves as the upper bound. As shown in Table 7, Qwen2.5-VL alone scores 70.9–77.0 without GT context; under GT guidance, scores rise to 86.2–88.0, approaching or matching human agreement.

Condition	Human	Short (w/o GT)	Middle (w/o GT)	Long (w/o GT)	Short (w/ GT)	Middle (w/ GT)	Long (w/ GT)
CLAIR score (/100)	88.4	70.9	76.2	77.0	**87.3**	**88.0**	**86.2**

Table 7: ***CLAIR scores on MSR-VTT against human references.*** Human inter-annotator agreement (88.4) serves as the upper bound. Our captions approach or match this bound across all granularities.

5 Conclusion

We present ViMix-14M, a large-scale video-text dataset comprising 13.7 million videos with multi-granularity high-quality captions and systematic per-video quality metadata. Through comprehensive re-captioning with Qwen and aggregation across seven complementary sources, our dataset achieves substantial improvements over existing counterparts, providing a stable and semantically rich resource that enables the broader research community to develop competitive video-language models. While ViMix-14M provides a large-scale, stably accessible corpus with high-quality captions, it inherits several constraints from its source datasets, which we discuss in Appendix.

AI USE STATEMENT

The models used to construct and evaluate ViMix-14M are described in the main text: Qwen2.5-VL-Instruct-7B for caption generation, Qwen3-4B-Instruct for the automated caption quality analysis, GPT-5 for the text-only VQA protocol, and Claude Sonnet 4.6 as the CLAIR judge; prompts and model settings are reported alongside each experiment and in the appendix. Caption quality was validated by human annotators and against human-written references. We additionally used AI tools to improve readability, and we have reviewed all AI-assisted work.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Anthropic. Claude sonnet 4.6 system card, 2026. URL `https://anthropic.com/claude-sonnet-4-6-system-card`.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 2023.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

David M Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. Clair: Evaluating image captions with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, Singapore, December 2023. Association for Computational Linguistics.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3558–3568, 2021.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36: 27092–27112, 2023.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850, 2017.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970, 2015. doi: 10.1109/CVPR.2015.7298698.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Hugging Face. Hugging Face. `https://huggingface.co`. Accessed: 2025.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10274–10284, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023b.

Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 19948–19960, 2023c.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 638–647, 2022.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2630–2640, 2019.

OpenAI. Sora. `https://openai.com/index/sora`, 2024.

OpenAI. Introducing gpt-5. Blog post, OpenAI, August 2025. URL `https://openai.com/index/introducing-gpt-5/`.

OpenDataLab. OpenDataLab. `https://opendatalab.com`. Accessed: 2025.

Pexels. Pexels. `https://www.pexels.com`. Accessed: 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3202–3212, 2015.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, and Hao Li. Vidgen-1m: A large-scale dataset for text-to-video generation. *arXiv preprint arXiv:2408.02629*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3), 2024.

The FastVideo Team. Fastvideo: A unified framework for accelerated video generation, April 2024. URL `https://github.com/hao-ai-lab/FastVideo`.

UmiMarch. Openvideo. `https://github.com/UmiMarch/OpenVideo`, 2025. GitHub repository.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Feng Wang, Manling Li, Xudong Lin, Hairong Lv, Alex Schwing, and Heng Ji. Learning to decompose visual features with latent textual prompts. In *The Eleventh International Conference on Learning Representations*, 2023a.

Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, pp. 315–332. Springer, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Wenhao Wang and Yi Yang. Videoufo: A million-scale user-focused dataset for text-to-video generation. *arXiv preprint arXiv:2503.01739*, 2025.

Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. 2023b.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023c.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5036–5045, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in neural information processing systems*, 34:23634–23651, 2021.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A LIMITATIONS

ViMix-14M unifies and recaptions the majority of stably-hosted, openly licensed video datasets currently available. We do not re-host or redistribute any original video, metadata or caption data, nor rely on fair use to justify their inclusion. However, we cannot guarantee with 100% certainty that none of these sources will be withdrawn in the future—copyright disputes affecting any individual source could render the corresponding portion of the corpus inaccessible. Users of ViMix-14M should comply with the original license terms of each constituent source dataset when using the corresponding subset. The multi-source design ensures that ViMix-14M does not depend on any single dataset, partially mitigating this risk. Additionally, although the multi-source strategy objectively broadens geographic coverage, content from less-developed regions remains under-represented, which may affect the generalizability of models trained on ViMix-14M to diverse cultural settings. To the best of our knowledge, ViMix-14M incorporated nearly all large-scale open video datasets available at the time of construction; the substantial size differences among these sources introduce a compositional imbalance that ViMix-14M inevitably inherits—a bias reflecting the current distribution of open-source video data in the community rather than our curation choices. Finally, our comparison with some prior datasets is limited. Datasets such as Panda-70M provide only video URLs and captions, while retrieving the raw videos would require crawling YouTube, which is inconsistent with its terms of service. We therefore did not obtain these videos, and limit our comparison to qualitative visualization and caption-level analysis rather than video-level evaluation.

B DATA STRUCTURE OF VIMIX-14M

We organize our curated corpus into standardized JSON for each source dataset. Each video entry follows a consistent schema:

```
{
  "segment_id": <unique video identifier>,
  "dataset_name": <source dataset name>,
  "split": <train|val|test>,
  "duration": <video length in seconds>,
  "use_gt_label": <True|False>,
  "start_time": <timestamp in HH:MM:SS.mmm>,
  "end_time": <timestamp in HH:MM:SS.mmm>,
  "caption_short_en": <10-20 words>,
  "caption_middle_en": <40-60 words>,
  "caption_long_en": <80-130 words>,
  "vbench_scores": {
    "subject_consistency": <0-1 score>,
    "background_consistency": <0-1 score>,
    "temporal_flickering": <0-1 score>,
    "aesthetic_quality": <0-1 score>,
    "imaging_quality": <0-1 score>
  }
}
```

The `use_gt_label` field indicates whether original ground-truth labels are retained from source datasets. Temporal boundaries (`start_time`, `end_time`) represent actual video segment positions, where `end_time - start_time = duration` always holds. This structure enables efficient data loading, granularity-specific caption retrieval, and quality-based filtering.

C VBENCH APPLICABILITY AND 10-FRAME SAMPLING

Detailed VBench dimensions. We evaluate each video across the following five VBench (Huang et al., 2024) dimensions:

- *Temporal Flickering* assesses local temporal inconsistencies in brightness and color across consecutive frames. Higher flickering scores indicate better stability, which is essential for training models that learn temporal patterns.

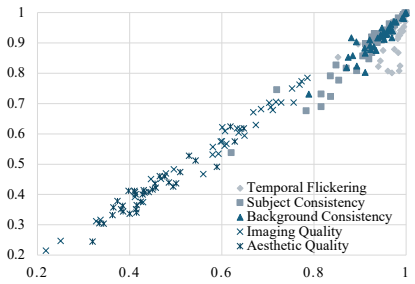


Figure 7: ***Impact of video compression on VBench quality metrics.*** X-axis: original scores; Y-axis: compressed scores. Imaging quality is normalized to 0-1. Strong correlation indicates minimal compression impact.

- *Subject Consistency* evaluates whether the main subject's appearance and identity remain stable throughout the video duration. This metric is critical for learning robust object representations across temporal sequences.
- *Background Consistency* measures the temporal stability of background elements and scene context. Stable backgrounds enable models to better distinguish foreground dynamics from static environmental features.
- *Aesthetic Quality* assesses the visual appeal and artistic value of video frames, including compositional balance, color harmony, and photorealistic rendering quality.
- *Imaging Quality* evaluates technical clarity aspects such as sharpness, proper exposure, color fidelity, and the absence of visual artifacts like blur or noise.

Applicability to natural videos. Although VBench (Huang et al., 2024) was originally introduced to evaluate generative video models, its underlying metrics are computed on raw frame statistics and are therefore domain-agnostic. VBench itself reports a *WebVid-Avg* baseline obtained by running all five dimensions on real WebVid-10M videos, and concurrent work VideoUFO (Wang & Yang, 2025) directly applies VBench to evaluate 1.09M real YouTube clips. This justifies our use of VBench as an automated, multi-dimensional quality estimator over the natural videos in ViMix-14M.

10-frame uniform sampling for tractability. To enable efficient processing of large-scale datasets, we uniformly sample 10 frames per video before VBench quality assessment. We use these scores only for *relative ranking* — quality-based filtering and per-video metadata — rather than as absolute quality measurements. Figure 7 shows strong correlation between original-video and 10-frame scores across all five dimensions, so the relative ordering on which our filtering decisions depend is preserved under this compression.

D Implementation Details

We process videos at 1 FPS and set maximum token limits of 30, 100, and 200 for short, middle, and long captions respectively. For videos longer than 20 seconds, we randomly sample a 5-15 second segment to balance computational cost and temporal information preservation. For video quality assessment, we sample up to 10 frames per video; videos with fewer than 5 frames are excluded from the dataset. We deploy the captioning pipeline across 36 NVIDIA A5000 GPUs with batch processing, where each video is processed through all three prompt levels simultaneously using batch inference with mixed precision (FP16) to accelerate generation. The complete pipeline, including caption generation and video quality assessment, takes approximately one month to process the entire dataset. For CLIP4CLIP, we use the pre-trained checkpoint from HuggingFace (*Searchium-ai/clip4clip-webvid150K*).

E Additional Video Generation Examples

Figure 8 presents additional examples demonstrating how caption granularity progressively improves video generation quality, from origin captions providing basic context, to short captions adding



Figure 8: ***Additional video generation examples across caption granularities.*** Similar to Figure 5 in the main paper, these examples demonstrate how increasing caption detail (from Origin to Short to Middle to Long) progressively improves generation quality. Each row shows: GT Video, Origin Caption generation, Short Caption generation, Middle Caption generation, and Long Caption generation, with the corresponding captions displayed below.

primary actions, middle captions incorporating visual details like colors and spatial relationships, and long captions achieving the highest quality through fine-grained attributes and contextual reasoning.

Across diverse scenarios—from tea preparation and childcare to hotel room settings and outdoor garden activities—we observe consistent quality improvements as caption length and detail increase. For example, the tea preparation scene demonstrates how long captions describing material properties ("transparent mug" versus "opaque jar"), motion dynamics ("leaves gently unfurl and float"), and symbolic meaning enable generation of videos with correct attributes and appropriate atmosphere. Similarly, the hotel bed example shows how descriptions of lighting, positioning, and ambiance translate directly to generation quality, with long captions producing videos that accurately capture color schemes, spatial layouts, and mood.



a.sthv2: *lifting salve up completely, then letting it drop down*
b.Ours(short): *A tube of lifting salve is shown being lifted and then dropped back onto a patterned surface.*

Q1: What object is shown in the image?
A1: It is a small tube of ointment or medicinal cream placed on the surface. (*a*|*b*)
Q2: Where is the tube placed?
A2: The tube is lying flat on a patterned carpet with decorative designs. (*a*|*b*)
Q3: What is the person doing with the tube?
A3: The person is reaching out to pick up the tube from the carpet. (*a*|*b*)
Q4: What happens in the final image?
A4: The person lifts the tube and holds it in their hand above the surface. (*a*|*b*)

a. Kinetics: *photocopying*
b.Ours(short): *A person places paper on a scanner, closes it, and prepares to copy documents.*

Q1: What device is shown in the image?
A1: It is a photocopier or multifunction printing machine. (*a*|*b*)
Q2: What part of the copier is visible here?
A2: The automatic document feeder located at the top of the copier. (*a*|*b*)
Q3: What is the main function of this part?
A3: Feeding multiple pages automatically for scanning or copying. (*a*|*b*)
Q4: Where should the papers be placed for copying?
A4: They should be placed on the top tray under the feeder. (*a*|*b*)



a.VidGen: *In this video, we see a raccoon and a dog interacting with each other in a playful manner. The raccoon is standing on its hind legs and appears to be hugging the dog, who is lying on the ground. The dog seems to be enjoying the attention and is looking up at the raccoon. The raccoon is moving its front paws around the dog's face and body, and the dog is responding by moving its head and body. The video captures a heartwarming moment between two unlikely animal friends.*
b.Ours(middle): *A raccoon and a white dog interact on a light-colored floor. The raccoon, with black and gray fur, nuzzles the dog's face, while the dog lies on its back, seemingly enjoying the affection. The scene is indoors, possibly in a pet-friendly area, with a few chairs visible in the background. The video captures a moment of interspecies friendship.*
c.Ours(long): *In this heartwarming scene, a playful raccoon and a white dog engage in a tender moment of companionship on a smooth, light-colored floor. The raccoon, with its distinctive black and gray fur and striped tail, stands protectively over the dog, which lies on its back, appearing relaxed and content. The dog's white fur contrasts sharply with the raccoon's darker tones, emphasizing their contrasting yet harmonious bond. The raccoon's gentle nuzzling and the dog's calm demeanor suggest a deep friendship, highlighting the unique and endearing relationship between these two unlikely animal friends.*

Q1: What animals are shown in the image?
A1: A raccoon and a white dog are seen together in the image. (*a*|*b*|*c*)
Q2: What is the raccoon doing to the dog?
A2: The raccoon is hugging or resting on the dog in a friendly manner. (*a*|*b*|*c*)
Q3: How does the dog appear to feel in the scene?
A3: The dog looks calm, relaxed, and comfortable with the raccoon. (*a*|*b*|*c*)
Q4: What is a notable feature of the raccoon's tail?
A4: The raccoon has a clearly visible, striped tail that stands out. *(a*|*b*|*c)*

a.VidGen: *The video shows a man sitting on a yellow inflatable duck in a pool. He is wearing a white t-shirt and shorts. The man is holding a chainsaw and appears to be cutting the duck in half. The pool is surrounded by a patio with red chairs and tables. There are trees and hills in the background.*

b.Ours(middle): *A man rides a yellow inflatable duck in a pool, while another person walks nearby. The setting is an outdoor patio with red chairs, a table, and palm trees in the background. The scene is bright and sunny, suggesting a leisurely day.*

c. Ours(long): *In a luxurious outdoor setting with a stunning mountain backdrop, a man is seen riding a large yellow inflatable duck float in a crystal-clear pool. He is casually dressed in a white t-shirt and dark shorts, exuding a relaxed vibe. Nearby, a woman in a plaid shirt and denim skirt walks along the poolside, her hand raised in a playful gesture. The scene captures a moment of leisure and fun, with the vibrant yellow duck contrasting beautifully against the serene blue water and lush greenery. The positioning of the man on the duck and the woman walking adds a dynamic element to the tranquil environment, highlighting the contrast between active play and calm observation.*

Q1: What is the man sitting on in the pool?
A1: He is sitting on a large yellow inflatable duck floating in the pool. (*a*|*b*|*c*)
Q2: What is the inflatable duck wearing?
A2: The inflatable duck is wearing a pair of black sunglasses. (*a*|*b*|*c*)
Q3: What is the man holding on the inflatable duck?
A3: Holding a motorized device or propeller (*a*|*b*|*c*)
Q4: What is the other person doing near the pool?
A4: Another person is standing by the poolside with his hands raised.(*a*|*b*|*c*)

Figure 9: ***Additional VQA examples comparing baseline captions with our captions across multiple datasets.*** For each example, we show: baseline caption from the original dataset, and our generated caption, followed by questions and answers. Color markers (a|b) or (a|b|c) indicate which captions can correctly answer each question, demonstrating our captions' superior coverage of visual details.

F Additional VQA Examples

To further demonstrate the richness of our captions for video question answering tasks, we provide additional qualitative examples across multiple datasets in Figure 9. These examples showcase how our multi-granularity captions enable accurate answers to diverse question types that require fine-grained visual understanding.

For action recognition datasets (Kinetics-700, Something-Something V2), our captions provide sufficient detail to answer questions about specific actions, object interactions, and temporal sequences that are typically missing from categorical labels. For instance, in the Something-Something V2 example, our caption describes not only the lifting action but also the object's appearance and surface placement, enabling accurate responses about object type and location. For large-scale video datasets (VidGen-1M), our captions capture rich contextual information including object counts, spatial arrangements, colors, and scene composition. The color-coded answer markers (a|b) or (a|b|c) clearly illustrate that original captions often fail to provide sufficient information for answering detailed questions, while our captions achieve near-complete coverage across all question types, validating their semantic richness.

G PER-SOURCE LICENSE TABLE

Table 8 summarizes the license under which each source dataset is publicly released. We do not re-host or redistribute any source video, metadata or caption data. Because the merged corpus inherits the most restrictive of the constituent licenses (e.g. CC BY-NC and CC BY-NC-SA), **ViMix-14M as a whole is intended for academic and non-commercial research**.

Source dataset	License
Kinetics-700 (Carreira et al., 2019)	CC BY 4.0
VideoUFO (Wang & Yang, 2025)	CC BY 4.0
VidGen-1M (Tan et al., 2024)	Apache 2.0
InternVid (Wang et al., 2023c)	CC BY-NC-SA 4.0
OpenVideo (UmiMarch, 2025) (Pexels)	CC BY 4.0
UCF-101 (Soomro et al., 2012)	CC BY-NC 4.0
Something-Something V2 (Goyal et al., 2017)	Qualcomm research license

Table 8: ***Per-source license summary.*** ViMix-14M aggregates seven openly licensed source datasets. The merged corpus inherits the most restrictive of the constituent licenses and is therefore intended for academic and non-commercial research only.

License definitions.

- **CC BY 4.0**: free use, modification, and distribution (including commercial) with attribution to the original authors.
- **Apache 2.0**: free use, modification, and distribution (including commercial); includes an explicit patent grant and requires preservation of license notices.
- **CC BY-NC 4.0**: non-commercial use only, with attribution.
- **CC BY-NC-SA 4.0**: non-commercial use only, with attribution; any derivative work must be released under the same license (share-alike).
- **Qualcomm research license**: academic and research use only; redistribution and commercial use are prohibited.

H BIAS ANALYSIS

We analyze the geographic and demographic bias of ViMix-14M at the per-dataset level. Our seven source datasets cover complementary content domains—action recognition benchmarks, user-generated content, and professional footage—that naturally span different demographics, geographic regions, and age groups. Multi-source aggregation itself serves as an effective bias mitigation strategy, as biases present in individual sources tend to cancel across sources.

Geographic coverage. As the largest source dataset in ViMix-14M, InternVid evenly distributes regional language content—Japanese (15%), Korean (16%), Chinese (18%), Thai (5%), English (18%), Russian (16%), French (8%)—covering Asia, Europe, and North America. OpenVideo (Pexels) contributes professionally produced footage from internationally diverse regions. UCF-101 and Kinetics-700, as action recognition benchmarks, contain videos filmed across a wide range of indoor and outdoor settings worldwide. UFO and Something-Something v2 further diversify the corpus with user-generated content spanning various cultural backgrounds.

Remaining limitations. Despite the broad coverage achieved through multi-source aggregation, content from certain regions (e.g., Africa, South America, and the Middle East) and age groups (e.g., elderly populations) remains under-represented. This imbalance is inherited from the source datasets themselves, which predominantly feature content created by young to middle-aged individuals in developed regions. Models trained on ViMix-14M may therefore exhibit reduced generalizability to these under-represented settings.