ADVANCING VIDEO-TEXT PRETRAINING WITH MULTI-VIEW CAPTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video-text pretraining has achieved remarkable progress through the scaling of models and datasets, yet the quality of language supervision remains underexplored. Existing web-scale datasets often provide only a single sparse caption per video that fails to capture rich spatiotemporal semantics, while directly using captioning models can generate noisy descriptions. We propose a large-scale multimodal large language model-based supervision generation framework that improves supervision diversity, fidelity, and semantic coverage. Starting from 10 million videos, our approach generates **m**ulti-**v**iew **c**aptions (MVC) through complementary summary and detailed captions, reasoning-based refinement, and semantic positive caption generation. To effectively exploit supervision at different granularities, we further introduce a granularity-aware text representation with separate `CLS` tokens for summary and detailed views. We pretrain video-text models using the resulting supervision corpus and evaluate them across standard, fine-grained and detailed text-to-video retrieval benchmarks. Our approach consistently improves both zero-shot and fine-tuned performance while using smaller pretraining corpora than existing methods, demonstrating the importance of rich and complementary textual supervision for video-text pretraining.

1 INTRODUCTION

Video-text pretraining has emerged as a fundamental paradigm for learning transferable video representations (Xu et al., 2021; Wang et al., 2022b; Bain et al., 2021; Wang et al., 2023; Li et al., 2023b; Lei et al., 2023; Wang et al., 2024b;c; Doughty et al., 2024). By aligning videos with natural language at scale, these approaches learn semantic representations that generalize across diverse downstream tasks, including retrieval, captioning, video question answering, etc. Inspired by the success of image-text pretraining (Radford et al., 2021; Jia et al., 2021; Li et al., 2021; 2022b), recent efforts have scaled both video-text corpora and model capacity (Wang et al., 2024b;c; Chen et al., 2024c), establishing scaling as a dominant paradigm for advancing video-text learning.

Despite this progress, most works primarily improve performance through larger datasets, stronger objectives, and increased model capacity, while the quality of language supervision itself remains comparatively underexplored. Existing video-text datasets are largely collected from web (Bain et al., 2021; Miech et al., 2019; Wang et al., 2024b) and rely on captions that are often short, sparse, weakly aligned, or noisy. As a result, rich video content is frequently compressed into a single high-level description that discards important visual cues. For example, a caption such as "a person playing football" captures the dominant activity while overlooking temporal progression, object interactions, and contextual details crucial for learning transferable representations. Since videos naturally contain multiple objects, actions, interactions, and evolving events, reducing them to a single description encourages models to learn coarse semantic correspondences while overlooking fine-grained dynamics and complementary relationships. This reveals a fundamental supervision bottleneck in current video-text pretraining: the limitation arises not only from insufficient information, but also from representing rich visual content through a single textual view. This naturally raises an important question: *can rich video content be effectively learned from a single textual view?* A video may admit multiple valid semantic interpretations, yet current supervision compresses these complementary signals into a single caption.

A natural solution is to leverage multimodal large language models (MLLMs) to generate richer supervision automatically. Recent studies have demonstrated the potential of large-scale recaptioning

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.

pipelines (Chen et al., 2024c; Zheng et al., 2024; Chen et al., 2024b;a). However, generating longer captions alone may not necessarily lead to better supervision, as such descriptions may emphasize irrelevant details, introduce incorrect information, or drift from the dominant visual content. Effective video-text supervision, therefore, requires signals that are not only richer, but also *diverse*, *visually faithful*, and *semantically comprehensive*.

In this work, we revisit video-text pretraining from the perspective of supervision quality and propose a fully automated MLLM-based framework for generating richer and more reliable textual supervision at scale. Given an input video, we generate summary and detailed descriptions using multiple instruction-tuned MLLMs, refine them through reasoning-based visual verification, and construct additional semantic positive captions capturing complementary perspectives. To effectively exploit supervision at different levels of granularity, we further introduce a granularity-aware text representation with separate `CLS` tokens for summary and detailed views, allowing them to be independently aligned with the video while learning complementary representations. Using this framework, we re-caption approximately 10 million videos to construct a large-scale video-text pretraining corpus. Extensive experiments across standard, fine-grained, and detailed video-text retrieval benchmarks demonstrate that **m**ulti-**v**iew **c**aptions (MVC) substantially improve video-text pretraining and provide a data-efficient alternative to simply scaling the number of video-text pairs.

2 Related Work

Video-Text Pretraining. Video-text pretraining has evolved from early approaches that learn aligned video-text representations through contrastive learning and cross-modal matching objectives (Bain et al., 2021; Lei et al., 2021; Xu et al., 2021). Building on these, subsequent works improved representation quality through larger-scale video-text datasets (Lei et al., 2023), stronger visual encoders and multimodal architectures (Wang et al., 2022a;b; Liu et al., 2022; Li et al., 2023b), scaling strategies (Wang et al., 2024c), and more effective masking (Zhuang et al., 2026; Wu et al., 2025). These developments have substantially advanced transfer performance across retrieval and video understanding tasks.

More recent efforts have further shifted toward large-scale video-text foundation models trained on millions or even billions of video-text pairs (Wang et al., 2024b; Chen et al., 2024c; Wang et al., 2024c). While these approaches largely improve performance through model and data scaling, relatively few works investigate the role of supervision design itself. Our work differs from existing approaches by focusing on the supervision source and exploring whether richer, multi-view textual supervision can provide stronger learning signals for video-text pretraining.

MLLM-Based Caption Generation. Several works have explored synthetic supervision to improve language quality by generating or rewriting captions using generative models and large language models, aiming to reduce noisy supervision and create more informative and grounded textual descriptions (Wang et al., 2024b; Yuan et al., 2025; Islam et al., 2024; Fan et al., 2023).

In the image-text domain, ShareGPT4V (Chen et al., 2024a) trains a captioning model on a curated set of image-caption pairs and uses it to generate large-scale detailed descriptions. SynthCLIP (Hammoud et al., 2024) creates synthetic image-text pairs with text-to-image generation while LaCLIP (Fan et al., 2023) rewrites existing captions as a form of textual augmentation. DreamLIP (Zheng et al., 2024) further demonstrates that recaptioning images with off-the-shelf MLLMs can significantly improve downstream performance by generating richer textual supervision. For the video domain, ShareGPT4Video (Chen et al., 2024b) extends large-scale recaptioning to videos by first collecting and annotating approximately 40K high-quality video-caption pairs, which are then used to train a video captioner for generating 4.8M detailed video-text pairs. InternVid (Wang et al., 2024b) constructs video annotations by generating frame-level captions using BLIP-2 (Li et al., 2023a) and subsequently summarizing them into short video-level descriptions with an LLM. Panda-70M (Chen et al., 2024c) further scales supervision generation by employing eight different MLLMs to generate caption candidates. However, a set of generated captions is verified and selected by human annotators before training an automatic caption-selection model for final data generation. Our framework is fully automated and generates multiple detailed and complementary textual views, providing richer supervision than short single-caption supervision.

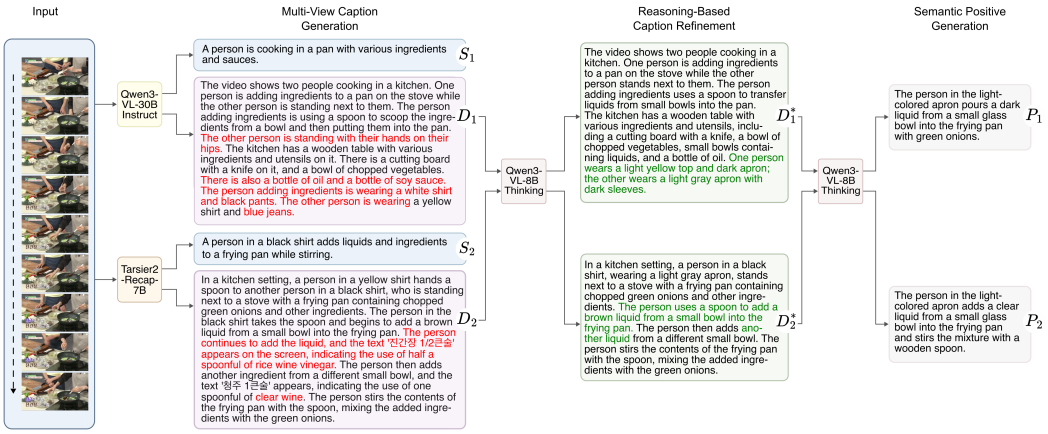

Figure 1: **Captioning pipeline.** Overview of our framework for improving supervision quality in video-text pretraining. Given an input video, we construct multiple complementary textual views, refine them through visual reasoning to improve grounding, and generate semantically consistent positives to enrich supervision. Red highlights the potentially incorrect segments in the detailed caption, while green represents the corrected segments after refinement. The resulting supervision corpus is used to learn richer video-text representations.

3 METHODOLOGY

3.1 MLLM-BASED MULTI-VIEW SUPERVISION GENERATION

Large-scale video-text pretraining critically depends on the quality of video-text supervision. Existing web-scale datasets typically associate each video with a single caption, which often captures only a limited aspect of the visual content and overlooks rich spatiotemporal semantics. While recent multimodal large language models (MLLMs) can generate more informative descriptions, they may introduce hallucinations and unsupported details. We therefore design our supervision generation framework around three complementary objectives: semantic diversity, visual fidelity, and semantic coverage. Specifically, we generate multiple complementary captions, refine them through visual verification, and create additional semantic positive captions that capture different aspects of the same video. Together, these stages produce a supervision corpus that is diverse, visually faithful, and semantically comprehensive.

Starting from InternVid-10M-FLT (Wang et al., 2024b), we construct a large-scale multi-view supervision corpus through three stages: (i) multi-view caption generation, (ii) reasoning-based caption refinement, and (iii) semantic positive caption generation. Given an input collection of videos $\mathcal{V} = \{v_i\}_{i=1}^{N}$ with $N \approx 10^7$, our goal is to generate multiple semantically complementary supervision signals for each video, as illustrated in Figure 1. We denote summary captions by S, detailed captions by D, refined detailed captions by D^*, and semantic positive captions by P. Each stage increases either the diversity, fidelity, or semantic coverage of the supervision.

3.1.1 MULTI-VIEW CAPTION GENERATION.

Different MLLMs naturally generate diverse descriptions of the same video. We aim to leverage such captions from different MLLMs as complementary semantic views, providing richer supervision than any individual caption alone. Let f_m denote the captioning function of model m. Given a video v_i, each model generates both a concise summary and a detailed description,

$$\begin{aligned} S_m &= f_m(v_i, q_s), \\ D_m &= f_m(v_i, q_d), \end{aligned} \tag{1}$$

where q_s prompts the model to summarize the primary activity, while q_d requests a richer description of visible objects, actions, interactions, and scene context.

We instantiate the caption generators using two pretrained video MLLMs, Tarsier2-Recap-7B (Yuan et al., 2025) and Qwen3-VL-30B-Instruct (Bai et al., 2025), producing two summary captions (S_1, S_2) and two detailed captions (D_1, D_2) for every video.

3.1.2 REASONING-BASED CAPTION REFINEMENT.

Detailed captions provide richer supervision but frequently contain unsupported actions and artifacts such as subtitles, text overlays, user-interface elements, or incorrect information. We therefore introduce a reasoning-capable refinement stage that treats generated captions as noisy supervision candidates. Rather than generating a new description from scratch, the model acts as a verifier checking consistency between textual and visual evidence.

Given a detailed caption D_m for $m \in \{1, 2\}$, we employ a reasoning-capable MLLM g based on Qwen3-VL-8B-Thinking (Bai et al., 2025):

$$D_m^* = g(v_i, D_m), \tag{2}$$

where the model jointly receives both the original video and the generated caption as input. We prompt the MLLM to preserve visually grounded content while removing unsupported details, potentially hallucinated actions, subtitles, watermarks, and non-visual inferences such as speech, intentions, or emotions. This stage, therefore, improves supervision fidelity while maintaining semantic richness. Although refinement substantially reduces incorrect information, some residual inaccuracies may still remain due to limitations of the underlying MLLM. For each video v_i, this stage generates two refined detailed captions D_1^* and D_2^* from D_1 and D_2, respectively.

3.1.3 SEMANTIC POSITIVE CAPTION GENERATION.

Although refined detailed captions provide stronger visual grounding, a single description still captures only one semantic interpretation of a complex video. Multiple valid descriptions may exist depending on the event, interaction, object, or temporal phase being emphasized. We therefore generate additional semantic positive captions that shift the semantic focus while preserving visual correctness.

Unlike conventional paraphrasing, which primarily changes the language of the same description, semantic positive captions intentionally describe different yet visually grounded aspects of the same video. One may emphasize the manipulated object, another the interaction, and another the resulting state, while all remain semantically consistent with the underlying visual content. We randomly select one refined detailed caption D_m^*, where $m \in \{1, 2\}$, and generate two semantic positive captions in one call:

$$(P_1, P_2) = h(v_i, D_m^*), \tag{3}$$

where h denotes Qwen3-VL-8B-Thinking (Bai et al., 2025), conditioned jointly on the video and the selected refined detailed caption. Each semantic positive caption explicitly describes the actor, action, interacting object, and resulting outcome while emphasizing a different grounded aspect whenever possible.

The final caption set for video v_i used for training is defined as:

$$\mathcal{C}(v_i) = \{S_1, S_2, D_1^*, D_2^*, P_1, P_2, O\}, \tag{4}$$

Here S_1 and S_2 are the summary captions, D_1^* and D_2^* are the refined detailed captions, P_1 and P_2 are the semantic positive captions, and O is the original InternVid-10M-FLT caption. The unrefined detailed captions D_1 and D_2 are intermediate outputs and are not included in the final training set. Detailed prompts, visual examples, and caption statistics are in the supplementary material.

3.2 VIDEO-TEXT PRETRAINING

We adopt a two-stage video-text pretraining framework following recent video-text models (Li et al., 2023b), while modifying the training pipeline to use the proposed multi-view supervision corpus.

Stage I: Video Representation Initialization. We initialize the visual encoder using SMILE (Thoker et al., 2025), a masked video representation learning framework that jointly captures spatial semantics and motion dynamics through masked reconstruction objectives. Compared with conventional masked video pretraining, SMILE provides stronger motion-aware representations, which are particularly beneficial for downstream video-text alignment.

Stage II: Video-Text Alignment. Starting from the pretrained visual encoder, we perform video-text pretraining using the supervision set $\mathcal{C}(v_i)$. We use two `CLS` tokens in the text encoder: one for

summary captions (including O) and one for refined detailed and semantic positive captions. During pretraining, we randomly sample for each video v_i one caption for each CLS token:

$$c_s \sim \{S_1, S_2, O\}, \quad c_d \sim \{D_1^*, D_2^*, P_1, P_2\}. \tag{5}$$

We encode the two captions separately, using the matching CLS token for each, and share the video representation between them. We train with three standard video-text objectives: video-text contrastive learning (VTC), video-text matching (VTM), and masked language modeling (MLM). For VTC and VTM, each caption is paired with the video using its corresponding CLS representation, and we apply MLM to both captions. For the summary view, the combined loss is

$$\mathcal{L}^s = \mathcal{L}^s_{\text{VTC}} + \mathcal{L}^s_{\text{VTM}} + \mathcal{L}^s_{\text{MLM}}. \tag{6}$$

The detailed-view loss $\mathcal{L}^d$ is computed in the same way, using the sampled refined detailed or semantic positive caption and its corresponding CLS token. We average the summary-view and detailed-view losses to obtain the overall objective:

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}^s + \mathcal{L}^d\right). \tag{7}$$

During zero-shot text-to-video retrieval, we encode each text query twice, once with the summary CLS token and once with the detailed CLS token. We compute each text representation's similarity to every candidate video and average the two similarity scores for each text-video pair. We then rank the videos by the averaged scores, following the UMT evaluation protocol.

Compared with conventional video-text pretraining pipelines that rely on a single caption per video, our framework exposes the model to multiple semantically consistent yet complementary textual views, enabling stronger alignment between visual content and language semantics.

4 Experiments

Implementation Details: For Stage I, we use publicly available SMILE (Thoker et al., 2025) checkpoints with ViT-B and ViT-L backbones pretrained on Kinetics-400 and Kinetics-700 (Kay et al., 2017), respectively. For Stage II, we initialize the video encoder from Stage I and use pretrained BERT-base/BERT-large as the text encoder following (Li et al., 2023b). We pretrain on 5M or 10M videos from InternVid-10M-FLT (Wang et al., 2024b), as indicated in each results table, using the supervision set $\mathcal{C}(v_i)$ defined in Section 3. Each video is represented using 16 randomly sampled frames of resolution 224×224 with a video token masking ratio of 10%, while captions are truncated to 128 tokens. Training is conducted on 16 GPUs with batch size 128 per GPU for 20 epochs using AdamW ($lr = 10^{-4}$, weight decay 0.02). We adopt cosine decay with one warmup epoch and a final learning rate of 1% of the initial value. The pretraining objective combines VTC, VTM, and MLM with equal weights, where VTM employs hard negative mining, VTC uses a temperature of 0.07, and MLM applies a token masking ratio of 0.5. During pretraining, we sample one summary caption from $\{O, S_1, S_2\}$ and one refined detailed or semantic positive caption from $\{D_1^*, D_2^*, P_1, P_2\}$, and encode them using their corresponding summary-view and detailed-view CLS tokens. Unless otherwise specified, all main results use this dual-CLS configuration.

4.1 Standard Text-to-Video Retrieval

Datasets. We evaluate on five standard video-text retrieval benchmarks: MSR-VTT (Xu et al., 2016), DiDeMo (Hendricks et al., 2017), ActivityNet (Caba Heilbron et al., 2015), LSMDC (Rohrbach et al., 2015), and MSVD (Chen & Dolan, 2011), following prior works. MSR-VTT and MSVD focus on general semantic alignment over open-domain videos with diverse activities and scenes. DiDeMo and ActivityNet place greater emphasis on temporal reasoning and activity understanding, requiring the model to capture dynamically evolving content and long-range temporal dependencies. LSMDC presents a more challenging movie-based setting with complex narratives and richer contextual semantics. We strictly follow (Li et al., 2023b) for both zero-shot and fine-tuning evaluation. We report text-to-video retrieval performance using Recall (R@1).

Zero-Shot Results. Table 1 presents zero-shot text-to-video retrieval results across five benchmarks. Under the 5M setting, our model outperforms prior approaches trained with comparable amounts of data. Compared with UMT-B, MVC-B improves by +8.6 on MSR-VTT, +25.5 on DiDeMo, and +30.3 on ActivityNet, while MVC-L similarly surpasses UMT-L by +27.2 on DiDeMo and +32.0 on ActivityNet. Compared with UMT pretrained on Panda-5M using captions generated

Table 1: **Zero-shot text-to-video retrieval.** We evaluate our method on MSR-VTT, DiDeMo, ActivityNet, LSMDC, and MSVD using Recall@1. Img3M = CC3M, Img15M = CC12M + SBU + COCO + VG. Within each pretraining-scale group, bold and underlined numbers represent the best and second-best models, respectively.

Method	#Samples	Source	MSR-VTT R@1	DiDeMo R@1	ActivityNet R@1	LSMDC R@1	MSVD R@1
Frozen (Bain et al., 2021)	5M	WebVid-2M+Img3M	18.7	20.2	–	–	–
Singularity (Lei et al., 2023)	5M	WebVid-2M+Img3M	28.4	36.9	30.8	–	–
UMT-B (Li et al., 2023b)	5M	WebVid-2M+Img3M	29.6	33.4	28.3	16.8	36.2
STM-B (Wu et al., 2025)	5M	WebVid-2M+Img3M	29.8	34.4	30.6	–	38.7
ClusterSTM-B (Zhuang et al., 2026)	5M	WebVid-2M+Img3M	31.2	36.5	31.4	–	40.3
UMT-L (Li et al., 2023b)[1]	5M	WebVid-2M+Img3M	33.3	34.0	31.9	<u>20.0</u>	<u>44.4</u>
UMT-L(Panda-5M) (Chen et al., 2024c)	5M	Panda-5M	<u>37.2</u>	<u>34.2</u>	–	–	<u>37.2</u>
Baseline-B	5M	InternVid-10M-FLT (original)	34.2	32.7	29.6	11.1	39.1
MVC-B (Ours)	5M	InternVid-10M-FLT + MVC	<u>38.2</u>	<u>58.9</u>	<u>58.6</u>	18.1	43.3
MVC-L (Ours)	5M	InternVid-10M-FLT + MVC	**40.6**	**61.2**	**63.9**	**21.0**	**48.4**
VIOLET (Fu et al., 2021)	138M	HowTo100M	25.9	23.5	–	–	–
Singularity (Lei et al., 2023)	17M	WebVid-2M+Img15M	34.0	37.1	30.6	–	–
OmniVL (Wang et al., 2022a)	17M	WebVid-2M+Img15M	34.6	33.3	–	–	–
InternVideo (Wang et al., 2022b)	646M	-	40.7	31.5	30.7	17.6	43.4
VideoCoCa (Yan et al., 2022)	4.8B	ALIGN+ALT200M	34.3	–	34.5	–	–
UMT-B (Li et al., 2023b)	17M	WebVid-2M+Img15M	35.5	41.9	33.8	18.1	41.4
UMT-L (Li et al., 2023b)	17M	WebVid-2M+Img15M	**42.6**	46.4	42.8	**25.2**	**49.9**
CLIP (Radford et al., 2021)	1B	DataComp-1B	30.4	12.7	9.1	13.9	40.5
ViCLIP-L (Wang et al., 2024b)	410M	CLIP-400M+InternVid-10M-FLT	<u>42.4</u>	18.4	15.1	20.1	<u>49.1</u>
MVC-B (Ours)	10M	InternVid-10M-FLT + MVC	39.0	<u>61.0</u>	<u>61.0</u>	18.6	44.2
MVC-L (Ours)	10M	InternVid-10M-FLT + MVC	41.1	**62.0**	**66.9**	<u>24.1</u>	49.0

by multiple cross-modality teachers, our approach achieves substantially stronger performance on MSR-VTT, DiDeMo, and MSVD using the same pretraining scale. These results suggest that richer supervision quality is more effective for learning video-text representations.

More importantly, Baseline-B provides a controlled comparison that isolates the effect of our proposed supervision. Baseline-B uses the same backbone (with single-`CLS`), training framework, and 5M InternVid videos as MVC-B, but is trained with only the original captions O. Replacing the original supervision with MVC improves R@1 from 34.2 to 38.2 on MSR-VTT, 32.7 to 58.9 on DiDeMo, 29.6 to 58.6 on ActivityNet, 11.1 to 18.1 on LSMDC, and 39.1 to 43.3 on MSVD. The largest gains occur on DiDeMo (+26.2) and ActivityNet (+29.0), where retrieval requires matching videos to comparatively detailed descriptions of events and activities. This controlled comparison shows that the improvements arise substantially from enriching the textual supervision, rather than simply increasing the number of pretraining videos.

We further compare with methods pretrained on substantially larger corpora, ranging from 17M to 646M samples. Despite using only 10M videos, MVC-L remains competitive across standard retrieval benchmarks and achieves particularly strong performance on DiDeMo and ActivityNet. Compared with UMT-L pretrained on 17M videos, MVC-L improves R@1 by 15.6 points on DiDeMo and 24.1 points on ActivityNet while using fewer pretraining videos. Notably, MVC-L reaches 62.0 R@1 on DiDeMo and 66.9 on ActivityNet, compared with 31.5 and 30.7 for InternVideo trained on 646M samples. Together with the controlled Baseline-B comparison, these results suggest that richer multi-view supervision can achieve strong performance with substantially fewer pretraining videos.

Fine-Tuned Results. Table 2 reports text-to-video retrieval after end-to-end fine-tuning. The improvements observed under zero-shot evaluation largely persist after task-specific adaptation. At the 5M scale, MVC-L substantially outperforms UMT-L on DiDeMo (+21.6), ActivityNet (+14.0), and LSMDC (+5.5), while also improving MSR-VTT (+2.2).

The controlled comparison with Baseline-B further shows that the benefits of our supervision persist after fine-tuning. With the same architecture and pretraining videos, MVC-B improves R@1 from 48.0 to 50.3 on MSR-VTT, 61.2 to 72.9 on DiDeMo, 55.7 to 64.4 on ActivityNet, 31.1 to 34.1 on LSMDC, and 45.0 to 46.0 on MSVD. Thus, the richer representations learned from multi-view supervision remain beneficial even after adaptation to individual downstream datasets.

[1] We report the corrected values from the official GitHub repository.

Table 2: **Fine-Tuned Text-to-video retrieval.** We evaluate our method on MSR-VTT, DiDeMo, ActivityNet, LSMDC, and MSVD. Img3M = CC3M Img15M = CC12M + SBU + COCO + VG CC = Conceptual Captions, VG = Visual Genome. Within each pretraining-scale group, bold and underlined numbers represent the best and second-best models, respectively.

Method	#Samples	Source	MSR-VTT R@1	DiDeMo R@1	ActivityNet R@1	LSMDC R@1	MSVD R@1
ClipBERT (Lei et al., 2021)	5.4M	COCO+VG	22.0	20.4	21.3	–	–
Frozen (Bain et al., 2021)	5M	WebVid-2M+Img3M	31.0	34.6	–	15.0	33.7
UMT-B (Li et al., 2023b)	5M	WebVid-2M+Img3M	46.3	54.8	52.1	30.3	47.4
STM-B (Wu et al., 2025)	5M	WebVid-2M+Img3M	48.5	56.9	53.6	–	–
ClusterSTM-B (Zhuang et al., 2026)	5M	WebVid-2M+Img3M	49.7	58.5	54.9	–	–
UMT-L (Li et al., 2023b)	5M	WebVid-2M+Img3M	53.3	59.7	58.1	<u>37.7</u>	<u>53.7</u>
UMT (Panda-5M) (Li et al., 2023b; Chen et al., 2024c)	5M	Panda-5M	**58.4**	60.6	–	–	57.5
Baseline-B	5M	InternVid-10M-FLT (original)	48.0	61.2	55.7	31.1	45.0
MVC-B (Ours)	5M	InternVid-10M-FLT + MVC	50.3	<u>72.9</u>	<u>64.4</u>	34.1	46.0
MVC-L (Ours)	5M	InternVid-10M-FLT + MVC	<u>55.5</u>	**81.3**	**72.1**	**43.2**	53.8
VIOLET (Fu et al., 2021)	138M	HowTo100M	34.5	32.6	–	16.1	–
All-in-one (Wang et al., 2023)	138M	WebVid-2M+HowTo100M	37.9	32.7	22.4	–	–
CLIP (Radford et al., 2021; Wang et al., 2024b)	400M	CLIP-400M	38.2	32.2	26.1	22.5	67.3
LAVENDER (Li et al., 2023c)	30M	WebVid-2M+Vid13M+Img15M	40.7	53.4	–	26.1	50.1
Singularity (Lei et al., 2023)	17M	WebVid-2M+Img15M	42.7	53.1	48.9	–	–
VINDLU (Cheng et al., 2023)	25M	WebVid-10M+Img15M	46.5	61.2	55.0	–	–
HiTeA (Ye et al., 2023)	17M	WebVid-2M+Img15M	46.8	56.5	49.7	28.7	–
OmniVL (Wang et al., 2022a)	17M	WebVid-2M+Img15M	47.8	52.4	–	–	–
UMT-B (Li et al., 2023b)	17M	WebVid-2M+Img15M	50.6	60.8	56.1	32.3	70.8
UMT-B (Li et al., 2023b)	25M	WebVid-2M+Img15M	51.0	61.6	58.3	32.7	<u>71.9</u>
ViCLIP-L (Wang et al., 2024b)	410M	CLIP-400M+InternVid-10M-FLT	52.5	49.4	49.8	33.0	53.1
CLIP-ViP (Xue et al., 2022)	500M	CLIP-400M+HD-VILA-100M	54.2	50.5	53.4	29.4	–
InternVideo (Wang et al., 2022b)	646M	-	55.2	57.9	62.2	34.0	**58.4**
UMT-L (Li et al., 2023b)	17M	WebVid-2M+Img15M	<u>56.5</u>	<u>66.6</u>	<u>66.6</u>	<u>41.4</u>	57.4
MVC-B (Ours)	10M	InternVid-10M-FLT + MVC	52.8	<u>74.8</u>	65.2	35.2	50.1
MVC-L (Ours)	10M	InternVid-10M-FLT + MVC	**58.1**	**82.9**	**75.2**	**45.6**	55.8

The advantage also extends to the 10M setup. MVC-L achieves 58.1, 82.9, 75.2, and 45.6 R@1 on MSR-VTT, DiDeMo, ActivityNet, and LSMDC, respectively, outperforming UMT-L pretrained on 17M videos by 1.6, 16.3, 8.6, and 4.2 points. Similarly, compared with ViCLIP-L, which also incorporates InternVid-10M-FLT in its substantially larger pretraining corpus, MVC-L improves DiDeMo from 49.4 to 82.9 (+33.5) and ActivityNet from 49.8 to 75.2 (+25.4). These results further highlight the data efficiency of our approach, with particularly strong gains on benchmarks requiring retrieval from richer event and activity descriptions.

4.2 Fine-Grained and Detailed Text-to-Video Retrieval

Datasets. Standard text-to-video retrieval benchmarks mainly evaluate alignment with summary-level descriptions, which may not adequately assess fine-grained visual understanding, temporal reasoning, or long-form semantics. Since our framework explicitly improves supervision quality through detailed, faithful, and diverse captions, we additionally evaluate zero-shot retrieval on CaReBench (Xu et al., 2025), DREAM-1K (Wang et al., 2024a), and Shot2Story20K (Han et al., 2023), which require richer text-to-video alignment. CaReBench evaluates spatial and temporal understanding through separate spatial and temporal retrieval settings. DREAM-1K focuses on fine-grained actions and event progressions, where we evaluate both detailed-description and event-based retrieval. Shot2Story20K consists of videos with multiple shots. We evaluate whole-clip retrieval and single-shot retrieval to assess long-form narrative semantics and local event understanding. We report text-to-video retrieval using Recall (R@1).

Results. Table 3 presents zero-shot retrieval results on fine-grained and detailed benchmarks. The direct comparison between Baseline-B and MVC-B reveals particularly large gains in these more challenging retrieval settings. Using the same 5M videos, MVC improves R@1 by 38.2 points on CARE-S, 30.5 on CARE-T, 34.6 on DREAM-D, 10.6 on DREAM-E, 32.0 on S2S-W, and 29.1 on S2S-S. The consistent improvements across spatial, temporal, detailed-description, event-level, and multi-shot retrieval demonstrate the effectiveness of our framework for learning richer video-text correspondences. We separately isolate the contributions of multi-view supervision and granularity-aware text representations in Section 4.3.

The gains remain substantial compared with models trained on considerably more data. MVC-B pretrained on a 5M video subset outperforms UMT-L pretrained on 25M videos across all six settings, by 7.9 points on CARE-S, 13.0 on CARE-T, 11.0 on DREAM-D, 2.4 on DREAM-E, 18.8 on S2S-W, and 8.4 on S2S-S, despite using a smaller backbone and fewer videos. Scaling to MVC-L at the same 5M video scale further increases these margins to 9.4, 19.7, 12.6, 6.4, 20.4, and 11.3

Table 3: **Fine-Grained and Detailed Zero-Shot Text-to-Video Retrieval.** We evaluate on CaReBench Spatial (CARE-S), CaReBench Temporal (CARE-T), DREAM-1K-Detailed (DREAM-D), DREAM-1K-Events (DREAM-E), and Shot2Story whole-clip (S2S-W) and single-shot (S2S-S) splits. Source labels use IV for InternVid and SG4V for ShareGPT4V. For prior methods, we use officially public checkpoints. Bold and underlined numbers mark the best and second-best results.

Method	#Samples	Source	CARE-S	CARE-T	DREAM-D	DREAM-E	S2S-W	S2S-S
			R@1	R@1	R@1	R@1	R@1	R@1
CLIP-B/16 (Radford et al., 2021)	400M	CLIP-400M	45.6	30.3	32.6	13.6	10.7	22.7
CLIP-L/14 (Radford et al., 2021)	400M	CLIP-400M	49.1	33.5	44.3	14.6	65.8	45.4
ViCLIP-B (Wang et al., 2024b)	400M+ 10M	CLIP-400M + IV-10M-FLT	56.0	31.3	54.9	10.4	52.2	43.9
ViCLIP-L (Wang et al., 2024b)	400M + 10M	CLIP-400M + IV-10M-FLT	55.7	34.5	55.5	23.3	55.1	43.6
Long-CLIP-L/14 (Zhang et al., 2024)	400M	CLIP-400M + SG4V-1M	65.6	33.3	58.3	24.3	74.7	47.7
UMT-B (Li et al., 2023b)	5M	WebVid-2M + Img3M	64.2	36.5	70.0	18.0	64.6	45.8
UMT-L (Li et al., 2023b)	5M	WebVid-2M + Img3M	68.0	39.6	75.0	21.1	59.2	47.3
UMT-B (Li et al., 2023b)	25M	WebVid-10M + Img15M	78.0	38.1	75.0	23.0	71.6	59.1
UMT-L (Li et al., 2023b)	25M	WebVid-10M + Img15M	81.5	47.1	83.1	26.7	75.6	65.1
Baseline-B	5M	IV-10M-FLT (original)	51.2	29.6	59.5	18.5	62.4	44.4
MVC-B (Ours)	5M	IV-10M-FLT + MVC	89.4	60.1	94.1	29.1	94.4	73.5
MVC-B (Ours)	10M	IV-10M-FLT + MVC	90.6	60.9	94.5	30.1	95.2	75.0
MVC-L (Ours)	5M	IV-10M-FLT + MVC	<u>90.9</u>	<u>66.8</u>	<u>95.7</u>	<u>33.1</u>	<u>96.0</u>	<u>76.4</u>
MVC-L (Ours)	10M	IV-10M-FLT + MVC	**92.9**	**67.4**	**96.4**	**33.6**	**96.4**	**77.3**

Table 4: **Comparison of supervision recipes.** We compare the original captions (O), our summary captions (S), our full generated supervision ($S + D^* + P$), and its combination with the original captions ($O + S + D^* + P$). We report text-to-video R@1 across six downstream benchmarks.

Supervision	MSVD	ANet	DiDeMo	LSMDC	CARE-T	Shot2Story-S
Original only (O)	35.4	25.9	29.3	11.8	27.1	42.2
Ours: summary only (S)	37.0	46.2	46.0	15.1	48.0	63.2
Ours: full ($S + D^* + P$)	38.5	**50.1**	52.4	16.7	**54.1**	67.4
Original + ours ($O + S + D^* + P$)	**38.8**	**50.1**	**53.0**	**18.2**	**54.1**	**67.9**

points, respectively. These results show that the benefits of our framework become particularly pronounced when retrieval requires finer-grained or more descriptive video-text alignment.

This trend also holds against models pretrained on substantially larger corpora. Compared with ViCLIP-L, which additionally leverages CLIP-400M pretraining, MVC-L trained on 10M videos improves R@1 by 37.2 points on CARE-S, 32.9 on CARE-T, 40.9 on DREAM-D, 10.3 on DREAM-E, 41.3 on S2S-W, and 33.7 on S2S-S. MVC-L also consistently outperforms Long-CLIP-L across all six settings despite Long-CLIP's specialized long-text modeling. Together, these results support the complementary contributions of caption granularity, reasoning-based refinement, and semantic positive captions to the effectiveness of our supervision pipeline.

4.3 Ablation Studies

For computational efficiency, we conduct all ablations on a 1M-video subset of our pretraining corpus, use a video masking ratio of 50%, while keeping the rest of the architecture, optimization, and training schedule fixed unless otherwise specified. We report zero-shot text-to-video R@1 on MSVD, ActivityNet, DiDeMo, LSMDC, CARE-T, and Shot2Story-S. The supervision ablations in Tables 4 and 5 use a standard single-`CLS` representation, while Table 6 separately evaluates our dual-`CLS` design under identical supervision.

Effect of Multi-View Supervision. Table 4 studies the effect of progressively enriching the textual supervision. Replacing the original captions (O) with our summary captions (S) improves performance across all six benchmarks, with particularly large gains on ActivityNet (25.9→46.2), DiDeMo (29.3→46.0), CARE-T (27.1→48.0), and Shot2Story-S (42.2→63.2). This shows that summary captions generated from complementary MLLMs provide substantially more effective supervision than the original captions. Enriching these summary captions with refined detailed captions and semantic positive captions ($S + D^* + P$) further improves all six benchmarks, reaching 50.1 on ActivityNet, 52.4 on DiDeMo, 54.1 on CARE-T, and 67.4 on Shot2Story-S. Finally, retaining the original caption alongside our generated supervision ($O + S + D^* + P$) yields the best or equal-best performance across all benchmarks, including a notable improvement from 16.7 to 18.2 on LSMDC. These results suggest that the multi-view captions provide rich complementary supervision compared to the single-view captions across diverse downstream settings.

Table 5: **Impact of caption complementarity, refinement, and semantic positive captions.** We compare summary captions (S), detailed captions (D), their combination, refined detailed captions (D^*), and semantic positive captions (P). Original captions (O) are excluded throughout. We report text-to-video R@1.

Supervision	MSVD	ANet	DiDeMo	LSMDC	CARE-T	Shot2Story-S
Summary only (S)	37.0	46.2	46.0	15.1	48.0	63.2
Detailed only (D)	26.4	42.7	38.4	12.4	41.1	62.3
Summary+detailed ($S + D$)	36.6	48.5	48.6	15.9	50.1	65.9
Summary+refined detailed ($S + D^*$)	37.5	49.6	49.7	16.1	51.8	66.7
Summary+refined+positive ($S + D^* + P$)	**38.5**	**50.1**	**52.4**	**16.7**	**54.1**	**67.4**

Table 6: **Impact of granularity-aware text representations.** We compare a standard single-`CLS` representation with our dual-`CLS` design that separately models summary and detailed views. The detailed view includes refined detailed and semantic positive captions. Both use the same supervision ($O + S + D^* + P$) and training setup. We report text-to-video R@1.

Text Representation	MSVD	ANet	DiDeMo	LSMDC	CARE-T	Shot2Story-S
Single `CLS`	38.8	50.1	53.0	18.2	54.1	67.9
Dual `CLS` (Ours)	**39.4**	**51.3**	**54.3**	**18.7**	**56.4**	**68.5**

Contribution of Supervision Components. Table 5 isolates the contributions of caption complementarity, refinement, and semantic positive captions. Detailed captions alone (D) perform worse than summary captions (S) across all benchmarks, indicating that detailed captions do not replace summary captions when used in isolation. Combining the two granularities ($S + D$), however, improves over S alone on five of the six benchmarks, including ActivityNet (46.2→48.5), DiDeMo (46.0→48.6), and CARE-T (48.0→50.1), demonstrating their complementary nature. Replacing the unrefined detailed captions with refined detailed captions ($S + D^*$) improves performance consistently across all six benchmarks, supporting the benefit of reasoning-based refinement. Adding semantic positive captions further improves every benchmark, with particularly clear gains on DiDeMo (49.7→52.4) and CARE-T (51.8→54.1). Together, these results validate the three objectives of our supervision pipeline: complementary caption granularities provide diverse semantic views, refinement improves visual fidelity, and semantic positive captions expand the semantic coverage of the supervision.

Granularity-Aware Text Representations. Finally, Table 6 evaluates whether summary captions and refined detailed or semantic positive captions benefit from separate text representations. Both variants use identical $O + S + D^* + P$ supervision, isolating the effect of the representation design. The dual-`CLS` model consistently outperforms the standard single-`CLS` representation across all six benchmarks, improving ActivityNet from 50.1 to 51.3, DiDeMo from 53.0 to 54.3, and CARE-T from 54.1 to 56.4, with the largest gain of 2.3 points on CARE-T. These results show that explicitly modeling the summary and detailed views with separate representations provides an additional benefit beyond multi-view supervision alone, motivating the dual-`CLS` design used in our final model.

5 Conclusion

We revisit video-text pretraining from the perspective of supervision quality and propose a large-scale MLLM-based multi-view supervision framework for generating richer video-text annotations. By combining multi-view caption generation, reasoning-based refinement, and semantic positive generation, our approach improves supervision diversity, visual fidelity, and semantic coverage beyond conventional single-caption annotations. We further introduce a granularity-aware text representation that separately models short and detailed descriptions, enabling the model to better exploit their complementary supervision. Extensive experiments across standard, fine-grained, and detailed video retrieval benchmarks demonstrate consistent gains in both zero-shot and fine-tuned settings, often surpassing approaches trained with larger datasets. Our results highlight the supervision quality and granularity as important dimensions of scalable video-text learning, demonstrating that richer textual supervision can significantly improve the data efficiency of video-text pretraining.

ETHICS STATEMENT

Our work uses the publicly available InternVid-10M-FLT dataset and collects no new videos, personal data, or human annotations. The textual supervision is generated automatically using existing MLLMs, and no crowdworkers or external annotators are employed. As InternVid originates from web-scale video data, its videos may inherit biases or potentially sensitive content from their original sources. Generated captions may similarly inherit biases or residual inaccuracies from the MLLMs, despite our reasoning-based refinement stage for improving visual grounding. We do not use the generated annotations to identify individuals or infer sensitive personal attributes. The resulting models and experiments are intended for research on video-text representation learning and retrieval, not for surveillance, identification, or high-stakes decision making. We use the source data and pretrained models according to their respective licenses and terms of use.

AI USE STATEMENT

In this work, we used AI tools to support research methodology and experimental design, software development and debugging. We also used generative AI tools for brainstorming, information search, and to polish and standardize the writing of the manuscript. Generative AI tools were not used to independently interpret experimental results or formulate the scientific conclusions of this work. All AI-assisted outputs, including code, analyses, and manuscript text, were reviewed and verified by the authors. The authors take full responsibility for the methodology, experiments, results, and content of the manuscript. We also used MLLMs for data processing and the generation and refinement of video captions, as described in Section 3.

REPRODUCIBILITY

Sections 3 and 4 describe the caption-generation pipeline, training setup, and evaluation protocol. The Appendix A provides the captioning prompts, dataset details, caption statistics, and fine-tuning hyperparameters. Our implementation builds on the public UMT codebase (`https://github.com/OpenGVLab/unmasked_teacher/tree/main/multi_modality`). Upon acceptance, we will release our training and caption-generation code, and generated captions indexed by InternVid-10M-FLT videos. The released captions will allow the training experiments to be reproduced without rerunning the computationally intensive caption-generation pipeline.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv*, abs/2511.21631, 2025. doi: 10.48550/arXiv.2511.21631. URL https://doi.org/10.48550/arXiv.2511.21631.

Max Bain, Arsha Nagrani, Gul Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 1708–1718, Montréal, Can., Oct. 2021. IEEE. doi: 10.1109/iccv48922.2021.00175. URL https://doi.org/10.1109/ICCV48922.2021.00175.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 961–970, Boston, MA, USA, Jun. 2015. IEEE. doi: 10.1109/CVPR.2015.7298698. URL https://doi.org/10.1109/CVPR.2015.7298698.

David Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proc. Conf. North Am. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, pp. 190–200, Portland, OR, USA, 2011. URL https://aclanthology.org/P11-1020/.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving large multi-modal models with better captions. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 15075 of *Lect. Notes Comput. Sci.*, pp. 370–387. Springer Nat. Switz., 2024a. doi: 10.1007/978-3-031-72643-9_22. URL https://doi.org/10.1007/978-3-031-72643-9_22.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. ShareGPT4Video: Improving video understanding and generation with better captions. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 37, pp. 19472–19495, Vancouver, Can., Dec. 2024b. Curran Assoc. Inc.

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 *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 13320–13331. IEEE, Jun. 2024c. doi: 10.1109/cvpr52733.2024.01265. URL https://doi.org/10.1109/cvpr52733.2024.01265.

Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. VindLU: A recipe for effective video-and-language pretraining. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 10739–10750, Vancouver, Can., Jun. 2023. IEEE. doi: 10.1109/cvpr52729.2023.01034. URL https://doi.org/10.1109/CVPR52729.2023.01034.

Hazel Doughty, Fida Mohammad Thoker, and Cees GM Snoek. Locomotion: Learning motion-focused video-language representations. In *Asian Conference on Computer Vision*, pp. 3–24. Springer, 2024.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving CLIP training with language rewrites. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 36, pp. 35544–35575, New Orleans, LA, USA, Dec. 2023. Curran Assoc. Inc. URL https://openreview.net/forum?id=SVjDiiVySh.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. VIOLET : End-to-end video-language transformers with masked visual-token modeling. *arXiv*,

abs/2111.12681, 2021. doi: 10.48550/arXiv.2111.12681. URL https://doi.org/10.48550/arXiv.2111.12681.

Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. SynthCLIP: Are we ready for a fully synthetic CLIP training? *arXiv*, abs/2402.01832, 2024. doi: 10.48550/arXiv.2402.01832. URL https://doi.org/10.48550/arXiv.2402.01832.

Mingfei Han, Linjie Yang, Xiaojun Chang, Lina Yao, and Heng Wang. Shot2Story: A new benchmark for comprehensive understanding of multi-shot videos. *arXiv*, abs/2312.10300, 2023. doi: 10.48550/arXiv.2312.10300. URL https://doi.org/10.48550/arXiv.2312.10300.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 5804–5813, Venice, Italy, Oct. 2017. IEEE. doi: 10.1109/iccv.2017.618. URL https://doi.org/10.1109/ICCV.2017.618.

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video ReCap: Recursive captioning of hour-long videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 18198–18208, Seattle, WA, USA, Jun. 2024. doi: 10.1109/cvpr52733.2024.01723. URL https://doi.org/10.1109/CVPR52733.2024.01723.

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 *Int. Conf. Mach. Learn. (ICML)*, volume 139 of *Proc. Mach. Learn. Res.*, pp. 4904–4916, 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv*, abs/1705.06950, 2017. doi: 10.48550/arXiv.1705.06950. URL https://doi.org/10.48550/arXiv.1705.06950.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: CLIPBERT for video-and-language learning via sparse sampling. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 7327–7337, Jun. 2021. doi: 10.1109/cvpr46437.2021.00725. URL https://doi.org/10.1109/CVPR46437.2021.00725.

Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Proc. Annu. Meet. Assoc. Comput. Linguistics*, pp. 487–507, 2023. doi: 10.18653/v1/2023.acl-long.29. URL https://doi.org/10.18653/v1/2023.acl-long.29.

Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven C.H. Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4943–4953, New Orleans, LA, USA, Jun. 2022a. doi: 10.1109/cvpr52688.2022.00490. URL https://doi.org/10.1109/CVPR52688.2022.00490.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 34, pp. 9694–9705, Virtual conference, Dec. 2021. Curran Assoc. Inc. URL https://openreview.net/forum?id=OJLaKwiXSbx).

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Int. Conf. Mach. Learn. (ICML)*, volume 162, pp. 12888–12900, Baltimore, Maryland USA, Jul. 2022b.

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 *Int. Conf. Mach. Learn. (ICML)*, volume 202, pp. 19730–19742, Honolulu, HI, USA, Jul. 2023a.

Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 19891–19903, Paris, Fr., Oct. 2023b. Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/iccv51070.2023.01826. URL https://doi.org/10.1109/ICCV51070.2023.01826.

Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. LAVENDER: Unifying video-language understanding as masked language modeling. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 23119–23129, Jun. 2023c. doi: 10.1109/cvpr52729.2023.02214. URL https://doi.org/10.1109/CVPR52729.2023.02214.

Ye Liu, Siyuan Li, Yang Wu, Chang Wen Chen, Ying Shan, and Xiaohu Qie. UMT: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3032–3041, New Orleans, LA, USA, Jun. 2022. IEEE. doi: 10.1109/cvpr52688.2022.00305. URL https://doi.org/10.1109/CVPR52688.2022.00305.

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 *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*. IEEE, Oct. 2019. doi: 10.1109/iccv.2019.00272. URL https://doi.org/10.1109/iccv.2019.00272.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn. (ICML)*, volume 139 of *Proc. Mach. Learn. Res.*, pp. 8748–8763, Virtual Conf., Jul. 2021. ML Res. Press. URL https://proceedings.mlr.press/v139/radford21a.html.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3202–3212, USA, Jun. 2015. doi: 10.1109/cvpr.2015.7298940. URL https://doi.org/10.1109/cvpr.2015.7298940.

Fida Mohammad Thoker, Letian Jiang, Chen Zhao, and Bernard Ghanem. SMILE: Infusing spatial and motion semantics in masked video learning. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 8438–8449, Nashville, TN, USA, Jun. 2025. IEEE. doi: 10.1109/cvpr52734.2025.00790. URL https://doi.org/10.1109/CVPR52734.2025.00790.

Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv*, abs/2407.00634, 2024a. doi: 10.48550/arXiv.2407.00634. URL https://doi.org/10.48550/arXiv.2407.00634.

Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023. doi: 10.1109/cvpr52729.2023.00638. URL https://doi.org/10.1109/cvpr52729.2023.00638.

Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. OmniVL:one foundation model for image-language and video-language tasks. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 35, pp. 5696–5710, New Orleans, LA, USA, Nov. 2022a.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. InternVideo: General video foundation models via generative and discriminative learning. *arXiv*, abs/2212.03191, 2022b. doi: 10.48550/arXiv.2212.03191. URL https://doi.org/10.48550/arXiv.2212.03191.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. InternVid: A

large-scale video-text dataset for multimodal understanding and generation. In *Int. Conf. Learn. Represent. (ICLR)*, pp. 1–25, Vienna, Austria, May 2024b. URL https://openreview.net/forum?id=MLBdiWu4Fw.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 15143 of *Lect. Notes Comput. Sci.*, pp. 396–416. Springer Nat. Switz., Nov. 2024c. doi: 10.1007/978-3-031-73013-9_23. URL https://doi.org/10.1007/978-3-031-73013-9_23.

Yue Wu, Zhaobo Qi, Junshu Sun, Yaowei Wang, Qingming Huang, and Shuhui Wang. Video language model pretraining with spatio-temporal masking. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 8557–8567, Nashville, TN, USA, Jun. 2025. IEEE. doi: 10.1109/cvpr52734.2025.00800. URL https://doi.org/10.1109/cvpr52734.2025.00800.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Conf. Empir. Methods Nat. Lang. Process.*, pp. 6787–6800. Assoc. Comput. Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.544. URL https://doi.org/10.18653/v1/2021.emnlp-main.544.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 5288–5296, Las Vegas, NV, USA, Jun. 2016. IEEE. doi: 10.1109/cvpr.2016.571. URL https://doi.org/10.1109/cvpr.2016.571.

Yifan Xu, Xinhao Li, Yichun Yang, Desen Meng, Rui Huang, and Limin Wang. CaReBench: A fine-grained benchmark for video captioning and retrieval. *arXiv*, abs/2501.00513, 2025. doi: 10.48550/arXiv.2501.00513. URL https://doi.org/10.48550/arXiv.2501.00513.

Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. CLIP-ViP: Adapting pre-trained image-text model to video-language representation alignment. *arXiv*, 2022. doi: 10.48550/arXiv.2209.06430. URL https://doi.org/10.48550/arXiv.2209.06430.

Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. VideoCoCa: Video-text modeling with zero-shot transfer from contrastive captioners. *arXiv*, 2022. doi: 10.48550/arXiv.2212.04979. URL https://doi.org/10.48550/arXiv.2212.04979.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 1666–1677, Montréal, Can., Oct. 2021. IEEE. doi: 10.1109/iccv48922.2021.00171. URL https://doi.org/10.1109/iccv48922.2021.00171.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 35, pp. 124–141, New Orleans, LA, USA, Nov. 2022.

Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. HiTeA: Hierarchical temporal-aware video-language pre-training. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 15359–15370, Paris, Fr., Oct. 2023. IEEE. doi: 10.1109/iccv51070.2023.01413. URL https://doi.org/10.1109/ICCV51070.2023.01413.

Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *arXiv*, abs/2501.07888, 2025. doi: 10.48550/arXiv.2501.07888. URL https://doi.org/10.48550/arXiv.2501.07888.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. MERLOT: Multimodal neural script knowledge models. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 34, pp. 23634–23651, Virtual conf., Dec. 2021.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-CLIP: Unlocking the long-text capability of CLIP. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 15109 of *Lect. Notes Comput. Sci.*, pp. 310–325. Springer Nat. Switz., 2024. doi: 10.1007/978-3-031-72983-6_18. URL https://doi.org/10.1007/978-3-031-72983-6_18.

Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. DreamLIP: Language-image pre-training with long captions. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 15076 of *Lect. Notes Comput. Sci.*, pp. 73–90. Springer Nat. Switz., Sept. 2024. doi: 10.1007/978-3-031-72649-1_5. URL https://doi.org/10.1007/978-3-031-72649-1_5.

Weijun Zhuang, Yuqing Huang, Weikang Meng, Xin Li, Ming Liu, Xiaopeng Hong, Yaowei Wang, and Wangmeng Zuo. Cluster-wise spatio-temporal masking for efficient video-language pretraining. *arXiv*, abs/2603.22953, 2026. doi: 10.48550/arXiv.2603.22953. URL https://doi.org/10.48550/arXiv.2603.22953.

Table 7: **Video Question-Answering Results.** We evaluate transfer to ActivityNet-QA, MSR-VTT-QA, and MSVD-QA. Our models show particularly strong performance on ActivityNet-QA while remaining competitive across the three benchmarks.

Method	#Samples	ANet-QA	MSR-VTT-QA	MSVD-QA
ClipBERT (Lei et al., 2021)	0.2M	–	37.4	–
ALPRO (Li et al., 2022a)	5M	–	42.1	45.9
JustAsk (Yang et al., 2021)	69M	38.9	41.5	47.5
VideoCLIP (Xu et al., 2021)	136M	–	–	–
All-in-one (Wang et al., 2023)	138M	–	44.3	47.9
MERLOT (Zellers et al., 2021)	180M	41.4	43.1	–
VIOLET (Fu et al., 2021)	138M	–	43.9	47.9
Singularity (Lei et al., 2023)	17M	44.1	43.9	–
OmniVL (Wang et al., 2022a)	17M	–	44.1	51.0
VINDLU (Cheng et al., 2023)	25M	44.7	44.6	–
FrozenBiLM (Yang et al., 2022)	400M	43.2	47.0	54.8
InternVideo (Wang et al., 2022b)	646M	–	47.1	55.5
VideoCoCa (Yan et al., 2022)	4.8B	–	46.0	56.9
UMT-B (Li et al., 2023b)	5M	43.5	44.3	49.1
	17M	44.9	44.9	48.9
UMT-L (Li et al., 2023b)	5M	45.1	45.5	51.3
	17M	47.3	46.4	53.4
MVC-B (Ours)	5M	47.3	45.1	49.5
	10M	48.3	45.5	50.2
MVC-L (Ours)	5M	49.1	46.0	51.5
	10M	50.0	46.3	51.9

A Appendix

A.1 VQA Results

Table 7 reports transfer performance on ActivityNet-QA (ANet-QA) (Caba Heilbron et al., 2015), MSR-VTT-QA (Xu et al., 2016), and MSVD-QA (Chen & Dolan, 2011). Unlike retrieval, video question answering requires interpreting visual content in the context of a natural-language query, providing a complementary evaluation of the learned video-text representations.

For the 5M entries in Table 7, MVC improves over the corresponding UMT models on ActivityNet-QA. MVC-B improves from 43.5 to 47.3 (+3.8), while MVC-L improves from 45.1 to 49.1 (+4.0). The gains on MSR-VTT-QA are more modest, with MVC-L improving from 45.5 to 46.0, while performance on MSVD-QA remains comparable (51.3 vs. 51.5). Increasing the number of video clips from 5M to 10M further improves all three benchmarks, with MVC-L reaching 50.0 on ActivityNet-QA, 46.3 on MSR-VTT-QA, and 51.9 on MSVD-QA. Notably, MVC-L achieves the strongest ActivityNet-QA result in the table despite using only 10M video clips, exceeding UMT-L trained on 17M reported pretraining videos (50.0 vs. 47.3) and models trained on substantially larger corpora. Overall, these results indicate that the benefits of our richer textual supervision extend beyond retrieval to downstream video-text understanding, with particularly strong improvements on ActivityNet-QA.



Figure 2: **Video frames.** Example of sub-sampled frames fed to the MLLMs for caption generation.

A.2 MLLM-BASED MULTI-VIEW SUPERVISION GENERATION

We detail the prompts used to generate the different captions hereafter. For all types of captions, the video frames and the corresponding textual prompt are jointly fed to the MLLM as input. We use separate prompts for summary captions, detailed captions, refined detailed captions, and semantic positive captions, matching the stages described in the Methodology section.

Summary caption prompts. Summary captions are intended to provide a short semantic anchor for each video. We therefore ask each captioning MLLM to describe only the dominant activity in a single concise sentence.

Qwen3-VL-30B-Instruct

Describe the main activity in this video, summarize the result in a single sentence of maximum 16 words.

Tarsier2-Recap-7B

Describe the main activity in this video in short.

Detailed caption prompts. Detailed captions aim to expose the pretraining model to richer visual evidence than summary captions. The prompt therefore encourages the MLLM to describe the visible content in detail while keeping the output bounded.

Qwen3-VL-30B-Instruct

Describe this video in detail, summarize the result in maximum of 128 words.

Tarsier2-Recap-7B

Describe the video in detail.

Refined detailed caption prompts. Refined detailed captions are generated from each detailed caption independently. In each call, Qwen3-VL-8B-Thinking receives the video and one draft detailed caption, either from Qwen3-VL-30B-Instruct or Tarsier2-Recap-7B, and is asked to verify the caption against the visual content. The prompt emphasizes factual correction rather than free-form regeneration.

Qwen3-VL-8B-Thinking

Analyze the visual content of this video focusing on the main action, objects, people, and environment. Be concrete and factual. Ignore text overlays, subtitles, watermarks, and UI elements. Do not infer audio, speech, emotions, or intentions.
Look at the following caption which has been generated by a video captioning model. Identify and correct any inaccuracies or hallucinations. Remove text overlays, watermarks, and UI elements mentioned in the caption. Here is the caption: $\boxed{D_m}$.
Write the new caption, based on the one already provided, but corrected to accurately reflect the visual content of the video only. Remove any foreign characters or symbols unrelated to the core video content.

Semantic positive caption prompts. Semantic positive captions are generated from a refined detailed caption and the video. Unlike generic paraphrases, these captions are constrained to describe specific visible events that remain semantically consistent with the same video. We ask Qwen3-VL-8B-Thinking to return two semantic positive captions by focusing on concrete actors, actions, objects, and outcomes. For each video, we randomly select one refined detailed caption D_m^* with $m \in \{1, 2\}$ and invoke the prompt once; the call returns P_1 and P_2.

Qwen3-VL-8B-Thinking

Based on the visual content of the video and on the following detailed caption $\boxed{D_m^*}$*, provide two positive captions. Each positive caption must describe one specific, meaningful event that is clearly visible in the video, not a generic scene summary.*

Table 8: **Caption statistics.** We report the average and standard deviation of caption length in words for each intermediate and supervision view.

Caption type	Avg. words	Std. words
Original caption	9.4	2.6
Qwen summary caption	13.8	2.6
Tarsier summary caption	16.2	4.5
Qwen detailed caption	96.0	27.1
Tarsier detailed caption	80.1	25.0
Qwen refined detailed caption	60.6	19.5
Tarsier refined detailed caption	60.3	22.5
Semantic positive caption	25.0	7.6

Each positive must explicitly include the actor, the key action, the object or person involved, and the visible outcome or state change. Make the two positives non-identical by focusing on different moments or phases of the event when possible, while remaining fully faithful to what is shown.
Use precise paraphrasing and action-level wording, and avoid vague text such as "someone does something" or purely static descriptions. Do not mention timestamps or frame numbers in the positives.

Caption statistics. Table 8 reports the mean and standard deviation of caption length in words for each intermediate and supervision view. The generated detailed captions are the longest and also show the largest spread, refinement reduces both their average length and variability, and the semantic positive captions remain concise while adding event-specific supervision.

These word-count differences reflect the roles of the caption types: summary captions provide concise anchors, detailed captions provide broader context, and semantic positive captions focus on specific visible events.

Effect of refined detailed captions. To complement the caption-length statistics, we analyze how refinement modifies detailed captions over a subset of 100,000 records. Given a detailed caption D_m and its refined version D_m^*, we compute the Jaccard similarity over their content-word sets:

$$J_{\text{content}}(D_m, D_m^*) = \frac{|C(D_m) \cap C(D_m^*)|}{|C(D_m) \cup C(D_m^*)|},$$

where $C(\cdot)$ denotes the set of content words. Higher values indicate greater lexical overlap. Since this metric ignores word order, frequency, and sentence structure, it measures lexical overlap rather than semantic correctness or claim preservation.

The average caption lengths observed on the audited subset are consistent with the aggregated statistics present in Table 8. Refinement reduces Qwen captions from 96.0 to 60.6 words and Tarsier captions from 80.1 to 60.3 words. No target length is imposed during refinement. Instead, as specified in the refinement prompt above, the model is instructed to remove text overlays, subtitles, watermarks, user-interface elements, unrelated symbols, hallucinated content, and non-visual inferences. The frequent shortening is therefore consistent with the intended removal of irrelevant or unsupported information. Qwen refinement yields a mean content-word Jaccard similarity of 0.405, with 0.47% exact matches, whereas Tarsier refinement is more conservative, yielding a similarity of 0.564 and 2.58% exact matches.

We further use Gemma-3-12B-IT as an independent text-only evaluator to assess the semantic relation between each detailed caption and its refined version. Gemma classifies 98.54% of Qwen and 98.72% of Tarsier transformations as semantically consistent or compatible, including omissions and changes in temporal focus. Specifically, 93.13% of Qwen and 87.99% of Tarsier refinements are classified as compatible with omissions, while only 0.79% and 0.90%, respectively, are classified as contradictory. These results indicate that refinement substantially compresses and rewrites the detailed captions while rarely introducing text-level semantic conflicts. Since Gemma observes only the captions, this analysis measures semantic consistency and transformation behavior rather than visual correctness.

Qualitative example. Figure 2 shows the video used in the example below. The original dataset caption is short and underspecified; each stage adds a complementary supervision signal with a different level of granularity or visual grounding.

Original dataset caption

a person walks on rocks overlooking the ocean

Summary captions

Qwen3-VL-30B-Instruct: A person walks along a rocky shoreline, enjoying the ocean view and natural surroundings.
Tarsier2-Recap-7B: A person walks along a rocky path on a coastline.

Detailed captions

Qwen3-VL-30B-Instruct: A woman wearing a hat and backpack walks along a rocky shoreline. She holds a drone controller in her hand. The ocean is a deep blue and waves crash against the rocks. She walks away from the camera along the shore. She turns and walks back toward the camera. The sun is shining and the sky is clear. She walks toward the camera and then turns and walks away again. She stops and looks out at the ocean. The camera follows her as she walks along the shore.
Tarsier2-Recap-7B: A person wearing a hat, red shirt, and carrying a black backpack stands on a rocky coastline, looking at the ocean. The person begins to walk along the rocky path, moving away from the camera. The person continues walking, occasionally glancing back at the ocean. The person keeps walking along the rocky path, with the ocean and cliffs in the background.

Refined detailed captions

Refined Qwen caption: A person wearing a hat and backpack walks along a rocky shoreline. They hold a stick in their hand. The ocean is a deep blue with waves crashing against the rocks. The person walks away from the camera along the shore, stops, and continues walking away. The sun is shining and the sky is clear.
Refined Tarsier caption: A person wearing a hat, red shirt, and carrying a black backpack stands on a rocky coastline, looking at the ocean. The person begins to walk along the rocky path, moving away from the camera. The person continues walking, occasionally glancing back at the ocean. The person keeps walking along the rocky path, with the ocean and cliffs in the background.

Semantic positive captions

Positive 1: The person shifts weight from one foot to the other while stepping onto a wet rock, maintaining balance on the rocky path.
Positive 2: The person walks away from the camera, continuing along the rocky coastal trail toward the distant grassy cliff.

Refinement effectiveness example. Figure 3 illustrates how refinement corrects object misidentification and removes unsupported actions in detailed captions. The original dataset caption and the Qwen3-VL-30B-Instruct captions incorrectly identify the animal as a dog, whereas the video shows a wet brown cat being dried with a pink towel. The detailed captions further introduce actions that are not supported by the visual content, such as lifting the animal's head or body. Refinement with Qwen3-VL-8B-Thinking corrects the animal category and removes these unsupported details in the detailed captions while preserving the visible drying activity. Incorrect or unsupported content in the input captions is highlighted in red, while corrected or visually grounded content in the refined detailed captions is highlighted in green.

Original dataset caption

a small dog is being toweled in a small bathroom



Figure 3: **Video frames.** Example of sub-sampled frames fed to the MLLMs for caption generation. This example shows the effect of refining the detailed captions to remove incorrect elements.

Summary captions

Qwen3-VL-30B-Instruct: A person dries a wet dog with a pink towel in a tiled bathroom.
Tarsier2-Recap-7B: A person dries a wet brown cat with a pink towel in a tiled bathroom.

Detailed captions

Qwen3-VL-30B-Instruct: A person in a blue and white striped shirt and green shoes is drying a wet brown dog with a pink towel in a bathroom with blue tiles. The person wraps the towel around the dog's body and rubs it gently. The dog walks around while the person continues to dry it. The person then lifts the dog's head and looks into its eyes. The video ends with the person still drying the dog.
Tarsier2-Recap-7B: In a tiled bathroom, a person is seen drying a wet brown cat with a pink towel. The person starts by wiping the cat's back and head, moving the towel across the cat's body. The cat appears to be standing on a tiled floor, and the person continues to dry the cat, moving the towel from the back to the sides and front. The person then wraps the towel around the cat's body, ensuring it is thoroughly dried. Finally, the person lifts the cat slightly, continuing to dry it with the towel.

Refined detailed captions

Refined Qwen caption: A person in a blue and white striped shirt and green shoes is drying a wet brown cat with a pink towel in a bathroom with blue tiles. The person rubs the towel over the cat's body and tail, moving it across the cat's fur. The cat stands on the floor while the person continues to dry it. The video ends with the person still drying the cat.
Refined Tarsier caption: In a tiled room, a person is drying a wet brown cat with a pink towel. The person starts by wiping the cat's back, moving the towel across the cat's body. The person continues to dry the cat's sides, then focuses on the tail area by moving the towel along the cat's back and sides. The person lifts the towel to dry the cat's tail.

Semantic positive captions

Positive 1: A person wearing a blue and white striped shirt rubs a pink towel against the head of a brown tabby cat, with the cat remaining stationary on the blue-tiled floor while the towel makes contact with the cat's fur.
Positive 2: The same person shifts the pink towel to the back of the brown tabby cat, rubbing the towel over the cat's fur as the cat takes a step forward on the blue-tiled floor.

A.3 DATASETS

All datasets used in this work are publicly available and were used in accordance with their original licenses and terms of use. We will release the generated captions for the InternVid-10M-FLT dataset after acceptance.

InternVid. InternVid (Wang et al., 2024b) is a large-scale video-text dataset designed for multi-modal understanding and generation. It contains over 7M web videos spanning approximately 760K hours, resulting in 234M video clips paired with automatically generated textual descriptions. In this work, we use the publicly released InternVid-10M-FLT subset containing 10M filtered video clips with original captions for large-scale video-text pretraining. We refer to these clips as videos in the method and results sections.

MSR-VTT. MSR-VTT (Xu et al., 2016) is a large-scale open-domain benchmark containing 10,000 web videos and approximately 200,000 captions covering diverse content such as sports,

music, cooking, and daily activities. It is widely used for evaluating general video-text alignment. It contains around 7000 training and 1000 test samples.

DiDeMo. DiDeMo (Hendricks et al., 2017) consists of around 10,000 videos paired with multiple descriptions associated with temporally localized events. The benchmark primarily evaluates retrieval under temporal event understanding. It contains around 8496 training and 1034 test samples.

ActivityNet Captions. ActivityNet Captions (Caba Heilbron et al., 2015) contains approximately 20,000 long videos with dense temporal annotations describing activities and events. The longer duration and multiple events per video make retrieval more challenging. It contains around 10K training and 5K test samples.

LSMDC. LSMDC (Rohrbach et al., 2015) contains around 118,000 movie clips paired with text derived from scripts and audio descriptions. The benchmark includes complex scenes and narrative content from movies.

MSVD. MSVD (Chen & Dolan, 2011) contains roughly 2,000 short videos with around 80,000 human-written captions. Despite its smaller size, it remains a commonly used benchmark for video-text evaluation.

CaReBench. CaReBench (Xu et al., 2025) is a fine-grained retrieval benchmark providing separate spatial and temporal annotations, enabling independent evaluation of appearance-focused and temporal retrieval. The benchmark contains 1,000 test samples.

DREAM-1K. DREAM-1K (Wang et al., 2024a) contains 1,000 test videos paired with detailed descriptions of actions and event sequences. We evaluate both detailed-description retrieval and event-based retrieval settings.

Shot2Story20K. Shot2Story20K (Han et al., 2023) focuses on multi-shot video understanding using shot-level captions and video-level summaries. We use the 2,000-sample test split and evaluate both whole-clip and single-shot retrieval settings.

A.4 Additional Ablations

Effect of Masking Ratio. Table 9 studies the effect of the video masking ratio while keeping the caption supervision fixed. Reducing the masking ratio from 50% to 10% improves performance on MSVD (38.8→39.7), ActivityNet (50.1→52.5), DiDeMo (53.0→54.3), and CARE-T (54.1→55.9), while performance on Shot2Story-S remains similar. In contrast, LSMDC favors the higher masking ratio, decreasing from 18.2 to 15.7 at 10%. Overall, a lower masking ratio is beneficial on most benchmarks, suggesting that retaining more visual tokens is advantageous when learning from richer textual supervision.

Table 9: **Effect of Masking Ratio.** We compare different video masking ratios while retaining the same caption supervision ($O + S + D^* + P$). We report text-to-video R@1; bold indicates the best result in each column.

Masking ratio	MSVD	ANet	DiDeMo	LSMDC	CARE-T	Shot2Story-S
50%	38.8	50.1	53.0	**18.2**	54.1	67.9
25%	39.5	51.9	53.0	15.8	53.3	**68.5**
10%	**39.7**	**52.5**	**54.3**	15.7	**55.9**	68.3

Effect of Training Schedule. Table 10 evaluates different training durations. Performance improves through 20 epochs, with smaller gains from 15 to 20 epochs than at earlier intervals. This differs from many previous video-text pretraining settings that commonly adopt shorter schedules, indicating that richer multi-view supervision can continue providing useful learning signals over longer optimization periods.

Table 10: **Effect of training duration.** We evaluate different pretraining durations while keeping the model, supervision, and training configuration fixed. We report text-to-video R@1 across six downstream benchmarks.

Epochs	MSVD	ANet	DiDeMo	LSMDC	CARE-T	Shot2Story-S
5	31.6	40.8	42.6	12.9	42.9	59.7
10	36.2	46.5	48.7	15.8	50.3	64.9
15	37.9	48.9	51.5	17.3	52.6	66.7
20	**38.8**	**50.1**	**53.0**	**18.2**	**54.1**	**67.9**

Table 11: **Fine-Tuning Hyperparameters.** Settings used for retrieval and VQA experiments across datasets.

Task	Dataset	LR	Epochs	DropPath
Retrieval	MSR-VTT	2e-5 (B), 1e-5 (L)	10(B),7(L)	0.2(B),0.3(L)
	DiDeMo	2e-5(B), 2e-5(L)	12(B),5(L)	0.1(B),0.3(L)
	ActivityNet	4e-5(B), 1e-5 (L)	20(B/L)	0.1(B),0.2(L)
	LSMDC	2e-5 (B), 1e-5 (L)	10(B),8(L)	0.1(B),0.2(L)
	MSVD	2e-5 (B), 1e-5 (L)	10(B/L)	0.2(B),0.3(L)
VQA	ActivityNet-QA	4e-5(B),1e-5(L)	12(B),10(L)	0.2(B),0.3(L)
	MSR-VTT-QA	2e-5(B),1e-5(L)	8(B/L)	0.2(B),0.4(L)
	MSVD-QA	2e-5(B),1e-5(L)	15(B),6(L)	0.2(B),0.4(L)
Optimizer		AdamW		
Momentum		$\beta_1, \beta_2 = 0.9, 0.999$		
Weight decay		0.02		
LR schedule		Cosine decay		
Batch size		256		
Warmup epochs		1		
Input frames		12		
Augmentation		MultiScaleCrop [0.5,1], Flip		

A.5 Fine-Tuning Hyperparameters

Table 11 shows the fine-tuning settings for text-to-video retrieval and VQA. All reported results correspond to a single training run following standard evaluation protocols used in prior works.