BEYOND STATIC CUES: MOTION-AWARE VIDEO RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

We identify a fundamental limitation in current video retrieval models: they rely heavily on static appearance cues and inadequately capture the motion dynamics that distinguish visually similar actions. To systematically study this limitation, we introduce MAVR, a motion-centric benchmark and training dataset built largely from newly collected videos for fine-grained text-to-video and video-to-text retrieval. MAVR combines motion-defined actions, scene-matched hard negatives, and a controlled caption scheme that separates action information from scene and viewpoint cues. Despite performing strongly on a broader multimodal embedding benchmark, existing models perform poorly on MAVR, revealing a substantial gap in motion-sensitive retrieval. As a first step toward closing this gap, we introduce MAVE, a motion-aware embedding architecture that augments a multimodal embedding model with temporally informed tokens from a dedicated video pathway. When trained on MAVR, MAVE substantially outperforms state-of-the-art retrieval baselines on actions represented during training, demonstrating the benefit of explicit temporal modeling. However, these gains do not consistently generalize to unseen actions, underscoring that robust motion-aware retrieval remains an open challenge.

1 INTRODUCTION

Video retrieval requires not only representing the content of individual frames, but also capturing the temporal dynamics reflected in the progression of frame content over time. Yet, state-of-the-art video embedding models can rely heavily on frame-level appearance, retrieving videos that are visually similar to a query but do not match its described action or motion. As a result, a retrieval system may return the correct person, object, or scene while missing what actually happens in the video. Figure 1 shows two clips of rolling the hand forward and backward. When sampled sparsely, the clips contain nearly identical poses and scene content, although their motion directions differ. A retriever that mainly represents individual frames may therefore assign similar embeddings to the two clips and fail to match them to the correct action queries.

Two factors encourage this behavior. First, many models encode frames independently. Although the language model can attend across frame tokens, neither the visual encoder nor the language model is explicitly trained to represent how visual content evolves over time. Second, many video-retrieval datasets contain examples that are visually distinguishable from individual frames, allow-



Figure 1: Visually similar frames can represent different actions. Rolling the hand forward and backward produces nearly identical poses and scene content when the videos are sampled sparsely, although the actions differ in their motion direction. Correct retrieval therefore requires modeling how visual content evolves over time rather than relying mainly on frame-level appearance.

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.

ing models to optimize contrastive objectives using objects, scenes, or isolated poses rather than fine-grained motion. Sparse frame sampling further compounds this problem by removing the intermediate evidence needed to distinguish visually similar actions.

To systematically measure this gap, we introduce MAVR (motion-aware video retrieval), a motion-sensitive video-retrieval benchmark and training dataset for text-to-video (T2V) and video-to-text (V2T) retrieval. MAVR is built largely from newly collected videos of motion-defined actions, where relevance depends on how visual content evolves over time. It combines scene-diverse evaluation, scene-matched hard negatives, and controlled captions that separate action information from scene and viewpoint cues. Together, these design choices make static appearance an unreliable shortcut and support controlled evaluation of motion-sensitive retrieval. Across several out-of-the-box models, performance ranges from 11–58% Recall@1 on the video tasks of MMEB-V3 (Huang et al., 2026), a widely used retrieval benchmark, yet no model exceeds 8% on MAVR. This discrepancy shows that strong performance on a broad multimodal embedding benchmark does not necessarily translate to fine-grained motion-sensitive retrieval.

As a first step toward closing the gap exposed by MAVR, we introduce MAVE, an embedding architecture that augments a multimodal embedding model with a dedicated temporal pathway. The pathway produces temporally informed tokens that are prepended to the language model context alongside standard vision encoder-based video tokens, allowing the model to jointly use appearance-level and motion-level representations when producing video embeddings. After MAVR adaptation, MAVE significantly outperforms MAVR-trained baselines on nearly all settings involving actions seen during training across both T2V and V2T retrieval. However, these gains do not consistently extend to unseen actions, highlighting the broader open problem of learning motion-aware video embeddings that generalize beyond the actions represented during training.

Our main contributions are:

- We introduce MAVR, a large-scale benchmark and training dataset for motion-sensitive text-to-video and video-to-text retrieval. MAVR is built largely from newly collected videos and combines motion-defined actions, controlled captions, scene-diverse evaluation, and scene-matched hard negatives.
- We benchmark existing video embedding models on MAVR and show that strong performance on broad multimodal retrieval benchmarks does not translate to fine-grained motion-sensitive retrieval.
- As a first step toward closing this gap, we introduce MAVE, a motion-aware embedding architecture with a dedicated temporal pathway. MAVE improves retrieval for actions represented during training, while its limited gains on unseen actions highlight an important direction for future work.

2 Related Work

Video Retrieval. Early video-retrieval methods adapt CLIP-based image–text encoders (Radford et al., 2021) to videos by independently encoding sampled frames and aggregating their representations (Luo et al., 2022; Ma et al., 2022; Reddy et al., 2025; Shen et al., 2025; Wu et al., 2023; Hur et al., 2025). Ma et al. (2022) extend video–text matching with multi-grained similarities between video, frame, sentence, and word representations. More recent methods use late interaction between text tokens and spatial and temporally contextualized video features (Reddy et al., 2025), query-conditioned frame weighting (Shen et al., 2025), generated-caption enrichment (Wu et al., 2023), and frame-level caption representations with query-irrelevant-information filtering (Hur et al., 2025).

CLIP-based embedding models have also been adapted to composed video retrieval (CoVR), where a reference video and a textual modification are combined to retrieve a target video (Ventura et al., 2024). Prior work improves this composed representation through generated captions of the reference video (Thawakar et al., 2024), decomposition of the modification into retained, injected, and excluded semantics (Wu et al., 2025), and modeling shared and differential frame semantics together with modification-aware object associations (Hu et al., 2026). Most closely related, Gupta et al. (2026) introduce TF-CoVR, a temporally fine-grained CoVR benchmark and an adapted AIM-style (Yang et al., 2023) video encoder for temporally discriminative embeddings.

Table 1: Comparison with representative video-retrieval benchmarks. MAVR combines direct bidirectional text–video retrieval with motion-defined relevance, newly collected videos, and explicit controls against appearance shortcuts.

Benchmark	Retrieval	Size	Temporal	Motion-defined	Source	Scene setup	Appearance control	Train	Unseen
MMEB-V3 (Huang et al., 2026)	Multiple multimodal retrieval tasks	190 tasks	Partial	No	Existing and newly introduced tasks	Task-dependent	No explicit control	No	Not reported
UVRB/UVRD (Guo et al., 2025)	Textual, composed, and visual retrieval	16 evaluation datasets; 1.55M training pairs	Yes	Partial	Public datasets and synthesized pairs	Heterogeneous	No explicit control	Yes	Not reported
CaReBench (Xu et al., 2026)	Text-to-video retrieval and captioning	1,000 videos	Yes	No	Curated videos	Diverse activities and environments	Separated spatial and temporal annotations	No	Not reported
TF-CoVR (Gupta et al., 2026)	Video + modification text → video	180K triplets	Yes	Yes	FineGym and FineDiving	Gymnastics and diving settings	Fine-grained labels; no explicit scene control	Yes	Not reported
MAVR (ours)	**Text ↔ video**	**357,195 clips; 287 actions**	**Yes**	**Yes**	**Largely newly collected crowdsourced videos**	**Action prescribed; scene unconstrained**	**Scene-diverse test set; scene-matched negatives**	**Yes**	**Yes**

"Temporal" denotes an explicit focus on fine-grained temporal information; "motion-defined" means that motion directly determines relevance. "Train" indicates a corresponding training resource, and "unseen" denotes evaluation on action classes excluded from training.

Recent work has repurposed Multimodal Large Language Models (MLLMs) as universal multimodal embedding models that support retrieval across text, images, videos, and visual documents. These methods adapt MLLM backbones through contrastive fine-tuning, staged training, hard-negative mining, distillation, and curricula over increasingly complex retrieval tasks (Meng et al., 2025; Jian et al., 2025; Li et al., 2026; Guo et al., 2025). Other approaches improve retrieval by incorporating retrieval-oriented reasoning (Lan et al., 2026; Jiang et al., 2026), text-based contrastive adaptation (Bagad & Zisserman, 2025), or Mixture-of-Experts LoRA adaptation (Wu et al., 2026). However, these approaches neither introduce a dedicated temporal component nor specifically adapt video embeddings to encode motion dynamics in video embeddings.

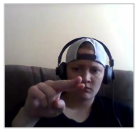
Video Retrieval Benchmarks. Several benchmarks and datasets support training or evaluating video embedding models. MMEB-V3 includes video-retrieval tasks within a broader multimodal embedding benchmark (Huang et al., 2026). UVRB evaluates retrieval across diverse video tasks and query modalities, while UVRD provides training data enriched with spatiotemporal captions (Guo et al., 2025). CaReBench evaluates retrieval using detailed annotations of static content, dynamic actions, and filming-related attributes (Xu et al., 2026). TF-CoVR (Gupta et al., 2026) instead targets temporally fine-grained composed video retrieval using videos from FineGym (Shao et al., 2020) and FineDiving (Xu et al., 2022).

MAVR complements these benchmarks by evaluating direct T2V and V2T retrieval of motion-defined actions, where fine-grained action dynamics themselves determine relevance rather than broad appearance cues, event-level temporal progression, or composed retrieval. It consists of single-person action clips with scene-matched hard negatives that reduce the usefulness of static appearance cues. Its caption levels distinguish action information from scene and viewpoint cues. Together, these design choices support both motion-sensitive training and controlled evaluation of whether video embeddings encode motion dynamics. Table 1 summarizes the differences in retrieval formulation, data source, and appearance control.

3 THE MAVR DATASET

We introduce *Motion-Aware Video Retrieval* (MAVR), a benchmark and training dataset for text-to-video (T2V) and video-to-text (V2T) retrieval of actions distinguishable by fine-grained motion dynamics rather than static appearance. Built largely from newly collected videos, MAVR contains single-person actions, such as gestures, exercise movements, and air-drawn characters, whose identity depends on how the video unfolds over time. MAVR also combines scene-matched hard negatives and a caption scheme to reduce appearance shortcuts and make motion more important for retrieval.

Data Collection. MAVR contains 357,195 short clips spanning 287 action classes, listed in full in Appendix A. Paid crowd-workers were shown an action label and example videos, then recorded themselves performing the action using their own devices. Although the action was prescribed, the room, lighting, viewpoint, clothing, and recording device were unconstrained, producing substantial scene variation within each class. Participants provided consent for research release; collection and consent details appear in Appendix B.



Action ($\mathcal{A}$)	Drawing the letter B in the air
Action + Context ($\mathcal{A} + \mathcal{C}$)	a person drawing the letter B in the air, indoors near a couch, under bright lighting with a static camera, wearing a black top, shown in third-person view
Context ($\mathcal{C}$)	indoors near a couch, under bright lighting with a static camera, wearing a black top, shown in third-person view

Figure 2: The MAVR caption scheme. The action-only caption ($\mathcal{A}$) names the action, the action-and-context caption ($\mathcal{A} + \mathcal{C}$) adds scene and viewpoint details, and the context-only caption ($\mathcal{C}$) describes the same video without naming the action.

Data Preprocessing. Before release, we removed exact duplicates and undecodable clips, stripped container metadata, and deleted audio. To retain only single-person clips, we applied the InsightFace face-detection pipeline (Deng et al., 2020) to sampled frames and discarded clips in which more than one face was detected. Further filtering and privacy details are are provided in Appendices B and 6.

Scene Attributes. We extract technical metadata with `ffprobe` (FFmpeg Developers, 2026) and automatically generate scene descriptors using Qwen3-VL-8B-Instruct (Bai et al., 2025). Using eight uniformly sampled frames, the model predicts the setting, background objects, lighting, camera motion, viewpoint, and coarse clothing and hair colors when visible. Unavailable attributes are omitted, and demographic attributes are excluded from the published annotations.

Caption Scheme. Each clip receives three deterministic captions: an *action-only caption* ($\mathcal{A}$), an *action-and-context* caption ($\mathcal{A} + \mathcal{C}$), and a *context-only* caption ($\mathcal{C}$). Throughout, *context* refers to the video's scene and viewpoint information. Figure 2 illustrates the three caption variants. The $\mathcal{A}$ and $\mathcal{A} + \mathcal{C}$ captions are used in both dataset construction and retrieval experiments, whereas $\mathcal{C}$ is used only for dataset construction. Additional caption details are provided in Appendix C.

Class Splits. We partition the 287 action classes into 274 *seen* classes used in both training and evaluation, 11 *unseen* classes held out entirely for zero-shot evaluation, and 2 *train-only* classes that are too small for reliable evaluation. The unseen set spans multiple action families, allowing us to test whether models generalize to actions not represented during training.

Scene-Diverse Test Selection. To prevent frequent classes from dominating evaluation, the test gallery is class-balanced: each of the 285 evaluated classes contributes 10 clips, yielding 2,850 test clips (2,740 from seen actions and 110 from unseen actions). Within each class, we embed the context-only captions using `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019), ℓ_2-normalize the embeddings, and cluster them with k-means using $k = 10$. We retain the clip nearest to each centroid, to cover a range of scenes and viewpoints without favoring outliers. The resulting test clips share an action while spanning different backgrounds, lighting conditions, viewpoints, and appearance, as illustrated in Figure 3. The remaining clips from seen actions form the training set. Further details appear in Appendix D.

Hard-Negative Mining. For each training clip $\mathcal{V}$, we mine ten hard negatives from clips with a different action but similar scene context, providing challenging negatives for contrastive train-



Figure 3: Scene-diverse test clips for the *waving* class. The six examples show the same action across different rooms, backgrounds, lighting conditions, and viewpoints. Clips are selected near cluster centroids in context-only caption space so that the test set covers common scene variations rather than repeatedly showing the same setting.

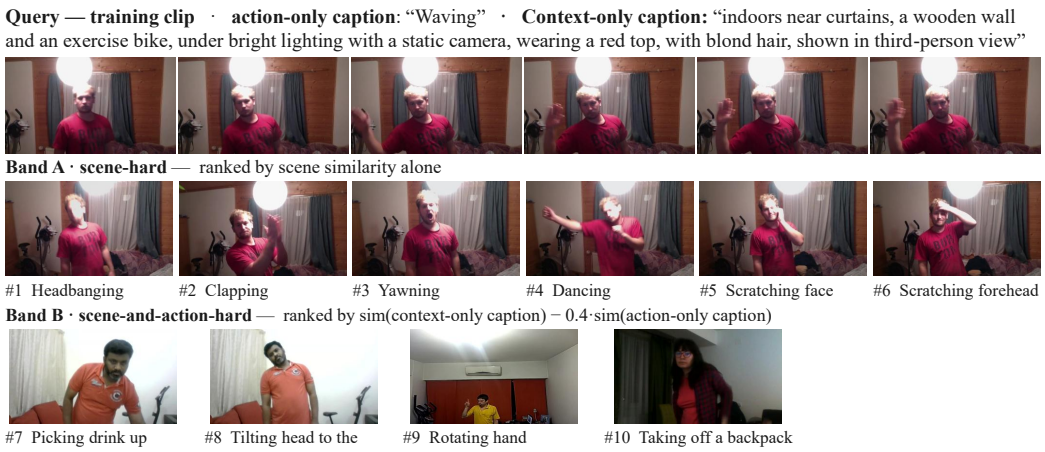


Figure 4: Hard negatives for a *waving* training clip. Band A contains clips of different actions recorded in the most similar scenes. Band B also favors actions that differ from the query. The selected clips preserve many of the same room and appearance cues while changing the action, making appearance alone unreliable during training.

ing (Oord et al., 2018). Mining scores each candidate relative to $\mathcal{V}$ using ℓ_2-normalized Sentence-BERT embeddings of two caption variants: the context-only caption measures scene similarity, $\mathrm{sim}_{\mathcal{C}}$, whereas the action-only caption measures action similarity, $\mathrm{sim}_{\mathcal{A}}$. The negative pool is restricted to training clips from a different action class. The ten negatives are drawn from two bands. Band A selects six *scene-hard* negatives with the highest $\mathrm{sim}_{\mathcal{C}}$, prioritizing clips with closely matched scene context and, in many cases, the same participant. Band B selects four additional candidates using Equation 1, favoring similar scenes with different actions. Figure 4 shows a query and its selected negatives. Because these clips look similar but show different actions, they make it harder for the model to rely on appearance alone. Near-duplicate filtering, per-class diversity guards, ordering, and coverage statistics are detailed in Appendix D.

$$sim(candidate, \mathcal{V}) = \mathrm{sim}_{\mathcal{C}}(candidate, \mathcal{V}) - 0.4\ \mathrm{sim}_{\mathcal{A}}(candidate, \mathcal{V}) \tag{1}$$

Dataset Summary. MAVR contains $354{,}345$ training clips and a class-balanced test set of $2{,}850$ clips. The videos are short, with a median duration of 3.0 seconds, and the number of clips per action varies considerably across classes. Most videos were recorded indoors from a third-person viewpoint using a static camera. Full split, video, and scene statistics are provided in Appendix E.

4 MAVE: Motion-Aware Video Embedding

Standard MLLM-based video embedding models face two limitations in capturing fine-grained motion. First, they rely on image-oriented ViTs that encode sampled frames independently. Although the LLM backbone attends across frame tokens and uses temporal positional encodings, neither component is explicitly trained to model how visual content evolves over time. Also, these models typically use low video sampling rates (1–2 FPS), which can discard intermediate cues needed to distinguish similar actions. Naively increasing the sampling rate can recover some of these cues, but it substantially increases the number of visual tokens and the associated computational cost. We propose MAVE as a first step toward addressing both limitations. MAVE augments a pretrained MLLM-based video embedding model with a dedicated *temporal pathway* that runs in parallel with its existing *visual pathway*, as illustrated in Figure 5. The visual pathway processes sparsely sampled frames, while the temporal pathway processes a higher-rate video stream and compresses its output into a fixed number of temporal tokens, which are inserted before the visual tokens in the LLM context. The LLM can then jointly attend to temporal and visual information when generating the video embedding.

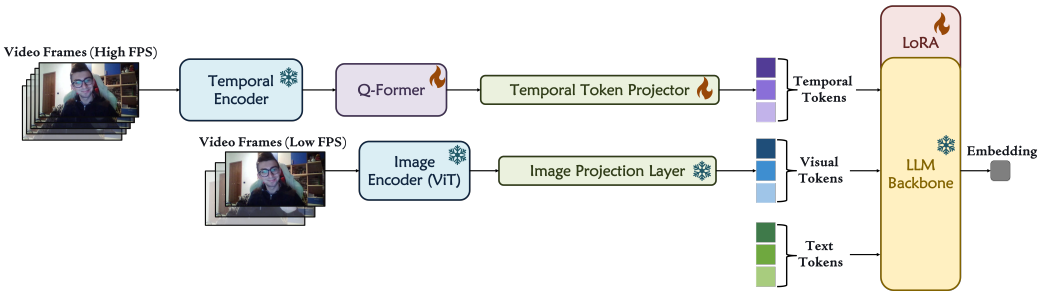


Figure 5: Overview of MAVE. The visual pathway (middle) processes sparsely sampled frames to produce visual tokens, while the temporal pathway (top) processes a higher-rate video stream and compresses its features into a fixed number of temporal tokens.

4.1 Visual Pathway

The visual pathway follows the base MLLM embedding model and remains frozen. The input video is sampled at a low rate (1 FPS) and encoded by the ViT into patch-level features. The visual projection module maps these features into the LLM embedding space, producing the visual tokens X_{vis} in Equation 2, where $\text{Frames}_{\text{low}}$ denotes the low-rate frames, L_{vis} is the resulting number of visual tokens, and d is the LLM embedding dimension.

$$X_{\text{vis}} = \text{Proj}_{\text{vis}}\big(\text{ViT}(\text{Frames}_{\text{low}})\big) \in \mathbb{R}^{L_{\text{vis}} \times d} \tag{2}$$

4.2 Temporal Pathway

The temporal pathway samples the input video at a higher rate ($\sim$ 5 FPS), producing $\text{Frames}_{\text{high}}$. A frozen temporal encoder maps these frames to temporal features, which are then compressed by a trainable Q-Former (Li et al., 2023) and mapped to the LLM embedding space by a trainable projection module. The temporal and visual pathways therefore use separate frame streams and encoder weights.

Frozen temporal encoder. A frozen pretrained temporal encoder maps the densely sampled frames to a sequence of temporal features, as shown in Equation 3, where L_{te} is the number of temporal encoder-output tokens and d_{te} is their feature dimension. MAVE can use any pretrained video encoder whose objective encourages temporal modeling. In our implementation, $\text{TempEnc}(\cdot)$ is a frozen V-JEPA2 model (Assran et al., 2025), whose training objective encourages the model to predict masked spatiotemporal representations in latent space. The encoder remains frozen throughout training to preserve its pretrained temporal representations.

$$Z_{\text{te}} = \text{TempEnc}(\text{Frames}_{\text{high}}) \in \mathbb{R}^{L_{\text{te}} \times d_{\text{te}}} \tag{3}$$

Q-Former compression. Because L_{te} can be large and varies across videos, a Q-Former compresses Z_{te} into a fixed set of K temporal embeddings. It uses K learnable queries $Q = [q_1, \ldots, q_K] \in \mathbb{R}^{K \times d_{\text{te}}}$ that cross-attend to Z_{te} to produce the temporal embeddings defined in Equation 4. Because K is independent of video length and sampling rate, the temporal pathway contributes a bounded number of tokens to the LLM context.

$$Z_{\text{q}} = \text{QFormer}(Q, Z_{\text{te}}) \in \mathbb{R}^{K \times d_{\text{te}}} \tag{4}$$

Temporal token projection. A learned linear projection maps the Q-Former outputs into the LLM embedding space, as illustrated in Equation 5. The Q-Former and temporal token projector are the only trainable components of the temporal pathway.

$$X_{\text{tmp}} = Z_{\text{q}} W_{\text{tmp}} + b_{\text{tmp}} \in \mathbb{R}^{K \times d} \tag{5}$$

4.3 Fusion of Temporal and Visual Tokens in the LLM

The projected temporal tokens X_{tmp} are enclosed by learned temporal tag embeddings, e_{start} and e_{end}, which mark their boundaries within the LLM context, analogous to existing visual tag tokens.

The temporal tokens precede the visual tokens X_{vis} and text tokens X_{text}, corresponding to a query or prompt. This gives the LLM input context in Equation 6, where $[\,\cdot\,;\,\cdot\,]$ denotes concatenation along the token dimension. Under causal attention, this ordering allows every subsequent visual and text token to attend to the temporal tokens.

$$C = \left[\, e_{\text{start}};\, X_{\text{tmp}};\, e_{\text{end}};\, X_{\text{vis}};\, X_{\text{text}} \,\right] \tag{6}$$

We adapt the LLM attention layers with Low-Rank Adapters (LoRA) (Hu et al., 2021) so that the model can learn to integrate temporal and visual information. The ViT, visual projection module, pretrained LLM weights, and temporal encoder remain frozen. The trainable components are the LoRA parameters, Q-Former, temporal token projector, and temporal tag embeddings. Text queries are encoded using the same adapted LLM backbone, and video–text relevance is computed as the cosine similarity between the resulting video and text embeddings.

5 Experiments

We first evaluate several embedding models on MAVR, showing that they are limited in their ability to encode motion dynamics, and that MAVR captures a dimension of video retrieval that is not adequately reflected in MMEB-V3 (Huang et al., 2026), a widely used embedding benchmark. We compare these results with performance on MMEB-V3 to determine whether success on a broad multimodal embedding benchmark carries over to motion-sensitive retrieval. We then evaluate whether MAVE improves motion-sensitive retrieval relative to visual-only embedding baselines trained on MAVR. This comparison assesses the contribution of MAVE's temporal pathway.

5.1 Diagnosing Motion Dynamics in Video Embeddings

Experimental Setup. We evaluate zero-shot multimodal embedding models on the video tasks of MMEB-V3 and on MAVR's test gallery for text-to-video retrieval. For MMEB-V3, we report mean Recall@1 across its video tasks. For MAVR, we report *target-sampled* Recall@1 for action-only ($\mathcal{A}$) queries and action-and-context ($\mathcal{A}+\mathcal{C}$) queries. Standard Recall@1 for an $\mathcal{A}$ query rewards the retrieval of any video from the queried action class, even if the remaining relevant videos rank poorly. For example, with ten positive videos, a model that ranks only one positive video first, with the rest given a low rank, receives the same Recall@1 score as a model that ranks all ten well. Target-sampled recall instead runs one retrieval trial per positive video. In each trial, a single positive video is used as the target, with the other positive videos being masked from the gallery, and recall is computed for that target. An average recall value is then obtained for these trials. Averaging these trials rewards models that retrieve every relevant video rather than only a favorable subset. This procedure is a no-op for $\mathcal{A}$+$\mathcal{C}$ queries, which have one positive video per query.

All models are evaluated without fine-tuning, with video inputs sampled at 1 FPS. We consider two groups of models. The first consists of dedicated multimodal embedding models: Qwen3-VL-Embedding (2B and 8B) (Li et al., 2026), GVE (3B and 7B) (Guo et al., 2025), RzenEmbed (7B) (Jian et al., 2025), and UME-R1 (2B) (Lan et al., 2026). The second consists of VLMs repurposed for retrieval by using the final-layer representation of the last input token as the text or video embedding. This group includes Qwen3-VL-Instruct (2B, 8B, and 32B) (Bai et al., 2025), Qwen3.8-27B (Qwen Team, 2026), and InternVL3-38B (Zhu et al., 2025). The 2B and 8B Qwen3-VL-Instruct models share the underlying model family with Qwen3-VL-Embedding, allowing us to examine the importance of retrieval-specific training. The 32B variant tests whether scaling the same VLM family improves retrieval, while Qwen3.8-27B represents a newer Qwen generation and InternVL3-38B provides a comparison outside the Qwen family.

Results. Table 2 reports results on both benchmarks. Models achieve a wide range of Recall@1 scores on MMEB-V3 (11–58%). In contrast, all models obtain low Recall@1 on MAVR: 0–3% for action-only queries and 0–8% for action-and-context queries. Action-and-context queries consistently produce slightly higher scores compared to action-only queries, because they provide scene and viewpoint details in addition to the action. This improvement is itself informative: when contextual cues are available, models can use them to narrow the search even though they still struggle to identify the motion. Overall, no model that performs well on MMEB-V3 achieves strong performance on MAVR. For example, the strongest models on MMEB-V3 reach 55–58% Recall@1 but

Table 2: Recall@1 (%) on the MMEB-V3 video tasks and MAVR text-to-video retrieval for action-only ($\mathcal{A}$) and action-and-context ($\mathcal{A} + \mathcal{C}$) queries, evaluated zero-shot. General-purpose VLMs are repurposed as embedding models using the final-layer hidden state of the final token.

Category	Model	Size	MMEB-V3	MAVR	
				$\mathcal{A}$	$\mathcal{A} + \mathcal{C}$
Embedding models	Qwen3-VL-Embedding	2B	55	1	5
	UME-R1	2B	42	1	2
	GVE	3B	31	0	1
	GVE	7B	36	0	1
	RzenEmbed	7B	56	3	8
	Qwen3-VL-Embedding	8B	58	2	7
Repurposed VLMs	Qwen3-VL-Instruct	2B	11	0	0
	Qwen3-VL-Instruct	8B	11	0	0
	Qwen3.8	27B	14	0	0
	Qwen3-VL-Instruct	32B	11	0	0
	InternVL3	38B	18	0	0

remain below 8% on MAVR. Increasing model size also does not resolve the gap, and the repurposed VLMs obtain 0% on both MAVR query types. These results suggest that current embedding models do not reliably encode motion dynamics for video retrieval, and that MAVR evaluates a complementary capability–fine-grained motion-sensitive retrieval–that is not adequately covered by MMEB-V3. MAVR exposes this gap by holding the action constant as the retrieval target while reducing the usefulness of objects, scenes, and isolated poses.

5.2 Evaluating MAVE

Experimental Setup. We compare MAVE, built on Qwen3-VL-Embedding-2B, with LoRA-adapted Qwen3-VL-Embedding baselines of 2B and 8B parameters. The baselines use frozen visual pathways with videos sampled at 1 FPS, and adapt the backbone attention layers using LoRA (rank 16, $\alpha = 32$). MAVE augments the 2B backbone with a temporal pathway comprising a frozen V-JEPA2-300M encoder, which processes videos at 5 FPS, followed by a 1-layer, 8-head Q-Former that compresses its output into $K = 32$ temporal tokens. The Q-Former, temporal projector, temporal tag embeddings, and backbone LoRA adapters are trainable.

All models are trained on MAVR's training split and evaluated on its test split using target-sampled variants of Recall@1 and nDCG@10, as described in Section 5.1. They optimize a symmetric multi-positive InfoNCE objective (Oord et al., 2018) with a fixed temperature of 0.03. Each training batch contains 768 videos and 32 text queries. In addition to in-batch negatives, we include the scene-matched hard negatives from Section 3 and dynamically mined non-positive videos with high similarity to each query. Training was executed on 8$\times$ H100 GPUs.

MAVE is trained in two stages: during the *warm-start* stage, we train only the Q-Former, temporal token projector, and temporal tag embeddings while keeping the LoRA adapters disabled. This stage allows the temporal pathway to learn a useful compression of the temporal encoder features before adapting the LLM to integrate them. During the *joint* stage, we activate the LoRA adapters and train them together with the temporal pathway, enabling the LLM to jointly attend to temporal and visual tokens when producing video embeddings. Each stage runs for up to 10,000 steps with early stopping. The baselines use a single LoRA-adaptation stage with the same training schedule.

Results. Table 3 reports results for MAVE and the LoRA-adapted baselines.

On seen actions, MAVE significantly outperforms both baselines in every retrieval setting except the V2T retrieval using $\mathcal{A} + \mathcal{C}$ captions against the 8B baseline, where its numerical advantage is not statistically significant. The largest gains occur for action-only queries, where scene and viewpoint cues are unavailable. MAVE improves Recall@1 over the matched 2B baseline from 19% to 45% for T2V and from 45% to 67% for V2T retrieval. These results show that MAVR provides a useful training signal for learning motion-sensitive embeddings and that MAVE's temporal pathway is particularly effective for actions represented during training.

On unseen actions, MAVE does not significantly outperform either baseline in any setting. For V2T retrieval with $\mathcal{A}$+$\mathcal{C}$ captions, both baselines significantly outperform MAVE in nDCG@10, although

Table 3: MAVR-training results. The numerical values represent target-sampled Recall@1 and nDCG@10 (%) for action-only ($\mathcal{A}$) and action-and-context ($\mathcal{A}$+$\mathcal{C}$) queries, evaluated on actions *seen* during training and actions held out from training (*unseen*).

	Action-only ($\mathcal{A}$)				Action-and-context ($\mathcal{A}+\mathcal{C}$)			
	Seen		Unseen		Seen		Unseen	
Model	R@1	nDCG@10	R@1	nDCG@10	R@1	nDCG@10	R@1	nDCG@10
Text-to-Video Retrieval								
Qwen3-VL-embedding-2B + LoRA	19	34	**6**	**16**	68	82	50	69
Qwen3-VL-embedding-8B + LoRA	24	38	5	**16**	74	86	**58**	**76**
MAVE-2B (ours)	**45**†§	**60**†§	4	14	**79**†§	**89**†§	55	73
Video-to-Text Retrieval								
Qwen3-VL-embedding-2B + LoRA	45	46	14	14	68	82	58	74*
Qwen3-VL-embedding-8B + LoRA	50	51	**19**	**20**	76	87	**60**	**77***
MAVE-2B (ours)	**67**†§	**68**†§	11	11	**78**†	**88**†	43	63

†: MAVE significantly outperforms Qwen3-VL-embedding-2B; §: MAVE significantly outperforms Qwen3-VL-embedding-8B; $*$: baseline significantly outperforms MAVE. Statistical significance is computed using a paired permutation test with Holm–Bonferroni correction across all 32 comparisons with MAVE ($\alpha = 0.05$).

their Recall@1 advantages are not statistically significant. This might be due to MAVE's trainable temporal modules overspecializing to the seen action distribution; temporal features remain useful for seen motions but may not align reliably with captions for novel motions. This effect may be amplified in V2T retrieval, where each video is contrasted with fewer text queries during training (32 queries per video, compared with 768 videos per query in T2V) and no explicitly mined hard queries are used for each video. Although the baselines share MAVE's training objective, they may retain more of the pretrained appearance-based video–text alignment that benefits retrieval of unseen $\mathcal{A} + \mathcal{C}$ captions. These results suggest that integrating the temporal pathway with the embedding model may benefit from pretraining on a broader video–text corpus before MAVR fine-tuning, rather than learning this alignment solely from MAVR's seen action distribution.

Additional analyses are provided in the appendix: an ablation study of MAVE's temporal-pathway design choices (Appendix H); two video-sampling-rate controls showing that the temporal pathway remains beneficial at low sampling rates and that denser visual sampling improves the 2B baseline while remaining below MAVE (Appendix I); and an analysis of cross-benchmark performance degradation after MAVR adaptation and mitigation through embedding anchoring (Appendix J). Appendix K discusses limitations of MAVR's construction, MAVE's design, and the experimental design.

6 Conclusion

This work introduces MAVR, a motion-sensitive video retrieval benchmark and training dataset built largely from newly collected videos of motion-defined actions. Experiments show that current embedding models perform poorly on MAVR despite strong performance on MMEB-V3, revealing that broad multimodal retrieval performance does not necessarily reflect the ability to distinguish actions based on their motion dynamics. By combining motion-defined relevance with scene-diverse evaluation, controlled captions, and scene-matched hard negatives, MAVR provides a focused setting for measuring this gap. We also propose MAVE as a first step toward motion-aware video embeddings. On actions seen during MAVR training, MAVE significantly outperforms baselines in nearly all retrieval settings, showing that an explicit temporal pathway can improve motion-sensitive retrieval. Its weaker performance on unseen actions, however, shows that these improvements do not yet generalize reliably beyond the actions represented during training. Developing motion-sensitive embeddings that generalize to new actions remains an open challenge. MAVR provides a benchmark and training resource for measuring progress on this problem.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, code development, and figure and table formatting. All experimental decisions, analyses, claims, citations, and final manuscript content were performed and verified by the authors.

ETHICS STATEMENT

We designed the collection, processing, and release of MAVR with participant privacy, informed consent, and responsible use as primary concerns. Because the benchmark's premise is that motion rather than appearance should drive retrieval, the same principle shapes what we release: we publish what is needed to study temporal understanding and withhold sensitive or identity-revealing signals.

Human Subjects, Consent, and Compensation All clips were created by paid participants recruited through crowdsourcing. Before recording, each participant was shown the target action as a label plus example videos, was informed of the purpose of the collection, and signed a consent form explicitly authorizing the release of their recordings for research. Participation was voluntary. The recording task was benign by construction: participants performed one specified everyday action (e.g. a gesture or an exercise movement) using their own device camera. No sensitive activities were requested, and no personally identifying information such as names or contact details was collected as part of the release.

Privacy Protection Because the recordings depict identifiable individuals, we apply several protections before release. Container-level metadata is stripped and audio is removed without re-encoding, eliminating device, timestamp, location, and voice information. To restrict the release to single-subject clips, we retain a clip only if the face-detection filter described in Appendix B detects at most one face in every sampled frame.

We do not blur or mask faces because many MAVR classes depend on motion of the head, face, or upper body (e.g. *shaking head*, *yawning*, *scratching forehead*, and *headbanging*). Masking would remove the signal these classes are intended to measure. Individuals therefore remain visually identifiable in released clips. Privacy rests on informed consent, the exclusion filter above, and the withholding of identity-linked annotations rather than visual de-identification. Appendix K discusses the remaining risk.

Sensitive and Demographic Attributes MAVR's captions are built from action labels and coarse scene attributes, and we deliberately exclude demographic and sensitive attributes from every released file. The only appearance cues retained, such as clothing and hair color, describe visual context rather than identity. This keeps protected attributes out of the retrieval task and prevents the benchmark from rewarding retrieval based on sensitive characteristics.

Automatically Generated Annotations Action-class labels are human-reviewed, but scene attributes and the resulting captions are entirely machine-generated and are not human-verified. They may contain errors, such as a mislabeled background object or a wrong clothing color, or reflect biases in the underlying model. To limit the impact of such errors, captions are assembled deterministically from validated attribute fields, uncertain fields are omitted rather than guessed, and grammatical post-processing is guarded so it cannot introduce or remove attribute content. Captions should nonetheless be treated as noisy annotations rather than ground-truth descriptions of individuals or environments.

Intended Use and Potential Risks MAVR is intended for research on motion-aware multimodal retrieval and related video understanding. The clips depict people performing benign actions and are anonymized at the metadata and filename level, but faces remain visible, so re-identification risk is reduced rather than eliminated. We discourage any attempt to re-identify participants, to infer sensitive attributes, or to use the data for surveillance or biometric applications. The benchmark is not appropriate for tasks requiring demographic labels and should not be used to train systems whose purpose is to identify or profile individuals. Users should respect the accompanying license and the intent under which participants provided consent.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. *URL https://arxiv.org/abs/2506.09985*, 2025.

Piyush Bagad and Andrew Zisserman. Adapting mllms for nuanced video retrieval. *arXiv preprint arXiv:2512.13511*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-stage dense face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5203–5212, 2020.

FFmpeg Developers. ffprobe documentation. `https://ffmpeg.org/ffprobe.html`, 2026. Accessed: 2026-08-15.

Zhuoning Guo, Mingxin Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Xiaowen Chu. Towards universal video retrieval: Generalizing video embedding via synthesized multimodal pyramid curriculum. *arXiv preprint arXiv:2510.27571*, 2025.

Animesh Gupta, Jay Parmar, Ishan Rajendrakumar Dave, and Mubarak Shah. From play to replay: Composed video retrieval for temporally fine-grained videos. *Advances in Neural Information Processing Systems*, 38, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Yupeng Hu, Zixu Li, Zhiwei Chen, Qinlei Huang, Zhiheng Fu, Mingzhu Xu, and Liqiang Nie. Refine: Composed video retrieval via shared and differential semantics enhancement. *ACM Transactions on Multimedia Computing, Communications and Applications*, 22(7):1–24, 2026.

Haohang Huang, Xuan Lu, Mingyi Su, Xuan Zhang, Ziyan Jiang, Ping Nie, Kai Zou, Tomas Pfister, Wenhu Chen, Wei Zhang, et al. Mmeb-v3: Measuring the performance gaps of omni-modality embedding models. *arXiv preprint arXiv:2604.23321*, 2026.

Chan Hur, Jeong-hun Hong, Dong-hun Lee, Dabin Kang, Semin Myeong, Sang-hyo Park, and Hyeyoung Park. Narrating the video: Boosting text-video retrieval via comprehensive utilization of frame-level captions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24077–24086. IEEE, 2025.

Weijian Jian, Yajun Zhang, Dawei Liang, Chunyu Xie, Yixiao He, Dawei Leng, and Yuhui Yin. Rzenembed: Towards comprehensive multimodal retrieval. *arXiv preprint arXiv:2510.27350*, 2025.

Haonan Jiang, Yuji Wang, Yongjie Zhu, Xin Lu, Wenyu Qin, Meng Wang, Pengfei Wan, and Yansong Tang. Embed-rl: Reinforcement learning for reasoning-driven multimodal embeddings. *arXiv preprint arXiv:2602.13823*, 2026.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. Ume-r1: Exploring reasoning-driven generative multimodal embeddings. In *International Conference on Learning Representations*, volume 2026, pp. 115779–115806, 2026.

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, 2023.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

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 and captioning. *Neurocomputing*, 508: 293–304, 2022.

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.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, et al. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

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.

Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, Kenton Murray, Reno Kriz, Celso M De Melo, Benjamin Van Durme, and Rama Chellappa. Video-colbert: Contextualized late interaction for text-to-video retrieval. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19691–19701. IEEE, 2025.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2613–2622. IEEE, 2020.

Leqi Shen, Sicheng Zhao, Yifeng Zhang, Pengzhang Liu, Yongjun Bao, and Guiguang Ding. Spatio-temporal attention for text-video retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(11):1–20, 2025.

Omkar Thawakar, Muzammal Naseer, Rao Muhammad Anwer, Salman Khan, Michael Felsberg, Mubarak Shah, and Fahad Shahbaz Khan. Composed video retrieval via enriched context and discriminative embeddings. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26886–26896. IEEE, 2024.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5270–5279, 2024.

Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4video: What can auxiliary captions do for text-video retrieval? In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10704–10713. IEEE, 2023.

Yebo Wu, Feng Liu, Ziwei Xie, Zhiyuan Liu, Changwang Zhang, Jun Wang, and Li Li. Tsembed: Unlocking task scaling in universal multimodal embeddings. *arXiv preprint arXiv:2603.04772*, 2026.

Yue Wu, Zhaobo Qi, Yiling Wu, Junshu Sun, Yaowei Wang, and Shuhui Wang. Learning fine-grained representations through textual token disentanglement in composed video retrieval. In *International Conference on Learning Representations*, volume 2025, pp. 18006–18028, 2025.

Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2939–2948. IEEE, 2022.

Yifan Xu, Xinhao Li, Yichun Yang, Desen Meng, Rui Huang, and Limin Wang. Carebench: A fine-grained benchmark for video captioning and retrieval. In *International Conference on Learning Representations*, volume 2026, pp. 51124–51150, 2026.

Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. *arXiv preprint arXiv:2302.03024*, 2023.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX

A COMPLETE LIST OF ACTION CLASSES

Below are all 287 action strings in the release. The nine groups below are editorial, introduced only to make the list readable. Within a group, actions are ordered alphabetically and separated by semicolons, and the group size is given in parentheses. We use the † marker to represent *unseen* actions that are held out of training (11 actions). We also use the ‡ marker to denote actions that are included only in the training set (2 actions). The remaining *seen* actions (274 actions) are used for both training and evaluation.

Air-drawn characters (35) *Drawing the letter* ⟨X⟩ *in the air* for each of the 26 letters X = A . . . Z (the letter W is the held-out unseen class†); *Drawing the digit* ⟨N⟩ *in the air* for each of the 9 digits N = 1 . . . 9.

Hand and arm gestures (44) *((*still*)/(*moving*) marks the still-/moving-background variant)* Backhand swiping multiple times with right hand; Calling someone closer with finger; Calling someone closer with hand; Clapping; Closing fist (moving); Closing fist (still); Cracking knuckles; Fist entering frame (moving); Fist entering frame (still); Fist leaving frame (moving); Fist leaving frame (still); Forehand swiping multiple times with right hand; Grabbing and tossing; Making a come here gesture; Making a go away gesture; Making an arc motion to the left; Making an arc motion to the right; Moving arm up with palm down; Moving fist around (moving); Moving fist around (still); Moving hand around (moving); Moving hand around (still); Moving left arm up with palm down; Moving right arm up with palm down; Opening fist (moving); Opening fist (still); Rolling hand backward; Rolling hand forward; Rotating hand clockwise; Rotating hand counterclockwise; Snapping fingers; Spinning hand clockwise; Spinning hand counterclockwise; Spreading hand (moving); Spreading hand (still); Swiping left; Swiping left (moving); Swiping left (still); Swiping right; Swiping right (moving); Swiping right (still); Waving; Waving and leaving; Waving with both hands.

Head and face actions (14) Bowing; Covering eyes; Covering eyes with hands then uncovering them; Headbanging; Nodding; Scratching face; Scratching forehead; Shaking head; Tilting head to the left; Tilting head to the right; Turning head to look left; Turning head to look right; Uncovering eyes; Yawning.

Locomotion and posture (37) Bending forward and putting hands onto knees; Coming closer; Coming closer and passing by at the left hand side; Coming closer and passing by at the right hand side; Coming closer then standing; Coming closer while jumping; Entering from the left then standing still; Entering from the right then standing still; Entering through a door; Getting into lying on back position; Getting into pushup position; Getting into sitting position; Going away; Going away while jumping; Going away while running; Going away while walking; Going away while walking and turning left; Going away while walking and turning right; Jumping forward; Jumping to the left; Jumping to the right; Jumping up in the air; Leaving through a door; Running forward; Running from left to right; Running from right to left; Standing up from crouching position; Taking one step back; Taking one step forward; Taking one step to the left; Taking one step to the right†; Turning around to look what is behind; Turning in circles; Walking by from left to right; Walking by from right to left; Walking forward; Walking in circles.

Object and clothing interactions (20) Cheering with drink; Jumping off [something]; Jumping onto [something]; Jumping over [something]; Kicking a ball; Picking [something] up; Picking drink up; Putting [something] down; Putting dishes in dishwasher; Putting drink down†; Putting on a backpack; Putting on a jacket or shirt; Putting on glasses; Putting on glasses upside down‡; Taking off a backpack; Taking off a jacket or shirt; Taking off glasses; Taking off glasses then putting them back on; Taking off upside down worn glasses; Unloading the dishwasher†.

Camera motions (3) Blocking the camera then releasing it; Tilting the camera downward; Tilting the camera upward†.

Vehicle motions (6) Using a [vehicle] and going away; Using a [vehicle] and moving forward; Using a [vehicle] passing by from left to right; Using a [vehicle] passing by from right to left; Using a [vehicle], going away and turning left; Using a [vehicle], going away and turning right†.

Dance (11) Dancing; Dancing - electricity through both arms; Dancing - grease finger point; Dancing - pulp fiction finger wag; Dancing - running man; Moving hips; Playing Air Guitar; Using left arm to tap left shoulder, then right shoulder, then up; Using left arm to tap left shoulder, then right shoulder, then up, then move hip; Using right arm to tap left shoulder, then right shoulder, then up; Using right arm to tap left shoulder, then right shoulder, then up, then move hip.

Fitness and exercise (117) *(prefix* Doing *omitted throughout;* (L)*/*(R) *marks the left-/right-limb variant,* (L bk)*/*(R bk) *the leg-in-back variant, and* $A + B$ *the combination* A with B*)* Air Jump Rope†; Alternating Lunges; Alternating Single Leg Glutes Bridge; Alternating V ups; Arm circles; Arm crosses; Backwards Windmills; Bending Windmill Stretch; Boxing bounce-steps; Bunny Hop; Burpee with pushup; Burpee without pushup; Butt kickers; Criss-Cross; Cross + Hook left; Cross + Hook right; Cross + Jab (L); Cross + Jab (R); Cross + Uppercut left (L); Cross + Uppercut right (R); Cross (L); Cross (R); Cross-legged hamstring stretch; Crouching; Curtsy Lunges (L bk); Curtsy Lunges (R bk)†; Dead Bugs; Diamond Pushups; Downward Dog; Fire Hydrant (L); Fire Hydrant (R); Floor Touches; Forward Stance Forward Bend; Forward Windmills; Front Lunge Kick (L); Front Lunge Kick (R); Glute Hamstring Walkout; Glutes Bridge; Good Morning; Heel Lift (L); Heel Lift (R); High Kicks; High Knees March; High Knees with Punches; High knees; Hip Abductions (L); Hip Abductions (R); Hip Circles; Hook Left (L); Hook Right (L); Hook left + Hook right (R); Hook left + Jab (R); Hook left + Uppercut right (R); Hook right + Cross (R); Hook right + Hook left (L); Hook right + Jab (L); Hook right + Uppercut left (L); Inchworm; Jab + Cross (L); Jab + Hook left (R); Jab + Hook right (L); Jab + Uppercut left (R); Jab + Uppercut right (L); Jabs; Jumping Lunges; Jumping jacks; Kickback (L); Kickback (R); Knee Circles; Knee-Up Rotations; Lunge Jumps; Lunges (L bk); Lunges (R bk); Mountain climbers; Moving plank; Neck Rolls; Oblique Twists; Plank; Plank Taps; Plie Squat; Puddle Jump; Pushups; Quick feet; Roll Down; Shoulder Gators; Shoulder Swipe; Side Plank; Skaters; Snowboarder; Spider Man; Spider Man Pushup; Squat Jabs; Squat Jacks; Squat Jump; Squat Kicks; Squat Punch; Squats; Standing Fire Hydrant (L); Standing Fire Hydrant (R)†; Standing Kicks; Standing Knee-to-Elbows†; Standing Oblique Crunches; Touching Toes; Tuck Jump; Uppercut Left‡; Uppercut Right; Uppercut left + Cross (L); Uppercut left + Hook right (L); Uppercut left + Jab (R); Uppercut left + Uppercut right (L)†; Uppercut right + Cross (R); Uppercut right + Hook left (L); Uppercut right + Jab (L); Uppercut right + Uppercut left (L); Walking Lunges; Wrist twists; Yoga Pushup.

B Data Collection and Annotation Details

This appendix records the implementation details of the pipeline summarized in Section 3 , in the order a clip passes through it, so that the release is reproducible and its automated stages are auditable.

Recording protocol. Clips were gathered through a crowdsourcing platform. Each task presented a participant with one target action, given as a short text label together with example videos demonstrating it, and asked them to perform that action once while recording themselves using their own device cameras. Participants were paid, and signed a consent form authorizing research-use of the recording before recording began, as discussed in the ethics statement. Nothing beyond the action was specified; not the room, framing, distance, lighting, clothing, or recording device. This is what produced the wide scene variation within an action-class that is exploited in the creation of the test split in Section 3. Collection was organized around action families (everyday actions, hand gestures, fitness movements, air-drawn characters and digits, and dance), each contributed by many participants; 7,599 distinct participants are identifiable in the collection logs, and a further 8,298 clips, which are predominantly egocentric (no face is visible), carry no recoverable participant identifier.

Deduplication and metadata stripping. Exact-duplicate source files were removed globally before any processing. Container-level metadata was then stripped with `ffmpeg -map_metadata -1 -c copy`, which discards device, creation-time, and location tags without re-encoding the video stream. At release, the audio stream was dropped entirely, again by stream copy, so no recompression artifacts are introduced and no incidental speech survives.

Face detection. Safety filtering used the InsightFace `buffalo_l` suite (Deng et al., 2020), specifically the `det_10g.onnx` detector. For each clip, we sampled up to ten evenly spaced frames and ran face detection on every sampled frame. A clip was retained only if no sampled frame contained more than one detected face. A separate decodability check removed clips that failed to decode.

Scene attribute extraction. Technical metadata (width, height, frame rate, duration, frame count) was probed per file with `ffprobe`. Semantic scene and actor attributes were extracted with Qwen3-VL-8B-Instruct (Bai et al., 2025), run on local GPUs with 8 sampled frames per clip, each frame capped at 360×360 pixels, and generation limited to 256 new tokens. The model was prompted to return a single JSON object with a fixed key set: `setting` $\in$ {indoor, outdoor, unknown}, `background_objects` (a short list), `lighting` $\in$ {bright, dim, unknown}, `camera_motion` $\in$ {static, handheld, unknown}, and the main actor's `top_color`, `bottom_color`, and `hair_color`. Values the model could not determine were returned as unknown or not-applicable and are omitted from captions rather than verbalized. Egocentric clips receive no actor-appearance attributes. The prompt also requested a perceived-gender field; this is retained internally for auditing and appears in no released file, as do the face counts from the filtering stage.

C Caption Construction

Captions are assembled using a deterministic template from the curated action label plus the extracted attributes; no generative model writes caption text, and no human edits it. Clause order is fixed across the corpus: action, setting, background objects, lighting, camera motion, actor top colour, actor bottom colour, hair colour, and finally viewpoint. A single validity rule governs every slot: an attribute is verbalized only if present and determinate, and is otherwise omitted along with its connective, so captions shorten gracefully rather than carrying placeholders. Egocentric clips therefore yield shorter captions than third-person ones without requiring any special-case logic, since their actor-appearance slots are simply absent.

Three caption variants are then derived from the templates: *action-and-context* ($\mathcal{A}+\mathcal{C}$) retains the full template, *action-only* ($\mathcal{A}$) retains only the action, and *context-only* ($\mathcal{C}$) retains scene and viewpoint attributes while omitting the action.

Templated assembly leaves grammatical artifacts, including article agreement, list conjunctions, and spacing. These are smoothed by an offline rule-based pass. That pass is guarded by a content-

preservation check that rejects any edit which would add or remove an attribute token, so post-processing can alter function words but never caption content. The pipeline is deterministic end to end: identical inputs yield byte-identical captions.

Object placeholders. Eleven action templates describe an interaction with an unspecified object and retain a bracketed placeholder in their caption text, such as *"picking [something] up"* and *"using a [vehicle] and going away"*. Rather than substituting into the caption, we keep the placeholder intact and expose the detected object separately in a `placeholder_object` column, grammatically normalized for substitution (e.g. *"a book"*, *"an id card"*, *"keys"*). This affects $5{,}544$ clips.

D Splits, Test Selection, and Hard-Negative Mining

The main paper describes the two selection objectives and shows representative outputs (Figures 3 and 4). Here we provide the embedding spaces, deterministic rules, guards, ordering, and class-selection details needed to reproduce them.

Scene and action embedding spaces. Both test selection and hard-negative mining rely on sentence embeddings from `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019), ℓ_2-normalized so that cosine similarity is an inner product. Two spaces are used. Embedding the context-only caption yields a *context-only* vector, and the cosine similarity between two clips' context-only vectors is $\mathrm{sim}_{\mathcal{C}}$. Embedding the action-only caption yields an *action-only* vector, giving $\mathrm{sim}_{\mathcal{A}}$. Mining in caption space rather than pixel space is a deliberate choice: it factorizes similarity into a context-only component and an action component that can be scored independently, which is what makes "same scene, different action" expressible as an explicit objective. A joint visual embedding entangles the two and cannot express it.

Test selection. Within each evaluated class, the captions for candidate clips are embedded in the context embedding space, clustered by k-means with $K{=}10$, and the candidate nearest to its cluster centroid is retained. Using each centroid's nearest candidate avoids promoting outliers that happen to sit alone in the embedding space. Classes with exactly K candidates fall back to singleton clusters, and classes that have less than K candidates are excluded from evaluation entirely and assigned train-only status. Selection is deterministic under a fixed random seed.

Hard-negative mining. For each training clip $\mathcal{V}$, the candidate pool contains training clips whose action label differs from $\mathcal{V}$'s. Band A selects six candidates with the highest $\mathrm{sim}_{\mathcal{C}}$. Band B selects four additional candidates from the remainder by maximizing $\mathrm{sim}_{\mathcal{C}} - 0.4\,\mathrm{sim}_{\mathcal{A}}$ (Eq. 1). This combination preserves fine-grained action siblings among the scene-hard negatives while also including scene-matched clips with semantically distant actions.

Guards, diversity, and ordering. Before ranking, candidates with $\mathrm{sim}_{\mathcal{C}}(\mathcal{V}, \cdot) > 0.98$ are removed because such extreme agreement is more consistent with a near-duplicate or mislabeled positive than a useful negative. A per-class cap then admits at most three negatives from any action class. Within each band, candidates are stored from hardest to easiest according to $\mathrm{sim}_{\mathcal{C}}$; training can therefore use any prefix without re-mining.

Unseen class selection. The eleven zero-shot classes were chosen to span semantic families and the full frequency range, rather than being sampled at random or taken from the tail, so that zero-shot difficulty is not confounded with class rarity. They comprise one air-drawn character (*drawing the letter W in the air*), one locomotion class (*taking one step to the right*), one camera motion (*tilting the camera upward*), two object interactions (*putting drink down*, *unloading the dishwasher*), one vehicle motion (*using a [vehicle], going away and turning right*), and five fitness movements (*air jump rope*, *curtsy lunges with right leg back*, *standing fire hydrant with right leg moving*, *standing knee-to-elbows*, and *uppercut left with uppercut right with left leg in front*). Holding out a single air-drawn character while retaining its sibling letter and digit classes in training is deliberate: it tests within-family generalization to an unseen trajectory.

Table 4: MAVR dataset statistics. Caption variants include *action-only* ($\mathcal{A}$), *action-and-context* ($\mathcal{A}+\mathcal{C}$), and *context-only* ($\mathcal{C}$)

	Train	Test	Total
Clips	354,345	2,850	357,195
– seen-class	–	2,740	–
– unseen-class	–	110	–
Action classes	276	285	287
– seen	–	–	274
– unseen	–	–	11
– train-only	–	–	2
Clips/class (median)		576	
Test clips/class	–	10	–
Caption variants/clip		3 ($\mathcal{A}$, $\mathcal{A}+\mathcal{C}$, $\mathcal{C}$)	
Hard negatives/training clip	10	–	–
Duration (s; median)		3.0	
Frames (median)		89	
Frame rate (fps; median)		30	

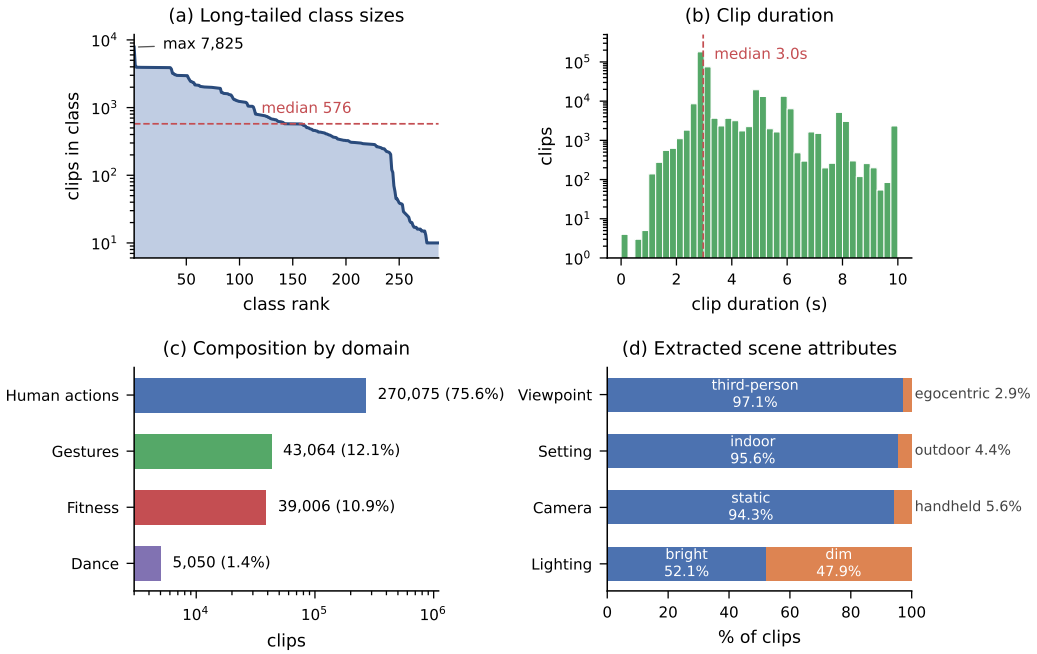


Figure 6: MAVR statistics: (a) long-tailed class sizes, (b) clip durations, (c) action-domain composition, and (d) learned scene-attribute distributions.

E Detailed Dataset Statistics

Table 4 summarizes the released splits and video characteristics, while Figure 6 shows the principal distributions. MAVR comprises approximately 104 GB of video and 360 MB of annotations. Clips are 640 pixels wide with heights of 344–640 pixels (median 360).

The corpus is strongly long-tailed with a median of 576 videos per class, and air-drawn character classes account for 38.5% of the clips. Scene attributes are also concentrated: 97.1% of clips are third-person, 95.6% are indoor, 94.3% use a static camera, and lighting is 52.1% bright versus 47.9% dim. These statistics motivate the balanced test set construction in Appendix D and the limitations discussed in Appendix K.

F RELEASE FORMAT AND RECOMMENDED USAGE

The detailed construction above is reflected in a compact public release. MAVR ships five artifacts joined by the clip identifier `video_id`, the only identifier present in any public file. Table 5 gives the column-level schema.

Table 5: Released files and their schemas. `video_id` joins every file.

File	Rows	Contents
`videos/`	357,195	One `<video_id>.mp4` per clip, video stream only (audio removed by stream copy).
`train.csv`	354,345	`video_id`, `category`, `action-only`, `action-and-context`, `context-only`, `hardneg_1` … `hardneg_10`, `placeholder_object`.
`test.csv`	2,850	`video_id`, `category`, `action-only`, `action-and-context`, `context-only`, `placeholder_object`, `split` ∈ {`seen`, `unseen`}.
`qrels.json`	—	{"action-and-context": {qid: [vid]}, "action-only": {qid: [vid, ...]}}, keyed by query `video_id`: instance-level for the clip itself, class-level for all ten test clips of the query's class.
`video_metadata.csv`	357,195	`video_id`, `width`, `height`, `fps`, `duration`, `num_frames`, `view`, `setting`, `lighting`, `camera_motion`, `background_objects`, `actor_top_color`, `actor_bottom_color`, `actor_hair_color`.

Recommended usage and protocol. Train on `train.csv` with either the action-only or the action-and-context caption as the query and the ranked hard negatives as the negative pool, taking a prefix of each band if fewer than ten negatives are wanted. Evaluate on `test.csv` as both gallery and query source, scoring against `qrels.json` at both granularities, and report seen and unseen partitions separately with macro-averaging over classes.

G ADDITIONAL QUALITATIVE EXAMPLES

Figure 7 shows all eleven zero-shot classes, and Figure 8 shows a broader sample of classes across the four action domains. Two properties are visible throughout. Within a class, scene, participant, framing, and lighting vary widely, because the collection prompt constrained only the action. Across classes, scenes are drawn from a narrow distribution of indoor consumer-device recordings, so appearance rarely separates one class from another. The class is legible in the frame sequence and generally not in any single frame.

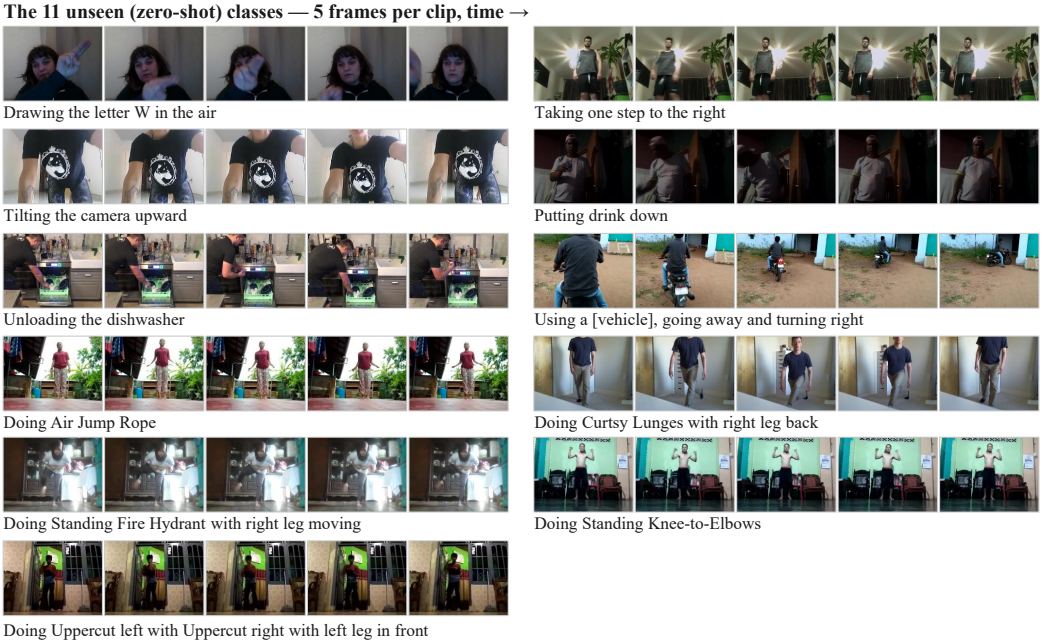


Figure 7: The eleven unseen (zero-shot) classes of the test split , five frames per clip. The set deliberately spans an air-drawn character, locomotion, camera motion, object interactions, a vehicle motion, and five fitness movements, across the full class-frequency range, so that zero-shot difficulty is not confounded with class rarity.

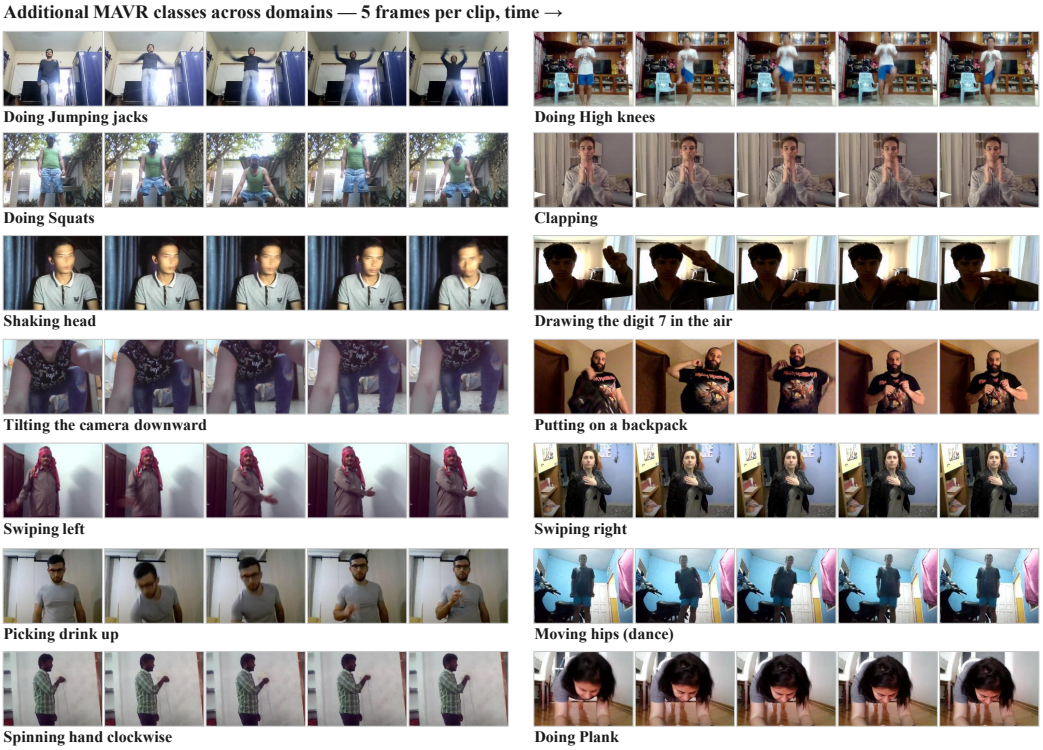


Figure 8: Additional MAVR classes spanning everyday actions, gestures, fitness movements, air-drawn characters, camera motions, and dance, five frames per clip. Note *swiping left* and *swiping right*: mirror-image temporal dynamics that no frame-level appearance feature can separate.

H Ablation Study for MAVE

We evaluate four MAVE variants to isolate the contributions of temporal sampling rate, learnable temporal-feature compression, temporal-token budget, and the visual pathway. Relative to the full model, the variants (i) reduce temporal sampling from 5 FPS to 1 FPS, (ii) replace the Q-Former with mean pooling, (iii) increase the number of Q-Former query tokens from 32 to 64, and (iv) remove the visual pathway. All variants otherwise use the full MAVE training configuration. Table 6 reports the results.

Removing the visual pathway reveals a trade-off between the two caption levels. For seen action-only retrieval, the no-visual-pathway variant improves performance in both directions, indicating that visual tokens can interfere with retrieval when using action-only queries. In contrast, removing the visual pathway significantly degrades all seen $\mathcal{A} + \mathcal{C}$ metrics compared to the main variant, showing that visual tokens are necessary to exploit the scene information in these more specific queries, even though they can compete with motion-focused representations for action-only retrieval.

The full model significantly outperforms the 1-FPS variant on every seen-action metric, showing that higher-rate temporal sampling improves in-distribution retrieval. Replacing the Q-Former with mean pooling similarly degrades every seen-action metric, supporting learned compression over uniform aggregation of temporal-encoder features.

Finally, increasing the Q-Former token budget from 32 to 64 yields no statistically significant improvement on any metric. We therefore use 32 temporal tokens as the more efficient configuration, with no detected loss in retrieval performance.

Table 6: Ablation results for MAVR-training. The numerical values represent target-sampled Recall@1 and nDCG@10 (%) for action-only ($\mathcal{A}$) and action-and-context ($\mathcal{A}$+$\mathcal{C}$) queries, evaluated on actions *seen* during training and actions held out from training (*unseen*).

	Action-only ($\mathcal{A}$)				Action-and-context ($\mathcal{A}+\mathcal{C}$)			
	Seen		Unseen		Seen		Unseen	
Model	R@1	nDCG@10	R@1	nDCG@10	R@1	nDCG@10	R@1	nDCG@10
Text-to-Video								
MAVE-2B (ours, full model)	$45^{\S\dagger}$	$60^{\S\dagger}$	4	14	$\mathbf{79}^{\S\dagger\psi}$	$\mathbf{89}^{\S\dagger\psi}$	55	73^{ψ}
– 1 FPS sampling§	29	46	**5**	**16**	74	86	54	71
– mean-pool compression†	34	51	3	**16**	72	85	49	66
– 64 query tokens‡	44	60	1	**16**	**79**	**89**	**58**	**76**
– no ViT features$^{\psi}$	**50**	$\mathbf{63}^{*}$	**5**	13	63	79	39	59
Video-to-Text								
MAVE-2B (ours, full model)	$67^{\S\dagger}$	$68^{\S\dagger}$	11	11	$78^{\S\dagger\psi}$	$88^{\S\dagger\psi}$	43	63^{ψ}
– 1 FPS sampling§	57	57	8	8	74	86	**51**	**69**
– mean-pool compression†	62	63	10	10	73	85	37	57
– 64 query tokens‡	69	69	15	15	**80**	**89**	48	68
– no ViT features$^{\psi}$	$\mathbf{71}^{*}$	$\mathbf{71}^{*}$	**19**	**20**	62	79	28	49

Markers denote significant differences versus the full model: § (1 FPS sampling); ‡ (64 query tokens); † (mean-pool compression); ψ (no ViT features). $*$ denotes an ablation significantly outperforming the full model. Statistical significance is computed using a paired permutation test with Holm–Bonferroni correction across all 64 comparisons with the full model ($\alpha = 0.05$).

I Effect of Video Sampling Rate

MAVE differs from the Qwen3-VL-Embedding baselines in two respects: (1) it adds a temporal pathway, and (2) feeds that pathway more densely sampled video frames. To isolate the contribution of the temporal pathway independently of sampling rate, we compare the 1-FPS 2B baseline with the 1-FPS MAVE variant used in the ablation study (Appendix H). For this variant, both the visual and temporal streams receive videos sampled at 1 FPS. As shown in Table 7, this variant significantly outperforms the 2B baseline on every seen-action metric in both T2V and V2T, mirroring the full

model's advantage over the matched 2B baseline in Table 3. Thus, MAVE's gains are not solely explained by denser temporal sampling; its temporal pathway itself provides a substantial architectural benefit for motion-sensitive retrieval over the typical MLLM-based model architecture.

Notably, the baseline does not significantly outperform the 1-FPS MAVE variant on unseen L_2 V2T retrieval, unlike its advantage over the full MAVE model in Table 3. This difference suggests that reducing the temporal sampling rate makes the temporal tokens less disruptive to pretrained visual–text matching for this out-of-distribution setting.

Table 7: 1-FPS rate-matched control on MAVR. The MAVE variant samples both its temporal and visual streams at 1 FPS, matching the 1-FPS visual-only 2B baseline and isolating the effect of the temporal pathway from denser temporal sampling. The numerical values represent target-sampled Recall@1 and nDCG@10 (%) for action-only ($\mathcal{A}$) and action-and-context ($\mathcal{A}$+$\mathcal{C}$) queries, evaluated on actions *seen* during training and actions held out from training (*unseen*).

	Action-only ($\mathcal{A}$)				Action-and-context ($\mathcal{A}+\mathcal{C}$)			
	Seen		Unseen		Seen		Unseen	
Model	R@1	nDCG@10	R@1	nDCG@10	R@1	nDCG@10	R@1	nDCG@10
Text-to-Video								
Qwen3-VL-Embedding-2B + LoRA	19	34	**6**	**16**	68	82	50	69
MAVE-2B-1FPS (ours)	**29**†	**46**†	5	**16**	**74**†	**86**†	**54**	**71**
Video-to-Text								
Qwen3-VL-Embedding-2B + LoRA	45	46	**14**	**14**	68	82	**58**	**74**
MAVE-2B-1FPS (ours)	**57**†	**57**†	8	8	**74**†	**86**†	51	69

†: difference between MAVE-2B-1FPS and Qwen3-VL-Embedding-2B is statistically significant. Statistical significance is computed using a paired permutation test with Holm–Bonferroni correction across all 16 comparisons with the MAVE ($\alpha = 0.05$).

Table 8: 3-FPS baseline-sampling control on MAVR. MAVE samples its temporal stream at 3 FPS and its visual stream at 1 FPS, whereas the visual-only 2B baseline samples video at 3 FPS. This tests whether denser visual sampling for the baseline can account for MAVE's gains. The numerical values represent target-sampled Recall@1 and nDCG@10 (%) for action-only ($\mathcal{A}$) and action-and-context ($\mathcal{A}$+$\mathcal{C}$) queries, evaluated on actions *seen* during training and actions held out from training (*unseen*).

	Action-only ($\mathcal{A}$)				Action-and-context ($\mathcal{A}+\mathcal{C}$)			
	Seen		Unseen		Seen		Unseen	
Model	R@1	nDCG@10	R@1	nDCG@10	R@1	nDCG@10	R@1	nDCG@10
Text-to-Video								
Qwen3-VL-embedding-2B-3FPS + LoRA	27	45	**4**	13	76	87	54	72
MAVE-2B-3FPS (ours)	**42**†	**58**†	**4**	**17**	**78**	**88**	**58**	**74**
Video-to-Text								
Qwen3-VL-embedding-2B-3FPS + LoRA	56	57	**16**	**17**	77	**88**	**61**†	**76**†
MAVE-2B-3FPS (ours)	**67**†	**67**†	**16**	16	**78**	**88**	43	64

†: difference between MAVE-2B-3FPS and Qwen3-VL-Embedding-2B-3FPS is statistically significant. Statistical significance is computed using a paired permutation test with Holm–Bonferroni correction across all 16 comparisons with the MAVE ($\alpha = 0.05$).

We also investigated whether increasing the frame rate of videos passed to the baseline would close its gap with MAVE. To this end, we trained a MAVE variant whose temporal and visual streams receive videos at 3 FPS and 1 FPS, respectively, and a Qwen3-VL-Embedding-2B baseline that receives videos at 3 FPS. As shown in Table 8, MAVE remains significantly better on seen $\mathcal{A}$ retrieval in both directions, although the numerical gap is smaller than in Table 3. This suggests that denser visual sampling helps the baseline either by enabling limited use of temporal relationships across frames or by exposing discriminative instantaneous motion states, such as subtle poses or object configurations, that distinguish an action from competing actions. However, the resulting improvement still falls short of the benefit provided by MAVE's temporal pathway and its compressed temporal

tokens. For seen $\mathcal{A} + \mathcal{C}$, MAVE's numerical advantage is no longer statistically significant. This suggests that the additional video evidence available to the baseline may be sufficient, when combined with the scene and viewpoint information in $\mathcal{A}+\mathcal{C}$ queries, to remove a statistically detectable difference.

J MAVR Overspecialization and Catastrophic Forgetting

The limited generalization of MAVE to unseen MAVR actions in Section 5.2 raises a related concern: adapting a pretrained embedding model to MAVR may also degrade its retrieval capability outside the benchmark. We assess this cross-benchmark transfer degradation on MMEB-V3. Indeed, we found that the MAVR-adapted 2B Qwen3-VL-Embedding baseline significantly underperforms its frozen, out-of-the-box counterpart on MMEB-V3, suggesting that MAVR adaptation has led to catastrophic forgetting.

To characterize the trade-off between MAVR adaptation and retention of pretrained retrieval ability, we sweep the LoRA scaling factor α. Figure 9 shows the effect of increasing α from 0, corresponding to no LoRA adaptation, to 32, which is the value used in our main experiments. As α increases, MAVR T2V performance improves substantially, particularly for $\mathcal{A} + \mathcal{C}$-seen queries, while MMEB-V3 Recall@1 decreases by approximately 20 percentage points.

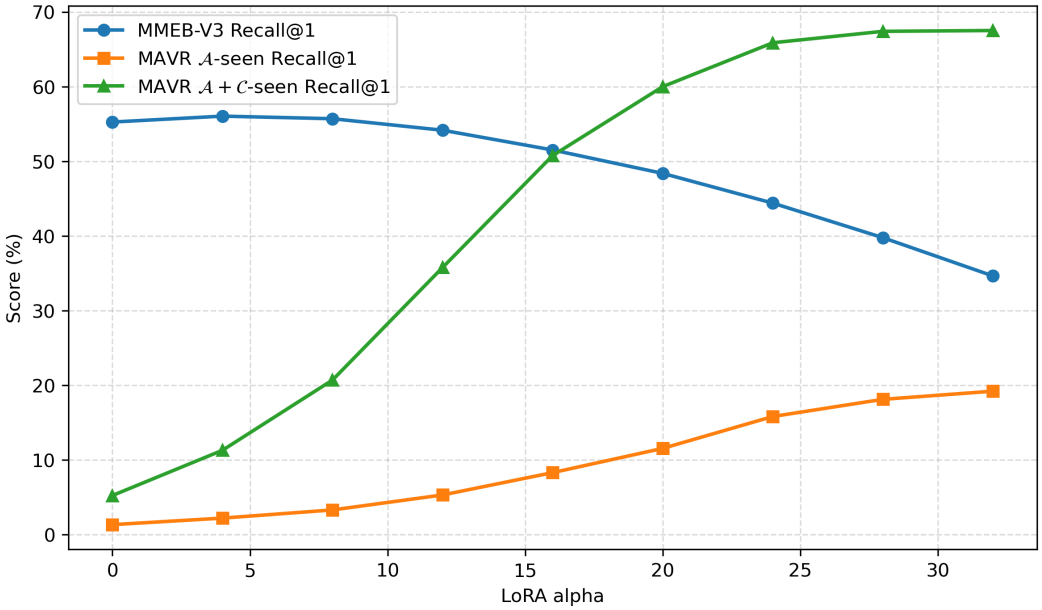


Figure 9: Effect of sweeping the LoRA scaling factor α on the MAVR-adapted 2B Qwen3-VL-Embedding model performance across both MMEB-V3 (video tasks) and MAVR's seen action-only ($\mathcal{A}$) and action-and-context ($\mathcal{A} + \mathcal{C}$) queries.

A related pattern appears for MAVE in Figure 10. Both MAVE curves use the same model trained with a 5-FPS temporal stream, but evaluate MMEB-V3 using temporal streams sampled at either 1 or 5 FPS. Although MMEB-V3 performance declines more gradually with increasing α for MAVE than for the visual-only baseline, both MAVE variants already underperform the frozen baseline at $\alpha = 0$. This initial gap indicates that introducing temporal tokens can itself disrupt the pretrained retrieval representation when the backbone is not adapted to attend to them. Moreover, evaluating MAVE with a 1-FPS temporal stream yields substantially higher MMEB-V3 performance than using a 5-FPS temporal stream across the sweep. This is consistent with the finding in Appendix I: lower-rate temporal tokens may be less disruptive to pretrained visual–text matching in some out-of-distribution scenarios.

We next compare the 5-FPS MAVE variant with the MAVR-adapted 2B baseline on MAVR seen actions, using $\mathcal{A}$ and $\mathcal{A}+\mathcal{C}$ queries, across the LoRA sweep (Figures 11 and 12). MAVE substantially

outperforms the baseline at $\alpha = 0$ on both metrics, showing that its temporal pathway provides useful MAVR retrieval signal even when LoRA adaptation of the backbone is disabled. Both models improve as α increases, while MAVE retains a large gap throughout the action-only sweep. For action-and-context queries, MAVE's performance begins to plateau beyond $\alpha = 16$, whereas the baseline continues improving until approximately $\alpha = 28$. This pattern suggests that MAVE obtains a larger share of its in-distribution $\mathcal{A}+\mathcal{C}$ performance from the temporal pathway at lower backbone adaptation strengths, while the visual-only baseline depends more strongly on backbone adaptation to achieve comparable retrieval performance.

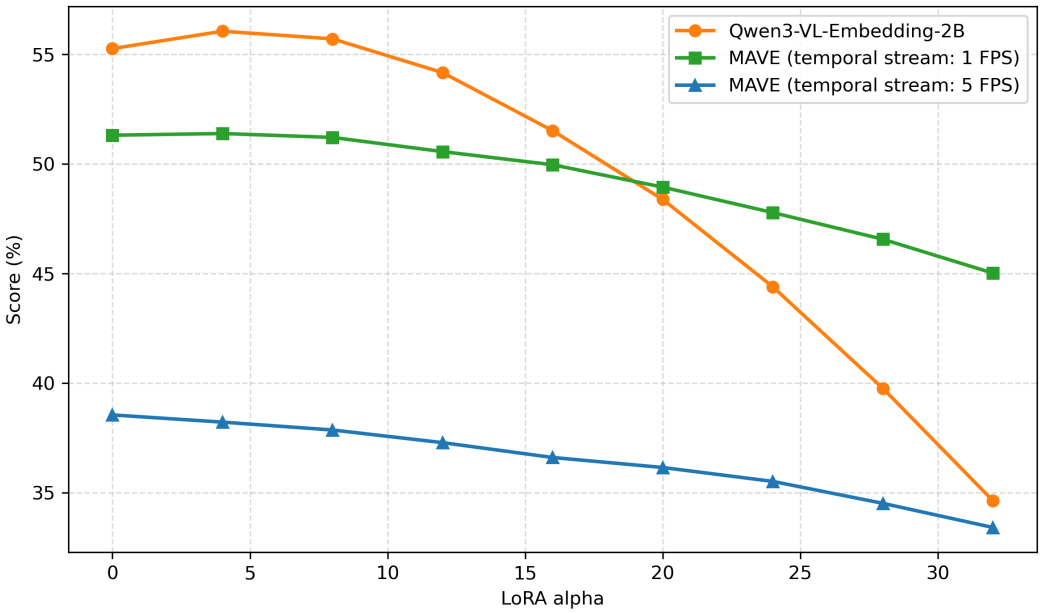


Figure 10: Effect of sweeping the LoRA scaling factor α on the performance of MAVE (temporal stream: 1 & 5 FPS) and the MAVR-adapted 2B Qwen3-VL-Embedding baseline on MMEB-V3's (video tasks) Recall@1.

To mitigate this cross-benchmark transfer degradation, we investigate anchoring the embeddings produced by the adapted baseline to those of its frozen pretrained counterpart during MAVR adaptation. For an adapted embedding $\hat{e}$ and the corresponding frozen embedding e, we augment the training objective with a cosine similarity-based regularization as illustrated in Equation 7, where $\lambda_{\text{anchor}} = 0.5$. This term penalizes deviations from the pretrained representation.

$$\mathcal{L} = \mathcal{L}_{\text{InfoNCE}} + \lambda_{\text{anchor}}\big(1 - \cos(\hat{e}, e)\big) \tag{7}$$

Table 9 compares the frozen model, the MAVR-adapted model without anchoring, and the MAVR-adapted model with anchoring. Relative to the frozen model, unanchored adaptation significantly improves seen-action MAVR retrieval (for both action-only and action-and-context queries) but reduces MMEB-V3 performance. Anchoring recovers much of this lost MMEB-V3 performance while retaining comparable MAVR performance; all numerical differences between the unanchored and anchored variants are statistically insignificant for MAVR's metrics. These results suggest that using anchoring regularization during MAVR-adaptation can partially mitigate catastrophic forgetting induced by MAVR adaptation.

K Limitations

K.1 MAVR Biases and Limitations

We document known limitations for MAVR so that its results are interpreted appropriately. Each follows directly from how the corpus was collected and annotated.

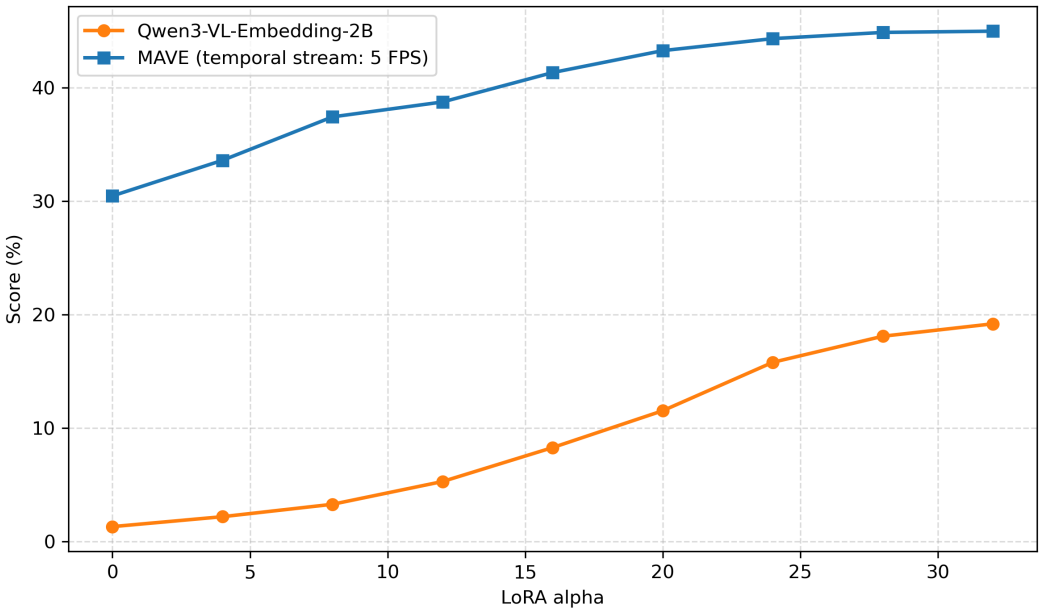


Figure 11: Effect of sweeping the LoRA scaling factor α on the performance of MAVE (temporal stream: 5 FPS) and the MAVR-adapted 2B Qwen3-VL-Embedding baseline on MAVR for the action-only ($\mathcal{A}$) queries on actions seen during training.

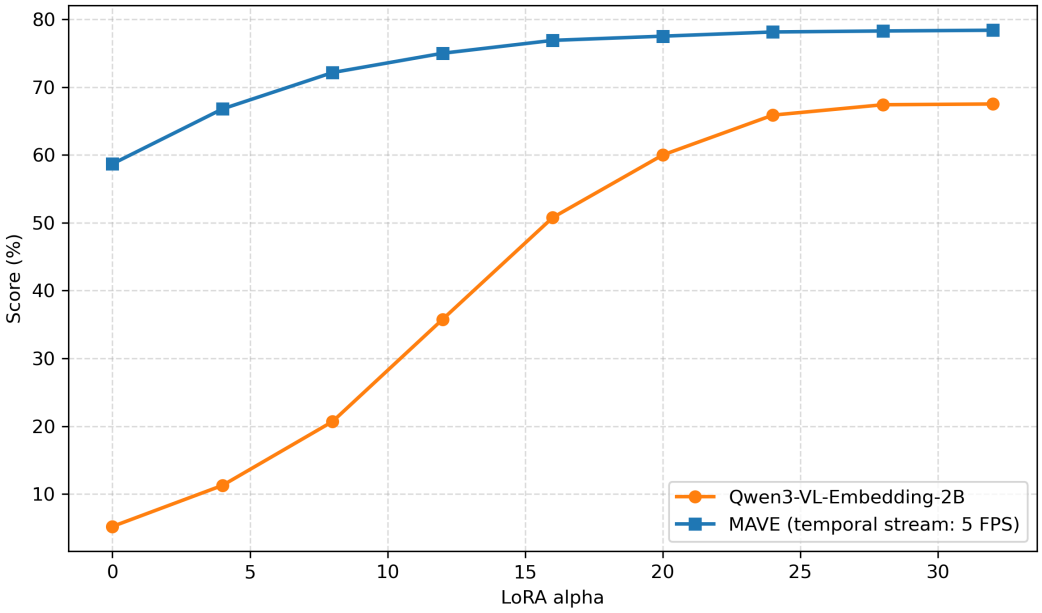


Figure 12: Effect of sweeping the LoRA scaling factor α on the performance of MAVE (temporal stream: 5 FPS) and the MAVR-adapted 2B Qwen3-VL-Embedding baseline on MAVR for the action-and-context ($\mathcal{A} + \mathcal{C}$) queries on actions seen during training.

Demographic and capture bias. Clips were contributed by crowd-workers whose demographic distribution was not controlled or balanced, so the corpus may be imbalanced across gender, ethnicity, and geography, and model performance may vary across populations. By construction it contains only single-person clips; models trained on it should not be expected to generalize to multi-person scenes. The footage is also strongly skewed toward indoor, third-person, static-camera consumer

Table 9: Effect of using anchoring regularization during MAVR adaptation of Qwen3-VL-Embedding-2B. All values are percentages (%). Standard recall@1 is computed for MMEB-V3, while MAVR reports target-sampled Recall@1 on action-only ($\mathcal{A}$) and action-and-context ($\mathcal{A}+\mathcal{C}$) queries for actions *seen* during training. Parentheses in the MMEB-V3 column indicate the change, in percentage points, relative to the frozen model.

Model	MMEB-V3	MAVR T2V		MAVR V2T	
		$\mathcal{A}$	$\mathcal{A}+\mathcal{C}$	$\mathcal{A}$	$\mathcal{A}+\mathcal{C}$
Frozen, unadapted	**55**	1	5	4	4
MAVR-adapted, no anchoring	36(-19)	**12**	49	**32**	53
MAVR-adapted, with anchoring	47(-8)	10	**50**	30	**55**

Statistical significance is computed across the MAVR metrics using a paired permutation test with Holm–Bonferroni correction. Both MAVR-adapted models significantly outperform the frozen model across all metrics, and the numerical differences between the two adapted models are statistically insignificant.

recordings (95.6% indoor, 97.1% third-person, 94.3% static camera), so coverage of outdoor scenes, egocentric viewpoints, professional capture, and unusual lighting is limited. This skew has a double-edged effect worth stating: it is part of why appearance is a weak discriminator on MAVR, which is the property the benchmark exploits, but it also means strong MAVR results do not establish robustness to visually diverse footage.

Long-tailed class distribution. Class sizes span nearly three orders of magnitude (10 to 7,825 clips, median 576), and air-drawn character classes alone account for 38.5% of the corpus. Although the test gallery is class-balanced (10 videos per action-class), the MAVR training set remains long-tailed; MAVR-adapted models may therefore receive substantially more supervision for frequent actions and action families, potentially affecting performance across them.

Limited action diversity. Although MAVR contains approximately $354{,}000$ videos, it covers only 287 action-classes. This limited action diversity may constrain training and generalization to motions outside MAVR's action-classes.

Small zero-shot test split. The zero-shot partition is small—eleven classes, 110 clips—so its computed metrics carry correspondingly wide error bars and should be read as indicative rather than precise.

Annotation quality. Action-class labels are human-reviewed, but captions are entirely machine-generated and have not been human-verified, so the scene attributes generated by a vision–language model might have errors in an unquantified fraction of rows. Because captions are template-assembled from a fixed clause order, they are also stylistically uniform and far less varied than free-form human queries; strong results under this caption style may not transfer directly to natural queries. A small number of clips retain minor article-agreement artifacts if the supplied placeholder object is substituted into the caption (Appendix C).

Hard-negative mining redundancy. Participants may contribute multiple similar takes of the same action under nearly identical scene and viewpoint conditions. Because hard-negative mining ranks candidate clips independently, several such near-duplicate takes of one action may be selected for the same anchor clip. Although each candidate remains a valid different-action negative, this can reduce the action and contextual diversity of each clip's hard-negative set.

Filtering limitations. The single-person constraints rely on automated face detection over sampled frames, not on identity documents or exhaustive human review, so errors are possible: a clip may include a second person visible only in unsampled frames. We also did not run automated logo, trademark, or incidental-PII detection (for example: visible documents, screens, or house numbers), so such content may appear incidentally in the background of some clips.

Task scope. MAVR isolates motion-sensitive retrieval by construction, which bounds what it can measure. Every clip is a single person performing a single prompted action in a short, roughly three-second window, so the benchmark says nothing about multi-agent interaction, long-horizon temporal structure, compositional multi-step activities, or retrieval over untrimmed video. It should be read as a targeted diagnostic for fine-grained motion sensitivity, complementary to general-purpose retrieval suites rather than a replacement for them.

K.2 MAVE and Experimental Limitations

We discuss the different limitations for the proposed MAVE architecture and the experiments run in this work.

Repurposed VLMs as embedding models. To broaden model coverage in our MAVR and MMEB-V3 benchmarking, we repurpose several general-purpose VLMs as embedding models using the final-layer hidden state of the final token. Because these representations are not trained specifically for retrieval, their near-zero MAVR performance may reflect the absence of retrieval training in addition to limited motion encoding.

Computational scalability of MAVE. MAVE's temporal pathway processes videos at a higher frame rate than its visual pathway, and this denser sampling contributes to its seen-action gains. To keep training and inference practical, we use a relatively lightweight temporal encoder (V-JEPA2-300M) and compress its outputs into a fixed number of temporal tokens. Scaling to a significantly larger temporal encoder would increase the computational cost of processing high-frame-rate videos and could diminish MAVE's efficiency advantage relative to processing more frames directly through the MLLM's visual pathway.

Training-schedule confound. MAVE uses a warm-start stage that trains its newly introduced temporal modules before LoRA adaptation of the embedding backbone, whereas the visual-only baselines use only the LoRA-active stage. Although the baselines and MAVE use the same LoRA-adaptation schedule, MAVE receives additional optimization of its temporal modules during the warm-start. Consequently, some portion of MAVE's observed gains may be attributable to the additional optimization provided by the warm-start stage.