SWIFTVMR: TOWARD EFFICIENT AND EFFECTIVE MLLM-BASED VIDEO MOMENT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

When repurposed for Video Moment Retrieval (VMR), Multimodal Large Language Models (MLLMs) deliver strong cross-modal alignment yet inherit significant efficiency burdens and a representational gap. Directly applying standard generative paradigms to VMR poses four major challenges to efficient and effective inference: 1) dense video sampling triggers a catastrophic visual token explosion; 2) auto-regressive decoding introduces inherent sequential overhead; 3) discrete-token formulations introduce a representational gap when modeling the continuous, ambiguous nature of action boundaries; and 4) massive parameter scales prohibit resource-constrained deployment. To address them, we propose SwiftVMR, a framework retaining the MLLM's alignment prowess while replacing iterative decoding with single-shot, continuous boundary prediction. To tackle the token explosion, a Zero-initialized Motion-aware Spatial Compressor (Z-MSC) combines content-adaptive grid pooling with a spatiotemporal residual path to reduce visual tokens while preserving spatial structures and training stability. To overcome the sequential decoding latency and representational mismatch of autoregressive generation, we introduce a Distribution-Aware Perception Head (DAPH), a single-shot logit-based architecture directly modeling probability distributions over temporal proposals. Finally, to overcome the deployment barrier, we introduce a Dual-Space Structured Distillation (DSSD) strategy to distill a high-capacity teacher into a lightweight student. Leveraging the continuous logit space provided by DAPH, DSSD performs foreground-weighted feature alignment and quality-gated distribution distillation. Extensive experiments on four benchmarks demonstrate that SwiftVMR matches state-of-the-art performance while accelerating post-ViT inference by an order of magnitude, proving that efficiency can be regained without sacrificing performance.

1 INTRODUCTION

The proliferation of unedited, lengthy video content has made fine-grained temporal localization a pressing demand in multimodal understanding (Wu et al., 2024; Zhou et al., 2025). At the core of this challenge lies Video Moment Retrieval (VMR) (Anne Hendricks et al., 2017; Gao et al., 2017; Wu et al., 2025a), a task dedicated to predicting the precise start and end boundaries of video segments that align with unconstrained natural language queries. Historically, the architectural evolution of VMR initially diverged into proposal-based (Zhang et al., 2020b; Yan et al., 2023) and proposal-free (Zhang et al., 2020a) paradigms. More recently, the field has transitioned towards elegant end-to-end set prediction paradigms. In particular, DETR-based frameworks (Lei et al., 2021) have gained significant traction. By conceptualizing learnable object queries as dynamic proposals, these models establish a robust foundation.

Driven by the revolutionary capabilities of Multimodal Large Language Models (MLLMs), a new wave of research (Wang et al., 2025a; Lu et al., 2024; Meinardus et al., 2024) eagerly attempts to unify VMR within these massive foundation models. However, applying standard generative paradigms to VMR introduces four major challenges: (1) Preserving sufficient temporal resolution necessitates dense video sampling, which triggers a catastrophic **visual token explosion**. (2) Auto-regressive decoding introduces inherent sequential overhead. (3) Standard discrete-token formulations introduce a representational gap when modeling the continuous, ambiguous nature of boundaries. (4) The massive parameter scales of foundation models prohibit resource-constrained

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.

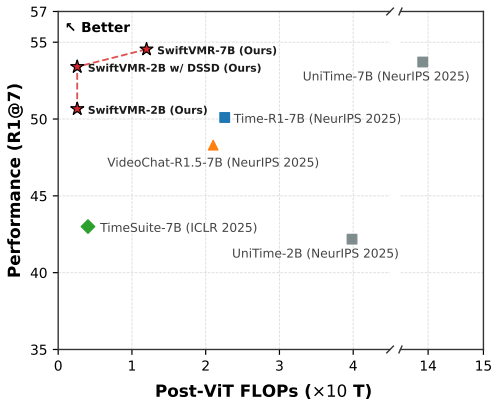


Figure 1: Performance vs. post-ViT efficiency trade-off on Charades-STA benchmark. Compared to recent generative MLLM baselines, our SwiftVMR achieves highly competitive performance while drastically reducing post-ViT computational overhead.

deployment. While prior VMR efforts have largely pursued accuracy within this generative regime, the efficiency dimension has been notably overlooked.

To systematically address them, we introduce SwiftVMR, an efficient MLLM framework that pursues both extreme post-ViT inference speed and high effectiveness (as highlighted in Figure 1). Fundamentally, SwiftVMR departs from generative paradigms by repurposing the MLLM into a unified, single-shot perception foundation. It gracefully retains the foundation model's powerful multimodal alignment capabilities while completely bypassing the cumbersome auto-regressive decoding process. Specifically, to tackle the token explosion (*Challenge 1*), a Zero-initialized Motion-aware Spatial Compressor (Z-MSC) is introduced. By combining content-adaptive grid pooling with a zero-initialized, multi-scale spatiotemporal residual path, Z-MSC reduces visual tokens before they enter the LLM. This preserves crucial spatial-temporal contexts without disrupting the foundation model's early training stability.

To circumvent the sequential decoding latency and discrete representational mismatch (*Challenges 2 & 3*), we leverage a causal `[Text][Video]` prompt and introduce the Distribution-Aware Perception Head (DAPH) for efficient and effective inference. By allowing visual tokens to attend to the preceding linguistic semantics, we obtain query-aware multimodal representations. Processing these features, DAPH predicts probabilities over temporal bins in a single forward pass and decodes continuous coordinates by expectation. This formulation bypasses autoregressive timestamp decoding while retaining the bin probabilities as a fine-grained interface for subsequent model-level distillation.

Finally, to overcome the deployment barrier (*Challenge 4*), we propose Dual-Space Structured Distillation (DSSD). Leveraging the continuous logit space natively produced by DAPH, DSSD distills a high-capacity MLLM teacher into a lightweight but highly competitive student via foreground-weighted feature alignment and quality-gated distribution distillation. This yields a deployable compact model that retains the precise performance of its heavy counterpart.

To sum up, the key contributions of our work are as follows:

- We propose SwiftVMR, a systematically efficient MLLM framework that explicitly addresses the token explosion, sequential latency, representational gap, and deployment barriers of generative VMR paradigms.
- We design the Zero-initialized Motion-aware Spatial Compressor (Z-MSC) to compress spatial redundancy through content-adaptive pooling, retaining temporal dynamics and MLLM fine-tuning stability.
- We introduce the Distribution-Aware Perception Head (DAPH) to replace autoregressive decoding with single-shot distribution prediction, decoding continuous boundaries by expectation and providing distributional targets for distillation.

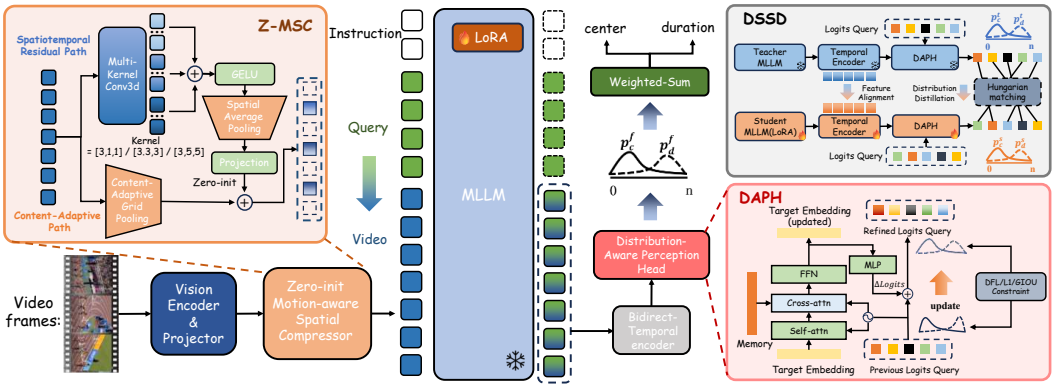


Figure 2: SwiftVMR framework, a single-forward MLLM pipeline for VMR. The Zero-initialized Motion-aware Spatial Compressor (Z-MSC) combines a Content-Adaptive Path with a Spatiotemporal Residual Path to compress spatial tokens. Compressed video tokens follow the instruction and query in causal `[Text][Video]` order through a LoRA-finetuned MLLM. A Bidirectional Temporal Encoder refines the final video hidden states to recover global temporal context. The Distribution-Aware Perception Head (DAPH) then predicts boundary probability distributions, decoded into timestamps by expectation. Dual-Space Structured Distillation (DSSD) transfers knowledge from a high-capacity teacher to a lightweight student through foreground-weighted feature alignment in the low-dimensional VMR-specific space and quality-gated logit-level distillation over DAPH probability curves.

- We develop a Dual-Space Structured Distillation (DSSD) strategy that exploits DAPH's continuous logit space, successfully transferring a high-capacity teacher into a lightweight student. SwiftVMR-7B achieves state-of-the-art results in our four-benchmark comparison, while delivers an order-of-magnitude post-ViT speedup.

2 Related Works

Video Moment Retrieval (VMR). VMR aims to localize specific temporal segments in untrimmed videos that semantically correspond to a given natural language query. Early literature primarily divides into two mainstream paradigms: proposal-based and proposal-free methods. Proposal-based methods (Zhang et al., 2020b; Shen et al., 2023; Sun et al., 2022) typically rely on pre-defining a dense set of candidate temporal segments and subsequently scoring their multimodal alignment with the query semantics. Conversely, proposal-free approaches (Zhang et al., 2020a; Huang et al., 2022; Lin et al., 2025; Dong et al., 2024) eliminate the need for heuristic anchors, directly predicting the start and end boundaries to bypass the heavy computational overhead of dense proposals. Recently, inspired by the success of object detection, DETR-style architectures (Lei et al., 2021; Moon et al., 2023; Zhao et al., 2025; Cao et al., 2025; Gordeev et al., 2026; Sun et al., 2024) have gained significant traction by formulating VMR as a direct set-prediction problem. However, most of these works predominantly focus on designing intricate text-video interaction modules or complex cross-attention fusion mechanisms before decoding. Relying on relatively small, task-specific backbones, they inherently lack the deep semantic reasoning capabilities offered by modern foundation models.

MLLM-based VMR. The advent of Multimodal Large Language Models (MLLMs) (Wang et al., 2024b; Li et al., 2024; Bai et al., 2025; Wang et al., 2025b) has introduced a new paradigm leveraging foundation models' massive reasoning capabilities for VMR (Jiang et al., 2024; Guo et al., 2025; Wu et al., 2025b; Tao & Huang, 2026; Pramanick et al., 2025; Yan et al., 2025; Wang et al., 2024c; Zeng et al., 2025). Early explorations like TimeChat (Ren et al., 2024) adapted LLMs for temporal retrieval via instruction tuning, while subsequent advancements like TRACE (Guo et al., 2024) and UniTime (Li et al., 2025b) significantly refined this generative paradigm. Recent pioneer works (Zhang et al., 2026; Li et al., 2025a; Wang et al., 2025c; Liu et al., 2026) further employ Reinforcement Learning (RL) to directly optimize retrieval policies. Ultimately, benefiting from

powerful innate multimodal understanding, these MLLM-based methods have robustly surpassed traditional VMR models.

Nevertheless, dense video sampling and auto-regressive decoding incur visual token explosion and sequential overhead. Next-token prediction also mismatches continuous, fuzzy action boundaries, while large model sizes restrict deployment. We propose **SwiftVMR**, which equips MLLMs with a lightweight, single-shot perception interface. It addresses these challenges through content-adaptive, motion-aware spatial compression, logit-based distribution perception, and structured dual-space distillation, retaining MLLMs' multimodal alignment capabilities while reducing inference and deployment costs.

3 METHOD

3.1 OVERVIEW

Problem Formulation. VMR aims to localize a specific temporal boundary $B = (\tau_c, \tau_d)$ that semantically matches a textual query $Q = \{w_i\}_{i=1}^{L_q}$ within an untrimmed video $V = \{v_t\}_{t=1}^{T_v}$. Here, w_i and v_t denote the i-th word and t-th video frame, respectively, while τ_c and τ_d represent the center and duration timestamps.

Overall framework. We present SwiftVMR, an efficient MLLM framework tailored for VMR. It combines visual token compression, non-autoregressive prediction, and model-level distillation to accelerate post-ViT inference and enable lightweight deployment (Figure 2). This unified pipeline addresses both computational bottlenecks and the parameter scale of conventional generative MLLMs. Unlike generative methods that quantize boundaries into discrete text tokens, it predicts temporal-bin probabilities, which support continuous coordinate prediction by expectation and provide an interface for downstream distillation.

The pipeline begins with the Zero-initialized Motion-aware Spatial Compressor (Z-MSC), which combines content-adaptive pooling with spatiotemporal residuals to reduce visual tokens. A causal `[Text][Video]` prompt then leverages the MLLM's cross-modal capabilities for multimodal alignment. The Distribution-Aware Perception Head (DAPH) performs logit-based temporal boundary decoding, and Dual-Space Structured Distillation (DSSD) transfers knowledge to lightweight student models.

3.2 ZERO-INITIALIZED MOTION-AWARE SPATIAL COMPRESSOR

In VMR, precise boundary retrieval strictly demands high temporal resolution and a strong sensitivity to motion dynamics. To alleviate this visual token explosion without sacrificing critical temporal cues, we propose the Zero-initialized Motion-aware Spatial Compressor (Z-MSC), which drastically compresses the spatial dimensions of each frame to a fixed size while effectively retaining inherent temporal dynamics and essential spatial structures.

The **Content-Adaptive Path** pools visual features X ($T \times H \times W \times C$) onto a regular $H' \times W'$ grid per time step. A lightweight query-agnostic function g scores tokens in each region $\mathcal{R}_j$:

$$
\begin{aligned}
s_{t,p} &= \kappa \tanh(g(\mathrm{LN}(x_{t,p}))/\kappa)\,, \\
a_{t,p}^{(j)} &= \frac{\exp(s_{t,p})}{\sum_{r\in\mathcal{R}_j}\exp(s_{t,r})}, \qquad (V_c)_{t,j} = \sum_{p\in\mathcal{R}_j} a_{t,p}^{(j)} x_{t,p}.
\end{aligned} \tag{1}
$$

A fixed $\kappa > 0$ smoothly bounds scores for stable normalization, retaining local linearity near zero. The **Spatiotemporal Residual Path** captures multi-scale motion through parallel depthwise 3D convolutions $\mathrm{Conv3d}_k$ with kernels $(3, k, k)$, $k \in \{1, 3, 5\}$, followed by summation, GELU, spatial average pooling, and projection:

$$
V_r = \mathrm{Proj}(\mathrm{Pool}(\mathrm{GELU}(\sum_{k\in\{1,3,5\}} \mathrm{Conv3d}_k(X)))). \tag{2}
$$

We output $V_{\text{comp}} = V_c + V_r$. To preserve the initial feature space, we **zero-initialize** the residual projection, following ControlNet (Zhang et al., 2023), and the scorer's final linear layer. Thus $V_r = 0$ and V_c uses uniform region weights at initialization. Z-MSC starts as spatial average pooling and learns content-adaptive aggregation and spatiotemporal residuals with an unchanged temporal length and spatial token budget.

3.3 Our Non-Autoregressive VMR Backbone

MLLM-based Multimodal Alignment. To efficiently leverage the MLLM's multimodal reasoning capabilities without introducing complex cross-attention networks, we adopt a causal `[Text][Video]` prompt paradigm. Governed by the MLLM's unidirectional causal attention, positioning the textual query ahead of the video sequence forces the subsequent visual tokens to naturally attend to the linguistic semantics during the forward pass. Consequently, these tokens are inherently transformed into query-aware multimodal representations.

Upon completing the forward pass, we extract the last-layer hidden states corresponding exclusively to the video tokens. However, the causal masking inherently restricts early video frames from perceiving future contexts, which is detrimental to precise boundary localization. To address this structural limitation, we feed the extracted states into a Bidirectional Temporal Encoder. This lightweight module reconstructs the unmasked global temporal dynamics, yielding a comprehensive and bidirectional sequence representation $\hat{V}$ ready for the subsequent decoding stage.

Distribution-Aware Perception Head. The refined multimodal representations $\hat{V}$ enter DAPH for single-shot, logit-based prediction. This design serves two objectives. First, single-pass prediction removes the sequential timestamp-decoding loop. Second, predicting probabilities over temporal bins retains information beyond the decoded coordinates, providing a fine-grained interface for model-level distillation. Distribution-based localization and iterative refinement have been explored in object detection (Li et al., 2020; Peng et al., 2025). Unlike D-FINE, DAPH directly represents normalized temporal center and duration on a shared uniform support. Corresponding bins thus have consistent temporal meanings across teacher and student, independently of their initial predictions.

Concretely, we depart from DETR-style decoders (Moon et al., 2023) that regress rigid coordinates with $N_q \times 2$ queries. We instead discretize the temporal duration into N_{bin} uniform bins and expand each query to $N_q \times N_{bin}$ logits. Each decoder layer iteratively refines these distributions by cross-attending to the memory $\hat{V}$, and the final boundary equals the expectation over the softmaxed logits.

L1 and GIoU supervise the expected coordinates but do not uniquely determine the predicted distributions: different probability curves can yield exactly the same expectation. We therefore adopt Distribution Focal Loss (DFL) (Li et al., 2020) to provide additional supervision directly on the bin probabilities. For a continuous ground-truth target y falling between its two nearest discrete bins y_l and y_r, DFL weights their cross-entropy losses according to the target's distance from each bin:

$$\mathcal{L}_{dfl} = (y_r - y)\mathcal{L}_{ce}(z_l, y_l) + (y - y_l)\mathcal{L}_{ce}(z_r, y_r) \tag{3}$$

where $\mathcal{L}_{ce}$ denotes the standard cross-entropy loss, while z_l and z_r represent the predicted logits of the respective bins.

These losses supervise complementary aspects of localization: L1 and GIoU constrain the decoded segment, while DFL constrains its underlying bin probabilities. The resulting distributions also enable quality-gated distillation: matching the teacher's soft probabilities transfers information across temporal bins that is not retained when only its expected coordinates are used as targets.

3.4 Dual-Space Structured Distillation

Beyond inference latency, the other barrier to deploying MLLM-based VMR is the model's parameter scale: a multi-billion-parameter backbone is impractical on resource-constrained devices. Dual-Space Structured Distillation (DSSD) targets exactly this axis, distilling a high-capacity MLLM teacher into a lightweight MLLM student so that the efficient single-forward pipeline above can run at a fraction of the cost. The transfer operates across two synergistic spaces, the encoded feature space and the decoded probability-distribution space, where the latter is made possible by the soft logit output of DAPH.

Foreground-Weighted Feature Alignment. We align in the low-dimensional VMR-specific space after the Bidirectional Temporal Encoder rather than the raw high-dimensional LLM hidden states: the latter carry broad open-world semantics that overburden the student without benefiting VMR, so transfer is both harder and less effective there (Table 2(c)). Since untrimmed videos contain substantial background noise, we further weight the alignment by a foreground mask M_{fg} (the core action with slight dilation) and its complement M_{bg}:

$$\mathcal{L}_{feat} = \alpha_{fg} \cdot \text{MSE}(\mathcal{F}^S_{fg}, \mathcal{F}^T_{fg}) + \beta_{bg} \cdot \text{MSE}(\mathcal{F}^S_{bg}, \mathcal{F}^T_{bg}) \tag{4}$$

where $\mathcal{F}^S$ and $\mathcal{F}^T$ denote the student and teacher features, respectively. By setting $\alpha_{fg} > \beta_{bg}$ (e.g., 0.8 and 0.2), we compel the student to prioritize the alignment of critical action dynamics while maintaining secondary awareness of the environmental context.

Quality-Gated Distribution Distillation. DAPH provides the teacher's probabilities over temporal bins in addition to the expected coordinates. Predictions and ground truths are paired via Hungarian matching. Distilling low-quality teacher predictions may introduce noisy targets, so we retain matched queries whose teacher-decoded boundary has IoU $> \tau_q$ with the ground truth. For these queries, we minimize the KL divergence between softened teacher and student distributions:

$$\begin{aligned}\mathcal{L}_{logit} = \sum_{i=1}^{B} \tau^2 (&KL(\sigma(p^s_c/\tau), \sigma(p^t_c/\tau)) \\ &+KL(\sigma(p^s_d/\tau), \sigma(p^t_d/\tau))),\end{aligned} \tag{5}$$

where B is the number of valid matched pairs passing the competence gate, τ represents the temperature hyperparameter to soften the distributions, $\sigma(\cdot)$ is the Softmax function, and (p^s_c, p^s_d) and (p^t_c, p^t_d) denote the center and duration boundary logits of the student and teacher, respectively.

In addition, following Lei et al. (2021), we introduce a saliency loss $\mathcal{L}_{sal}$ applied to the refined multimodal representations $\hat{V}$. Furthermore, to explicitly align the predicted confidence scores of individual queries with their true localization quality, we incorporate a quality loss $\mathcal{L}_{quality}$ (Lee & Byun, 2024). Ultimately, the overall training objective $\mathcal{L}_{total}$ is formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{ret} + \lambda_1 \mathcal{L}_{distill} + \lambda_2 \mathcal{L}_{sal} + \lambda_3 \mathcal{L}_{quality} \tag{6}$$

where the retrieval loss $\mathcal{L}_{ret}$ comprises L1, GIoU, and the adopted DFL loss. The distillation loss $\mathcal{L}_{distill}$ combines feature alignment $\mathcal{L}_{feat}$ and distribution distillation $\mathcal{L}_{logit}$. The hyperparameters λ_1, λ_2, and λ_3 balance distillation, saliency supervision, and quality estimation, respectively.

4 Experiment

4.1 Experimental Setups

Datasets. We conduct comprehensive evaluations of our proposed framework on four widely adopted video moment retrieval benchmarks: Charades-STA (Gao et al., 2017), ActivityNet Captions (Krishna et al., 2017), QVHighlights (Lei et al., 2021), and TACoS (Gao et al., 2017).

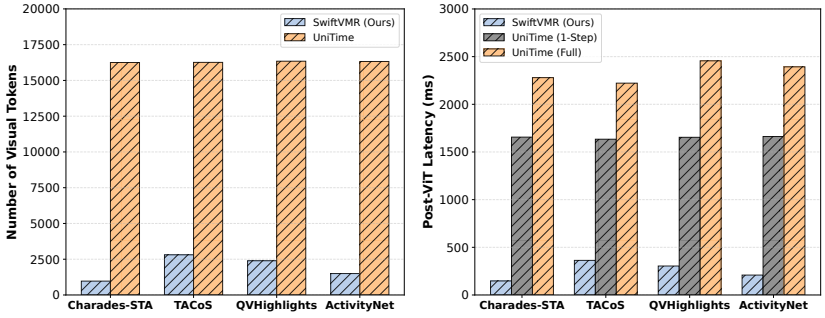
Dataset descriptions and statistics are provided in Appendix D.

Evaluation Metrics. Following the previous works (Gao et al., 2017; Liu et al., 2018; Yuan et al., 2019), we evaluate our model using Rn@μ and mean IoU (mIoU). Specifically, Rn@μ measures the percentage of test queries where at least one of the top-n retrieved moments achieves a temporal Intersection over Union (IoU) with the ground truth strictly exceeding a predefined threshold $\mu * 0.1$. Meanwhile, mIoU represents the average temporal overlap across all testing samples. In our experiments, we report the performance at $n = 1$ with thresholds $\mu \in \{3, 5, 7\}$.

Implementation Details. Our SwiftVMR framework is implemented in PyTorch, adopting Qwen2-VL (Wang et al., 2024a) as the foundation model for fair comparison. During training, the visual encoder is strictly frozen to preserve the pre-trained spatial semantics. To achieve efficient multimodal alignment, the LLM backbone is fine-tuned using LoRA (Hu et al., 2021) (rank=128,alpha=256). We adopt a differentiated learning rate strategy optimized by AdamW (Loshchilov & Hutter, 2017):

Table 1: Comparison with state-of-the-art methods on four benchmarks. The top section presents traditional high-performance models, including those utilizing DETR-style decoders. The middle section details recent MLLM-based generative approaches. For a fair and rigorous comparison, we adopt MLLM baselines equipped with **7B**-parameter backbones and cite their fine-tuned performance on the corresponding benchmark as reported in their respective original papers. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	QVHighlights			ActivityNet Captions			Charades-STA			TACoS		
	R1@5	R1@7	mIOU	R1@5	R1@7	mIOU	R1@5	R1@7	mIOU	R1@3	R1@5	mIOU
QD-DETR (CVPR 2023)	62.40	44.98	-	-	-	-	55.51	34.17	-	-	-	-
TR-DETR (AAAI 2024)	64.66	48.96	-	-	-	-	53.47	30.81	-	-	-	-
LD-DETR (ArXiv 2025)	66.80	51.04	-	-	-	-	62.58	41.56	53.44	57.61	44.31	40.30
Flash-VTG (WACV 2025)	73.10	57.29	-	-	-	-	70.32	49.87	-	53.71	41.76	37.61
SG-DETR (WACV 2026)	72.80	59.50	-	-	-	-	71.10	52.80	60.70	-	46.40	42.40
TimeChat (CVPR2024)	-	-	-	-	-	-	46.70	23.70	-	-	-	-
HawkEye (ArXiv 2024)	-	-	-	34.70	17.90	39.10	58.30	28.80	49.30	-	-	-
ED-VTG (ICCV 2025)	-	-	-	45.10	22.70	44.90	62.10	35.00	52.60	46.00	31.50	32.40
TimeSuite (ICLR 2025)	-	-	-	-	-	-	67.10	43.00	-	-	-	-
VideoChat-R1.5 (NeurIPS 2025)	-	-	-	-	-	-	71.60	48.30	60.60	-	-	-
Time-R1 (NeurIPS 2025)	-	-	-	55.60	34.00	-	72.20	50.10	-	-	-	-
AdaVTG-LLM (CVPR 2026)	-	-	-	45.20	22.90	45.50	63.30	35.60	53.00	46.70	32.30	31.80
UniTime (NeurIPS 2025)	77.76	63.29	-	54.81	36.62	-	74.33	53.71	63.15	61.18	48.31	45.02
SwiftVMR-2B	77.32	63.06	69.80	55.17	35.70	51.94	71.77	50.67	60.74	62.61	47.14	43.56
SwiftVMR-2B w/ DSSD	78.35	65.19	71.08	56.18	36.83	52.10	73.25	53.39	62.05	64.23	49.24	45.85
SwiftVMR-7B	**79.74**	**66.71**	**72.03**	**58.39**	**38.36**	**54.26**	**74.76**	**54.54**	**63.32**	**68.08**	**53.44**	**48.74**



(a) Visual token comparison. (b) Post-ViT inference latency per query.

Figure 3: Efficiency comparison of visual token counts and post-ViT inference latency under the same 7B foundation model. (a) shows the comparison of the number of visual tokens fed to LLM, and (b) represent the post-ViT inference latency per query. Our SwiftVMR achieves an order-of-magnitude reduction in post-ViT inference time compared to the SOTA method, UniTime. Both methods use identical acceleration settings (BF16 and FlashAttention).

the LoRA parameters in the MLLM are updated with a learning rate of 5e-5, while all newly initialized structural modules (i.e., Z-MSC, DAPH, and Bidirectional Temporal Encoder) are trained at a higher rate of 1e-4. All experiments are conducted with a total batch size of 32.

4.2 Comparison Results

Table 1 compares SwiftVMR with traditional VMR and recent MLLM-based methods. SwiftVMR-7B achieves state-of-the-art results across all four benchmarks within this comparison, leading all metrics with available external baseline values.

Furthermore, to validate the effectiveness of our proposed distillation strategy, we report the performance of our student model after distillation. A direct comparison reveals that our Dual-Space Structured Distillatio (DSSD) consistently improves the student across all four benchmarks (Table 1), demonstrating effective transfer of structural knowledge and localization capabilities from teacher to student.

Table 2: Ablation studies of our proposed components on QVHighlights.

Line ID	Approach	R1@3	R1@5	R1@7	mIOU
Baseline (7B): Avg. Pooling + Regressive Head		88.02	78.42	63.91	70.27
(a) Z-MSC (7B)					
#1	w/o Spatiotemporal Residual Path	89.35	79.28	65.34	71.50
#2	w/o Content-Adaptive Path	88.28	79.31	64.97	70.64
#3	w/o Zero-Init	87.52	77.45	64.35	69.66
#4	Ours	**89.74**	**79.74**	**66.71**	**72.03**
(b) Prompt paradigm and DAPH (7B)					
#1	`[Video][Text]`	87.35	75.74	60.26	68.13
#2	Regressive Head	89.58	79.03	65.53	71.45
#3	w/o DFL	88.91	78.97	65.74	71.63
#4	Ours	**89.74**	**79.74**	**66.71**	**72.03**
(c) DSSD (2B)					
#1	w/o DSSD	86.84	77.32	63.06	69.80
#2	w/o logits	88.81	**78.38**	64.88	70.55
#3	w/o feature	88.03	78.29	64.40	70.34
#4	Coordinate Distillation Only	87.62	77.73	63.97	70.18
#5	Feature + Coordinate Distillation	88.50	78.01	64.92	70.07
#6	feature w/o weighted	88.55	77.30	64.21	70.29
#7	feature in LHS	87.87	76.78	63.53	69.98
#8	Ours	**89.26**	78.35	**65.19**	**71.08**

4.3 Efficiency Verification

While Figure1 demonstrates our macro-level superiority in computational cost versus localization accuracy, this section delves into the granular metrics that dictate real-world deployability: visual token footprint and online inference latency.

Visual Token Reduction. As illustrated in Figure 3(a), generative baselines typically map raw videos into massive token sequences to retain spatiotemporal details, inherently bottlenecking the foundation model. In stark contrast, our Z-MSC module achieves a dramatic token reduction by aggressively compressing the spatial dimensions while fully preserving essential spatial structures and temporal dynamics. Consequently, SwiftVMR consistently reduces the number of visual tokens by approximately 90% across all four evaluated benchmarks compared to UniTime, the state-of-the-art generative models. It also reflects that SwiftVMR achieves a particularly low memory footprint.

Accelerated Post-ViT Inference. Beyond memory footprint, post-ViT inference latency is the ultimate crucible for real-world deployment. As shown in Figure 3(b), evaluated on a single NVIDIA A100-80G GPU, SwiftVMR radically outpaces UniTime, achieving a $15.5\times$ post-ViT speedup (147.2ms vs. 2279.7ms per query on Charades-STA, from the stage totals in Table 3).

Importantly, this extreme efficiency is the result of a meticulously designed synergistic pipeline rather than a single architectural trick. To isolate the impact of the decoding strategy, we constrained UniTime to generate only a single token. Even under this strict setting, UniTime still consumes approximately 1656ms in the post-ViT stages. Restricting generation to one token reduces post-ViT latency by 27.4%. The majority of the speedup is unlocked by Z-MSC, which shrinks the massive attention computation matrix within the LLM. Finally, by coupling these architectural efficiencies with Dual-Space Structured Distillation (DSSD), SwiftVMR becomes highly compact for real-world deployment.

4.4 Ablation Studies

As shown in Table 2, to validate the efficacy of our proposed components, we conduct extensive ablation studies on QVHighlights with 7B foundation models, while DSSD ablations are conducted with 2B foundation models. The full 7B model outperforms the average-pooling baseline with a regressive head on all four metrics.

Design of Z-MSC. The Spatiotemporal Residual Path captures multi-scale motion, while the Content-Adaptive Path aggregates spatial details. Removing either path lowers all four metrics (Table 2(a)). The w/o Zero-Init variant retains both paths but disables zero-initialization in both, also reducing performance.

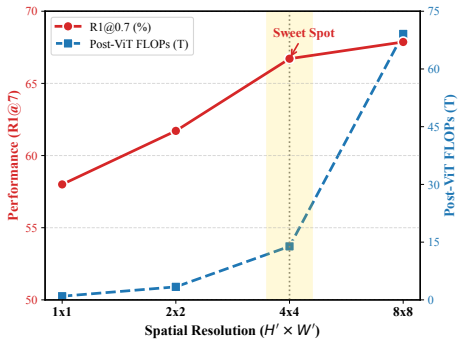


Figure 4: Ablation studies of spatial resolution of Z-MSC.

Spatial Resolution of Z-MSC. To explore the optimal trade-off between visual token compression and the preservation of semantic details, we ablate the target compressed spatial dimensions ($H' \times W'$) within the Z-MSC module. We evaluate a range of output resolutions, varying from an aggressive 1×1 downsampling to a more detailed 8×8 feature map per frame. The results, illustrated in Figure 4. While increasing the spatial resolution generally leads to improved performance due to richer visual features, it simultaneously incurs a significant rise in computational cost due to the increased visual token counts. Conversely, excessive compression harms performance by discarding crucial spatial structures. Empirically, a resolution of 4×4 emerges as the distinct "sweet spot" It offers the best balance, delivering high performance while maintaining the high efficiency.

Early Fusion Paradigm and DAPH. Next, we compare the architectural paradigm and perception-head configurations. Table 2(b) includes a decoupled `[Video][Text]` paradigm with QD-DETR cross-attention for late visual-textual interaction, a standard deterministic regression head from QD-DETR, and DAPH without DFL. These variants examine the prompt order, boundary parameterization, and distribution supervision, respectively. Our full model scores higher than these variants on all four metrics, supporting query-aware early fusion and the combination of distribution prediction with DFL supervision.

Efficacy of DSSD Components. Table 2(c) supports foreground weighting and low-dimensional alignment over unweighted alignment and last-hidden-state (LHS) alignment. Both coordinate baselines replace KL with L1/GIoU losses between decoded teacher and student segments, using the same student DAPH and quality gate. Coordinate-only omits feature alignment; Feature + Coordinate Distillation retains it. Distribution targets outperform coordinate targets on all four metrics both without and with feature alignment. This supports the additional value of distribution targets alongside feature alignment. Feature alignment yields larger gains than distribution-only distillation, possibly because the student already receives boundary supervision while feature alignment provides additional semantic guidance. Their combination achieves the highest R1@7 and mIoU. Qualitative analyses appear in Appendix F.

5 Conclusion

In this paper, we revisit VMR from an efficiency perspective that has been largely overlooked in the MLLM era. To systematically overcome the challenges of visual token explosion, sequential generation overhead, the discrete representational mismatch, and massive parameter scales, we propose SwiftVMR. Within this framework, Z-MSC combines content-adaptive spatial pooling with spatiotemporal residuals to compress visual tokens. DAPH reformulates localization into a single-shot logit-based architecture that directly models probability distributions over temporal proposals. Crucially, the logit space enables DSSD to distill a high-capacity teacher into a lightweight student, making the efficient pipeline deployable at a fraction of the cost. SwiftVMR achieves state-of-the-art results in our four-benchmark comparison, while delivers an order-of-magnitude post-ViT speedup.

AI USE STATEMENT

We used generative AI tools for translation, language polishing, and assistance with the analysis and interpretation of experimental results. The authors reviewed all AI-assisted analyses and edits to verify their accuracy and ensure that they preserve the scientific meaning, numerical results, uncertainty, and limitations. The authors take full responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe the model architecture and experimental setup in Sections 3 and 4.1. Dataset details and video sampling settings are provided in Appendices D and C, respectively. Additional hyperparameter analyses and the training prompt template are included in the appendix.

REFERENCES

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pp. 5803–5812, 2017.

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.

Zhuo Cao, Bingqing Zhang, Heming Du, Xin Yu, Xue Li, and Sen Wang. Flashvtg: Feature layering and adaptive score handling network for video temporal grounding. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 9226–9236. IEEE, 2025.

Jianfeng Dong, Xiaoman Peng, Daizong Liu, Xiaoye Qu, Xun Yang, Cuizhu Bao, and Meng Wang. Temporal sentence grounding with relevance feedback in videos. *Advances in Neural Information Processing Systems*, 37:43107–43132, 2024.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pp. 5267–5275, 2017.

Aleksandr Gordeev, Vladimir Dokholyan, Irina Tolstykh, and Maksim Kuprashevich. Saliency-guided detr for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 907–916, 2026.

Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*, 2024.

Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3302–3310, 2025.

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.

Jiabo Huang, Hailin Jin, Shaogang Gong, and Yang Liu. Video activity localisation with uncertainties in temporal boundary. In *European Conference on Computer Vision*, pp. 724–740. Springer, 2022.

Yiyang Jiang, Wengyu Zhang, Xulu Zhang, Xiao-Yong Wei, Chang Wen Chen, and Qing Li. Prior knowledge integration via llm encoding and pseudo event regulation for video moment retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7249–7258, 2024.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017.

Pilhyeon Lee and Hyeran Byun. Bam-detr: Boundary-aligned moment detection transformer for temporal sentence grounding in videos. In *European Conference on Computer Vision*, pp. 220–238. Springer, 2024.

Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in neural information processing systems*, 33:21002–21012, 2020.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025a.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. *Advances in Neural Information Processing Systems*, 38:64426–64455, 2025b.

Junan Lin, Daizong Liu, Xianke Chen, Xiaoye Qu, Xun Yang, Jixiang Zhu, Sanyuan Zhang, and Jianfeng Dong. Audio does matter: Importance-aware multi-granularity fusion for video moment retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6027–6036, 2025.

Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 15–24, 2018.

Wenqi Liu, Yunxiao Wang, Shijie Ma, Meng Liu, Qile Su, Tianke Zhang, Haonan Fan, Changyi Liu, Kaiyu Jiang, Jiankang Chen, et al. Videotemp-o3: Harmonizing temporal grounding and video understanding in agentic thinking-with-videos. *arXiv preprint arXiv:2602.07801*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Weiheng Lu, Jian Li, An Yu, Ming-Ching Chang, Shengpeng Ji, and Min Xia. Llava-mr: Large language-and-vision assistant for video moment retrieval. *arXiv preprint arXiv:2411.14505*, 2024.

Boris Meinardus, Hector Rodriguez, Anil Batra, Anna Rohrbach, and Marcus Rohrbach. Chrono: A simple blueprint for representing time in mllms. *arXiv preprint arXiv:2406.18113*, 2024.

WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23023–23033, 2023.

Yansong Peng, Hebei Li, Peixi Wu, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. D-FINE: Redefine regression task of DETRs as fine-grained distribution refinement. In *International Conference on Learning Representations*, 2025.

Shraman Pramanick, Effrosyni Mavroudi, Yale Song, Rama Chellappa, Lorenzo Torresani, and Triantafyllos Afouras. Enrich and detect: Video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24297–24308, 2025.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024.

Xingyu Shen, Xiang Zhang, Xun Yang, Yibing Zhan, Long Lan, Jianfeng Dong, and Hongzhou Wu. Semantics-enriched cross-modal alignment for complex-query video moment retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4109–4118, 2023.

Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4998–5007, 2024.

Xin Sun, Xuan Wang, Jialin Gao, Qiong Liu, and Xi Zhou. You need to read again: Multi-granularity perception network for moment retrieval in videos. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 1022–1032, 2022.

Te Tao and Shao-Lun Huang. Adavtg-llm: A videollm-based efficient video temporal grounding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3631–3642, 2026.

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

Qiyi Wang, Senda Chen, and Ying Shen. Causalvtg: Towards robust video temporal grounding via causal inference. *Advances in Neural Information Processing Systems*, 38:103609–103633, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *Advances in Neural Information Processing Systems*, 38:83330–83364, 2025c.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pp. 396–416. Springer, 2024b.

Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. Hawkeye: Training video-text llms for grounding text in videos. *arXiv preprint arXiv:2403.10228*, 2024c.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Jianlong Wu, Wei Liu, Ye Liu, Meng Liu, Liqiang Nie, Zhouchen Lin, and Chang Wen Chen. A survey on video temporal grounding with multimodal large language model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13754–13765, 2025b.

Shen Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David Ross, and Cordelia Schmid. Unloc: A unified framework for video localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13623–13633, 2023.

Ziang Yan, Yinan He, Xinhao Li, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *Advances in Neural Information Processing Systems*, 38:119152–119184, 2025.

Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9159–9166, 2019.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. Timesuite: Improving mllms for long video understanding via grounded tuning. In *International Conference on Learning Representations*, volume 2025, pp. 38057–38081, 2025.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*, 2020a.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10419–10429, 2026.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 12870–12877, 2020b.

Pengcheng Zhao, Zhixian He, Fuwei Zhang, Shujin Lin, and Fan Zhou. Ld-detr: Loop decoder detection transformer for video moment retrieval and highlight detection. *arXiv preprint arXiv:2501.10787*, 2025.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025.

SUPPLEMENTARY MATERIAL OF "SWIFTVMR: TOWARD EFFICIENT AND EFFECTIVE MLLM-BASED VIDEO MOMENT RETRIEVAL"

A COMPUTING COST EVALUATION

Following the computing-cost formulation established in DyCoke, we analytically estimate the post-ViT FLOPs of MLLM-based video moment retrieval models by decoupling the computation into the prefilling and decoding stages. Let n, d, and m denote the number of input tokens, the hidden state size, and the intermediate size of the Feed-Forward Network (FFN), respectively. For a single transformer layer, the total FLOPs in the prefilling phase can be estimated as:

$$\mathcal{F}\text{prefill} = 4nd^2 + 2n^2d + 2ndm. \tag{7}$$

For the decoding phase, considering the significant computational contribution of the KV cache, the consumption for R total auto-regressive iterations (i.e., predicting R tokens) per layer is calculated as:

$$\mathcal{F}\text{decode} = R\Big((4d^2 + 2dm) + 2\Big(d, n + \tfrac{d(R+1)}{2}\Big)\Big). \tag{8}$$

Consequently, for an LLM backbone comprising T total transformer layers, the overall computational cost is aggregated as follows:

$$\text{FLOPs} = T \times (\mathcal{F}\text{prefill} + \mathcal{F}\text{decode}). \tag{9}$$

To ensure a rigorous and fair comparison, all variables utilized in these FLOPs measurements (such as the input sequence length n and the generation steps R) are strictly derived from the actual benchmark.

Table 3: End-to-end computational profiling on Charades-STA (7B foundation model). Both methods use identical acceleration settings (BF16 and FlashAttention).

Component	UniTime	SwiftVMR	Approach	Δ
Vision Encoder (ms)	991.8	646.0	-	-345.8
Token Compression (ms)	0.0	12.1	Z-MSC	+12.1
LLM Forward (ms)	1577.8	124.1	-	-1453.7
Decoder (ms)	701.9	11.0	BiTE + DAPH	-690.9
Total (ms)	3271.5	793.2	-	-2478.3
GPU Memory (MB)	66119	25962	-	-40157

B END-TO-END LATENCY COMPARISON

To provide a more comprehensive analysis of the computational efficiency, we further compare the end-to-end inference latency and GPU memory consumption between UniTime and our proposed SwiftVMR. As summarized in Table 3, SwiftVMR reduces total inference latency and peak GPU memory, accelerating both the LLM forward and decoding stages despite the added token-compression overhead.

Specifically, the overall inference latency is reduced from 3271.5 ms to 793.2 ms, corresponding to a 75.8% reduction, while the peak GPU memory decreases from 66.1 GB to 26.0 GB (60.7% reduction). The largest latency improvement comes from the LLM forward stage, whose latency is reduced by more than 92%, indicating that the computational bottleneck of MLLM-based VMR is effectively alleviated. Moreover, replacing autoregressive generation with our Bidirectional Temporal Encoder (BiTE) and Distribution-Aware Perception Head (DAPH) further reduces the decoding latency from 701.9 ms to only 11.0 ms, almost eliminating the sequential decoding overhead.

The post-ViT latency excludes vision encoder cost. The key reason for the post-ViT reduction is that the proposed Zero-initialized Motion-aware Spatial Compressor (Z-MSC) aggressively compresses the spatial resolution before the visual tokens are processed by the LLM. On Charades-STA, the spatial feature map is compressed from approximately 12×18 to 4×4, corresponding to about

a 92.6% reduction in visual tokens. Consequently, the LLM processes substantially fewer visual tokens, which dramatically reduces both the computational complexity and memory consumption of subsequent transformer layers. Therefore, the post-ViT efficiency gain of SwiftVMR spans the LLM forward computation and decoding stages, rather than originating from a single component.

Vision-encoder latency is measured under each method's respective video sampling and spatial-resolution settings; see Appendix C for details.

Table 4: Accuracy-latency trade-off on Charades-STA with 7B models.

Method	E2E Latency (ms)	R1@7
Baseline (Avg. Pooling + Regressive Head)	**780.5**	52.10
SwiftVMR	793.2	**54.54**

Trade-off against a simple baseline. Table 4 compares SwiftVMR with a 7B baseline using average pooling and a regressive head. SwiftVMR improves R1@7 from 52.10 to 54.54 (+2.44 percentage points), while end-to-end latency increases only from 780.5 to 793.2 ms (+12.7 ms, or 1.6%). This shows that the accuracy gain comes with little additional inference cost.

C Video Sampling Settings

For SwiftVMR, we use dataset-specific video sampling settings: Charades-STA is sampled at 4 fps with a maximum of 128 frames; QVHighlights at 2 fps with at most 150 frames; TACoS at 2 fps with at most 256 frames; and ActivityNet Captions at 2 fps with at most 128 frames.

For UniTime, we retain its original sampling and adaptive frame-scaling protocol (Li et al., 2025b), rather than imposing SwiftVMR's frame limits. Its adaptive frame scaling adjusts spatial resolution or token allocation according to video length under a fixed token budget. The two methods are therefore not constrained to use identical visual-encoder inputs. Differences in frame sampling and adaptive spatial resolution can contribute to the vision-encoder latency difference in Table 3; this difference should not be attributed to Z-MSC, which operates after visual encoding. Post-ViT latency excludes the vision encoder, whereas end-to-end latency includes it. The reported peak GPU memory is measured over each method's complete inference pipeline, not only the post-ViT stages.

D Dataset Details

- **Charades-STA** focuses on daily indoor activities. It is built upon the Charades dataset and contains 12,408 moment annotations for training and 3,720 for testing.
- **ActivityNet Captions** consists of diverse, open-domain videos. Following the previous works (Zhang et al., 2020a), we utilize 37,421 query-moment pairs for training and 17,505 for testing.
- **QVHighlights** is a challenging benchmark designed for retrieving specific moments and highlights across varied genres (e.g., daily vlogs, news). It comprises approximately 7.2k training queries and 1.5k testing queries.
- **TACoS** centers on fine-grained, multi-step cooking activities recorded in a fixed-camera kitchen environment. It provides a densely annotated corpus with 9.8k training and 4.0k testing annotations.

E More analysis of the DSSD

Computational efficiency during the training phase. Fundamentally, DSSD allows the semantic features and continuous logit distributions generated by the high-capacity teacher to be precomputed and stored offline. As a result, the distillation process completely eliminates the need for the heavy teacher model to perform expensive forward passes during student optimization. To leverage these cached offline targets, the student model only requires a lightweight linear projection layer within its low-dimensional feature space to perform feature alignment. Consequently, incorporating

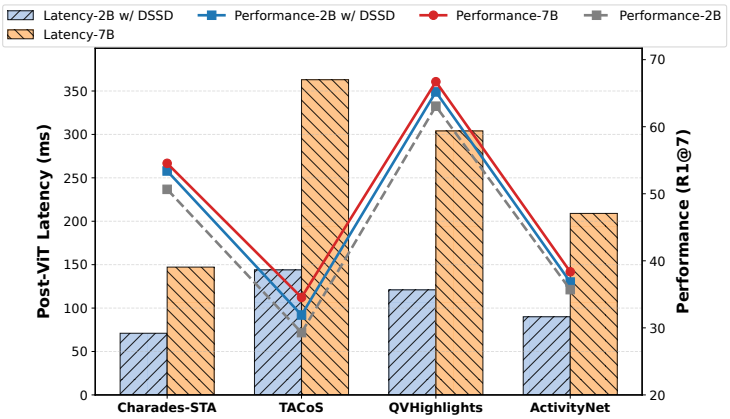


Figure 5: Efficiency-performance comparison of the DSSD strategy. The bar chart (left y-axis) illustrates the post-ViT latency comparison between the high-capacity 7B teacher model and the lightweight 2B student model. The line chart (right y-axis) denotes the corresponding performance.

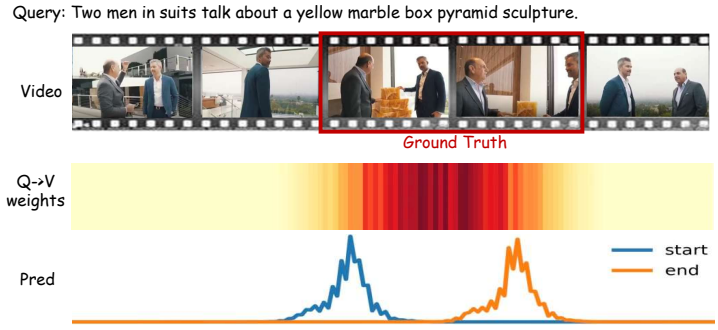


Figure 6: Qualitative visualization of the SwiftVMR localization process. From top to bottom: (1) Language query. (2) Video frames. (3) The Q→V cross-attention heatmap (4) The predicted boundary distributions. For visualization, the predicted center–duration distributions are mapped to start–end boundaries via $\tau_{s/e} = \tau_c \pm \tau_d/2$.

the DSSD strategy introduces virtually negligible parameter overhead and computational burden to the student model.

Furthermore, since this auxiliary projection layer is strictly discarded after training, DSSD incurs absolutely zero additional parameter or computational cost to the student model during the inference phase.

Efficiency-Performance Ratio of DSSD. As illustrated in Figure 5, we compare the post-ViT inference latency and localization accuracy between the massive 7B teacher model and the lightweight 2B student model. The transition to the compact 2B architecture drastically reduces the post-ViT inference latency by over 50%. Crucially, empowered by the knowledge transfer of DSSD, the lightweight student model retains the vast majority of the teacher's grounding capabilities, suffering only a marginal average performance degradation of approximately 5% compared to its heavy counterpart. This compelling efficiency-performance trade-off explicitly demonstrates that our framework successfully overcomes the massive parameter scale barrier, rendering the high-accuracy MLLM-based VMR pipeline highly practical and accessible for resource-constrained real-world deployment.

F Qualitative Analysis

Localization Process.

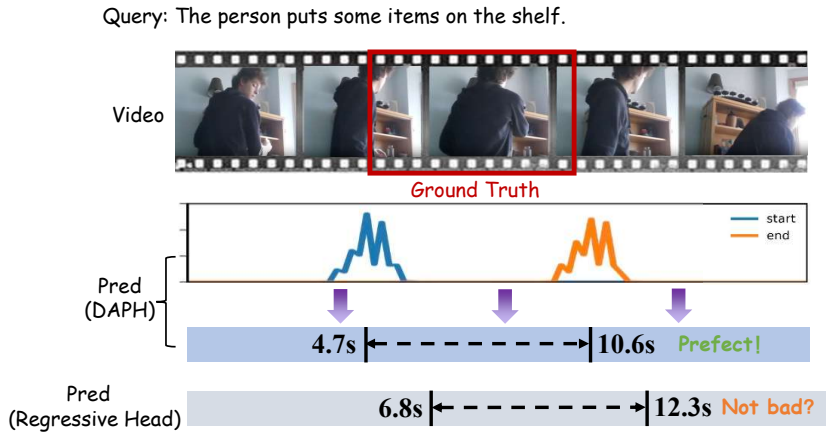


Figure 7: Qualitative comparison between the proposed DAPH and a standard regressive head under visually ambiguous conditions.

Table 5: Flexibility validation for different MLLMs on QVHighlights benchmark.

Method	R1@3	R1@5	R1@7	mIOU
SwiftVMR (Qwen2-VL-2B)	86.84	77.32	63.06	69.80
SwiftVMR (Qwen2-VL-7B)	89.74	79.74	66.71	72.03
SwiftVMR (Qwen2.5-VL-7B)	90.26	80.52	68.19	72.96
SwiftVMR (Qwen3-VL-8B)	90.39	81.16	68.34	73.13
SwiftVMR (Qwen3.5-4B)	90.77	82.39	69.10	74.03

To intuitively demonstrate the cross-modal grounding and boundary modeling capabilities of our SwiftVMR framework, Figure 6 visualizes the end-to-end prediction process on a representative sample. Given the language query and input video frames, we extract the cross-attention weights to illustrate how the learnable logits queries dynamically aggregate the multimodal representations $\hat{V}$. The resulting heatmap reveals a highly concentrated semantic alignment, indicating that the queries successfully suppress background noise and focus on the relevant action segments. The model then predicts boundary distributions over temporal bins and decodes continuous coordinates by expectation.

Comparison with a Regressive Head.

Figure 7 presents a qualitative comparison between DAPH and a standard regression head under visual ambiguity. For the query *"the person puts some items on the shelf,"* the actor's back faces the camera and obscures the exact contact frame. The regression head predicts 6.8–12.3 s, shifting both boundaries later than the annotated interval. DAPH predicts 4.7–10.6 s, which is closer to the annotation. This example illustrates a localization difference when the action is recognizable but its precise onset and completion are less directly visible.

The plotted curves show how DAPH represents the prediction through probabilities distributed over temporal bins. The final coordinates are obtained by expectation over these probabilities, so the localization output depends on the distribution rather than a single selected bin. The curves visualize this intermediate representation, while the predicted intervals show the resulting localization. This case complements the quantitative comparisons with the regression head and the w/o DFL variant in Tables 2 and 6, which evaluate the head parameterization and distribution supervision across the two benchmarks.

G Flexibility Verification

To evaluate the flexibility of SwiftVMR, we instantiate our framework on multiple vision-language foundation models with different architectures and parameter scales. As shown in Table 5, SwiftVMR consistently achieves competitive temporal localization performance across all evalu-

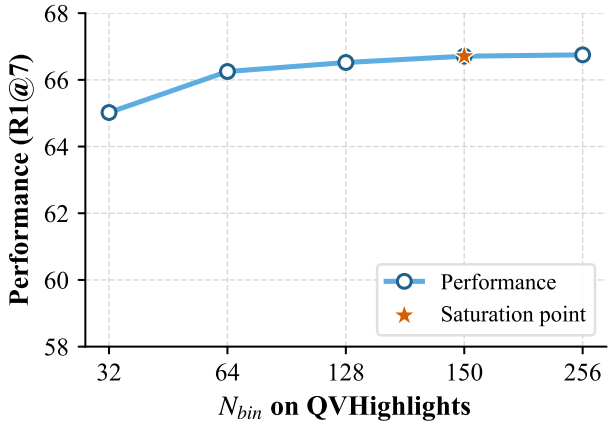
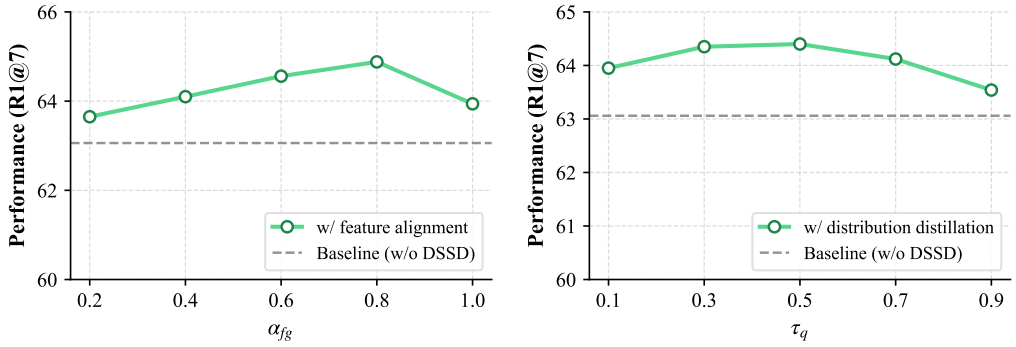


Figure 8: Ablation studies of N_{bin} in DAPH.



(a) Ablation studies of α_{fg} in feature alignment. (b) Ablation studies of τ_q in distribution distillation.

Figure 9: We conducted ablation studies on two critical hyperparameters in DSSD on QVHighlights benchmark

ated backbones. Qwen3-VL-8B outperforms both Qwen2-VL-7B and Qwen2.5-VL-7B on all four reported metrics. Qwen3.5-4B achieves the highest scores, including 69.10 R1@7 and 74.03 mIoU, despite its smaller parameter scale. These results support the flexibility of SwiftVMR across the evaluated MLLMs rather than dependence on a single backbone.

H Ablation Studies of Hyperparameters

Number of Bins in DAPH. To investigate the impact of temporal discretization granularity, we ablate the hyperparameter N_{bin} in DAPH. As shown in Figure 8, the overall performance variation across the entire range from 32 to 256 is remarkably small, demonstrating that N_{bin} is not a highly sensitive hyperparameter and our framework is intrinsically robust. Within this narrow variance, increasing N_{bin} from 32 to 150 yields a steady, relatively more noticeable improvement as the model acquires sufficient temporal resolution. Performance then plateaus at $N_{bin} = 150$. Crucially, this choice is empirically grounded: most video durations in the QVHighlights benchmark are approximately 150 seconds, meaning $N_{bin} = 150$ intuitively allocates roughly one bin per second. Further increasing N_{bin} to 256 yields virtually no performance change, suggesting that a 150-bin discretization already captures the finest necessary temporal details without introducing redundant dimensions. Given this insensitivity at sufficient granularities, we heuristically set N_{bin} to 128 for both Charades-STA and ActivityNet Captions, and 256 for TACoS, purely to maintain a natural alignment with their specific temporal distributions.

DSSD Hyperparameters. We perform ablation studies on its two critical hyperparameters in DSSD: the foreground balancing weight (α_{fg}) in the feature alignment space, and the gating thresh-

Table 6: Ablation studies of our proposed components on Charades-STA.

Line ID	Approach	R1@3	R1@5	R1@7	mIOU
Baseline (7B): Avg. Pooling + Regressive Head		82.15	71.99	52.10	61.38
(a) Z-MSC (7B)					
#1	w/o Spatiotemporal Residual Path	83.17	73.17	53.04	62.47
#2	w/o Content-Adaptive Path	82.72	72.45	52.88	62.10
#3	w/o Zero-Init	81.24	71.37	49.57	60.10
#4	Ours	**84.14**	**74.76**	**54.54**	**63.32**
(b) Prompt paradigm and DAPH (7B)					
#1	`[Video][Text]`	80.48	68.68	46.83	58.70
#2	Regressive Head	82.50	72.31	53.01	61.72
#3	w/o DFL	82.90	72.53	53.15	61.93
#4	Ours	**84.14**	**74.76**	**54.54**	**63.32**
(c) DSSD (2B)					
#1	w/o DSSD	81.51	71.77	50.67	60.74
#2	w/o logits	82.82	72.74	52.77	61.76
#3	w/o feature	82.34	71.72	51.67	60.98
#4	Coordinate Distillation Only	82.18	70.97	51.10	60.51
#5	Feature + Coordinate Distillation	82.93	72.82	52.59	61.51
#6	feature w/o weighted	82.04	71.91	52.31	61.33
#7	feature in LHS	81.61	72.15	51.67	60.99
#8	Ours	**83.52**	**73.25**	**53.39**	**62.05**

old (τ_q) in the distribution distillation space. As shown in Figure 9a, we evaluate the feature alignment hyperparameter α_{fg}, with the corresponding background weight β_{bg} complementarily set to $1 - \alpha_{fg}$. This formulation dictates the optimization focus between the foreground action and the background context. We observe that setting $\alpha_{fg} = 0.8$ yields the optimal retrieval effectiveness. A larger value ($\alpha_{fg} > 0.8$) overly concentrates the distillation on the foreground, causing the student model to ignore valuable contextual background knowledge that aids in holistic scene understanding. Conversely, a smaller value leads that the background information overshadows the primary action and acts as visual noise, which distracts the lightweight student model. As shown in Figure 9b, we ablate the quality-gated threshold τ_q. This gate controls the quality of the teacher's soft labels passed to the student. The performance peaks at $\tau_q = 0.5$. If the threshold is set too low, the gate fails to filter out unconfident predictions, inadvertently introducing the teacher's prediction noise into the student's optimization process. On the other hand, an excessively high threshold makes the gating mechanism overly strict, which severely bottlenecks the distillation pipeline and prevents the effective transfer of valuable "dark knowledge" from the teacher.

I Ablation Studies on Charades-STA

To assess robustness and generalization on Charades-STA, Table 6 reports the ablation studies of our proposed components with 7B foundation models, while DSSD ablations retain 2B configuration.

The full 7B model outperforms the average-pooling baseline with a regressive head on all four metrics. The ablations show similar trends to QVHighlights:

- **Design of Z-MSC:** Table 6(a) ablates the Spatiotemporal Residual Path for multi-scale motion and the Content-Adaptive Path for spatial aggregation. The w/o Zero-Init variant retains both paths but disables zero-initialization in both. The full model outperforms all three variants on all four metrics.
- **Early Fusion Paradigm and DAPH:** As shown in Table 6(b), our full model scores higher than the decoupled `[Video][Text]`, deterministic-regression, and w/o DFL variants on all four metrics. These comparisons favor the full configuration.
- **Efficacy of DSSD Components:** Table 6(c) supports the complementary roles of feature alignment and distribution distillation. Foreground weighting and low-dimensional alignment improve performance. With the same student DAPH and quality gate, distribution-

System:
You are a helpful assistant.
User:
Your task is to ground the temporal window of the query in the following video.
Query: <Query>
Video: <Video frames>

Figure 10: Prompt template for training.

only distillation outperforms the L1/GIoU coordinate-only baseline on all four metrics. When feature alignment is retained, full DSSD also outperforms Feature + Coordinate Distillation on all four metrics. Together with QVHighlights, this supports the additional value of distribution targets alongside feature alignment. Feature alignment contributes larger gains than distribution-only distillation, suggesting useful semantic guidance beyond the student's boundary supervision. Combining both spaces achieves the highest scores on all four metrics.

J PROMPT TEMPLATE FOR SWIFTVMR

As shown in Figure 10, to effectively guide the MLLM within the SwiftVMR framework to perform precise temporal grounding, we design a structured and succinct prompt template. By coupling this query-aware prompt structure with our highly compressed visual tokens, SwiftVMR achieves robust cross-modal alignment with minimal input overhead.

K CODE AVAILABILITY

Our complete codebase, including the implementations of Z-MSC and DAPH, DSSD, along with the full prompt templates and training, evaluation scripts, will be open-sourced immediately upon the acceptance of this manuscript.