WHEN TO STOP: TRAINING-FREE VIDEO TEMPORAL GROUNDING VIA ENDPOINT ACTIVATION STEERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video temporal grounding (VTG) aims to localize the start and end times of queried events in untrimmed videos. Despite substantial progress in event understanding, video large language models still produce unreliable timestamps. Existing improvements often rely on post-training with temporal annotations or additional inference, incurring annotation or computational costs. Contrastive activation steering offers a complementary training-free approach, but applying it to timestamp generation requires identifying the temporal components within activation differences that influence numerical boundary decisions. To address this question, we investigate how video evidence contributes to timestamp generation. We find that temporal information persists during generation, yet visual evidence provides weaker discrimination among end-time candidates than among start-time candidates. Models can recognize an event and locate its onset while still struggling to determine when it terminates. To identify the temporal evidence influencing this decision, we use endpoint-score Jacobians to locate directions that alter end-time preferences and project activation differences induced by temporal perturbations onto these directions. This construction selects components according to native endpoint sensitivity rather than activation differences alone. Aggregating the projected components across samples yields a low-dimensional subspace in late decoder layers. Causal interventions further establish that this subspace mediates the influence of temporal evidence on end-time prediction. Building on this intervention interface, we propose training-free endpoint activation steering. Using unlabeled samples, we extract a shared direction within the subspace and calibrate its relative strength offline for each model–dataset pair. At inference, we apply a single residual update immediately before the first end-time numeric token, preserving the generated start and native decoding without parameter updates or additional video forward passes. Evaluated across seven frozen backbones on TimeLens-Bench, our method improves mIoU by up to 3.62 points. Across the 4B and 8B variants of Qwen3-VL, InternVL3.5, and TimeLens2, it consistently improves both mIoU and R@0.5 on all three datasets, with average gains of 2.19 and 1.46 points respectively.

1 INTRODUCTION

Video Temporal Grounding (VTG) aims to localize the start and end times of a queried event within an untrimmed video. Recent Video Language Models (VLMs) approach this task through a unified generative interface, translating video content and language queries into timestamp sequences (Ren et al., 2024; Huang et al., 2024). However, recognizing an event does not ensure that its boundaries are expressed accurately. Boundary-related visual responses can remain misaligned with predicted timestamps (Jia et al., 2026). Reliable grounding therefore requires not only representing temporal information, but also using it effectively when generating numerical boundaries.

We investigate what temporal information frozen video VLMs retain during generation and how it contributes to timestamp prediction. Hidden-state probes in Qwen3-VL-4B, TimeLens2-4B, and InternVL3.5-4B recover temporal information both before answer generation and immediately be-

Anonymous code: `https://anonymous.4open.science/r/When-to-Stop-Training-Free-Video-Temporal-Grounding-via-Endpoint-Activation-Steering-4A14/`

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.

fore end-time prediction. We next examine how this retained information supports boundary selection by comparing timestamp preferences with and without video. Video evidence distinguishes end-time candidates less reliably than start-time candidates across all three backbones, with the strongest asymmetry in Qwen (Figures 1 and 2). This pattern helps explain why a model can locate an event's onset yet still misjudge when it ends.

We therefore seek the activation components through which temporal evidence influences end-time prediction, constructing a Jacobian-guided low-rank projection framework to identify them. Endpoint-score Jacobians measure how small activation changes alter the model's preference between earlier and later endings. We obtain temporal contrasts by comparing activations from the original video and a temporally reordered version that preserves its visual content. Each contrast is projected onto its own query's endpoint-sensitive directions before the resulting components are aggregated into a low-dimensional subspace in late decoder layers. This extends contrastive steering approaches such as CAA and VTI (Rimsky et al., 2024; Liu et al., 2025) by selecting temporal components according to the corresponding numerical endpoint decision. Removing the projected evidence component shifts endpoint preferences toward those under temporal perturbation, while reinjecting it into perturbed states partially restores localization, establishing the subspace as a causal intervention interface. Building on this interface, we develop training-free endpoint activation steering: unlabeled calibration yields a shared direction and relative strength for each model–dataset pair, followed by one residual update immediately before the first end-time numeric token. Subspace construction and calibration remain offline, without temporal annotations or parameter updates. Inference preserves the generated start and native decoding without additional video forward passes or generation calls.

Across seven frozen backbones on Charades-TimeLens, QVHighlights-TimeLens, and ActivityNet-TimeLens, our method achieves mIoU gains of up to 3.62 points over native decoding (Table 2). Both mIoU and R@0.5 improve on all three datasets for the 4B and 8B variants of Qwen3-VL, InternVL3.5, and TimeLens2, with average gains of 2.19 and 1.46 points, respectively. Gains extend to temporally post-trained models: TimeLens2-8B reaches 71.59 mIoU and 77.19 R@0.5 on QVHighlights-TimeLens, demonstrating complementarity with temporal post-training. Filtering temporal activation changes through each query's own endpoint-sensitive directions improves localization over low-rank compression alone (Table 4).

Our contributions are threefold:

- We identify a key bottleneck in video temporal grounding, where models can recognize an event and localize its onset yet struggle to determine its termination. Temporal information persists during generation, while visual evidence supports event termination less reliably than event onset.
- We construct a low-dimensional late-decoder subspace by filtering each query's temporal contrast through its own endpoint-sensitive directions before global aggregation, and establish its causal role in end-time prediction via causal intervention experiments.
- We propose a training-free activation steering method that uses this causal subspace to improve timestamp generation. It extracts a shared direction from projected temporal contrasts, calibrates relative strength on unlabeled samples, and applies one residual update before end-time generation, without temporal annotations, parameter updates, or additional video forward passes.

2 Related Work

Video temporal grounding. TimeChat and VTimeLLM support temporal grounding through timestamp-aware encoding and boundary-aware training, respectively (Ren et al., 2024; Huang et al., 2024). Time-R1 and TimeLens improve localization through reinforcement-learning-based post-training (Wang et al., 2025c; Zhang et al., 2026). CAVE aligns boundary-specific visual evidence with timestamps during training (Jia et al., 2026). Training-free alternatives change how temporal evidence is presented or queried. NumPro adds numerical frame identifiers (Wu et al., 2025), and FV-Action searches clips through binary visual questions (Huang et al., 2026). Du et al. (2026) use prefill attention to guide re-inference. In contrast, we target the native end-time readout of frozen

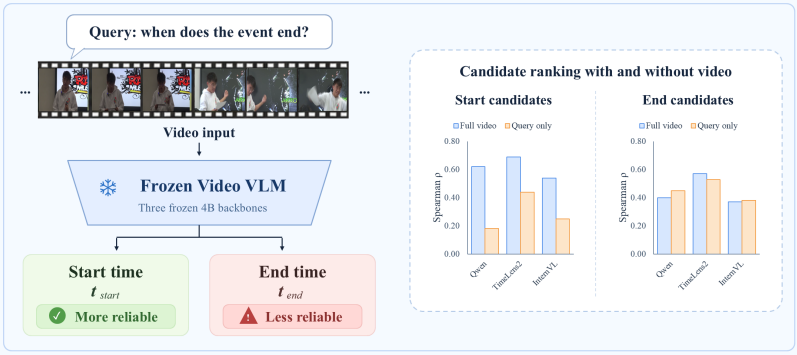


Figure 1: **Uneven visual support for temporal boundaries.** Native timestamp generation (left) and start/end candidate-ranking correlations with and without video in Qwen3-VL, TimeLens2, and InternVL3.5 (right, all 4B).

models. We retain the original video input and timestamp-generation interface, without external temporal search or additional generation calls.

Activation steering and evidence interventions. CAA constructs shared residual-stream directions from contrastive examples (Rimsky et al., 2024), while ITI shifts activations in selected attention heads to elicit truthful responses (Li et al., 2023). VTI extends steering to visual and textual representations for hallucination mitigation (Liu et al., 2025). STEAR combines layer-aware evidence reinjection with temporally perturbed counterfactual decoding (Fan et al., 2026). We filter each temporal contrast through its own query's endpoint sensitivity before aggregation, enabling a shared activation update before end-time generation.

3 Diagnosing Endpoint Readout in Frozen Video VLMs

We study frozen Qwen3-VL-4B-Instruct, TimeLens2-4B, and InternVL3.5-4B on Charades-TimeLens to trace how temporal representations contribute to endpoint predictions. We denote the positions immediately before the first answer token and the first end-time numeric token by P_0 and P_2. Probes are fitted and selected on video-disjoint training and validation sets, then evaluated on the test set. Coordinate errors are normalized by video duration T.

3.1 Temporal Information Persists During Generation

Position-specific probes retain their end-time accuracy from P_0 to P_2 in all three models, whereas transferring an unchanged P_0 probe increases error (Table 1). These probes use a canonical model-own-start prefix at P_2. Subspace analysis and steering use the exact native prefix. End information remains decodable, while its linear readout changes across generation positions. We next fix the query, video duration, generated start, and P_2 prefix across video conditions. In Qwen and TimeLens2, each probe receives identical prior features, isolating the contribution of the hidden-state input. Normal video lowers end MAE relative to both no video and temporal roll in all three models (Table 1). Thus, P_2 states retain video evidence beyond the generated start, with a contribution sensitive to temporal arrangement. Appendix A.1 provides the prior and remaining-time controls.

3.2 Video Evidence Provides Uneven Support for Temporal Boundaries

Around each annotated interval $[s^*, e^*]$, we move one boundary while fixing the other. We teacher-force each candidate and measure the mean negative log-likelihood $C(y)$ of its numeric timestamp tokens. Within each model, we compare start and end on queries eligible for both boundaries. Query-level Spearman correlations between cost and temporal distance measure preference for closer candidates: higher values indicate better discrimination (Figure 2a). Appendix A.2 gives candidate details and offset curves.

Comparing full-video and query-only rankings, video increases start correlation by +0.44, +0.25, and +0.29 in Qwen, TimeLens2, and InternVL, respectively, while end correlation changes by

Table 1: **End-time information and video contribution.** Position-specific decoding (top) and video controls under a fixed P_2 prefix (bottom).

Probe / comparison	Qwen3-VL	TimeLens2	InternVL3.5
Position-specific readout: normalized end MAE (↓)			
P_0 position-specific	0.12	0.11	0.16
P_2 position-specific	**0.11**	**0.10**	**0.15**
$P_0 \to P_2$ unchanged transfer	0.33	0.16	0.27
Fixed-prefix controls: paired MAE reduction (↑)			
No-video − Full	+0.03	+0.03	+0.12
Roll − Full	+0.02	+0.02	+0.06

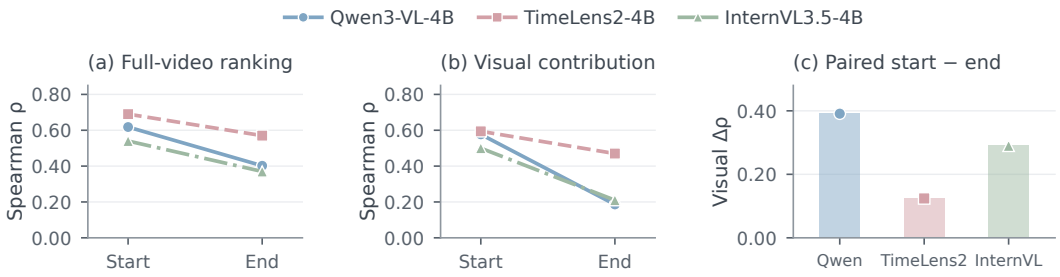


Figure 2: **Boundary discrimination on common queries.** (a) Full-video ranking. (b) Matched visual contribution. (c) Paired start-minus-end gap in visual contribution.

-0.05, +0.04, and -0.01 (Figure 1). Video thus improves onset discrimination substantially more than termination discrimination across all three models.

We next measure the candidate-level visual contribution by subtracting the query-only cost from the full-video cost:

$$C_{\text{visual}}(y) = C_{\text{full}}(y) - C_{\text{query}}(y). \tag{1}$$

Negative values indicate that video lowers a candidate's numeric-token cost. With integer offsets and matched numeric-token counts, video supports start discrimination more strongly in all three models (Figure 2b). We compute $\Delta\rho_i = \rho_{i,\text{start}} - \rho_{i,\text{end}}$ within each query before averaging. The paired visual-support gaps are +0.39, +0.12, and +0.29 for Qwen, TimeLens2, and InternVL, respectively (Figure 2c). For the same queried event, incremental video evidence therefore distinguishes its end less reliably than its start, motivating endpoint readout as the intervention target.

3.3 Endpoint Subspace Construction and Causal Validation

Endpoint subspace construction. To isolate temporal arrangement while holding visual content and token layout fixed, we now compare the original video with a temporally rolled version. For each calibration query, we generate with the original video and retain the model's own predicted start, end, and exact token prefix π_i up to P_2. Both branches reuse this prefix throughout state extraction, scoring, and intervention. Temporal roll reorders complete temporal tubelet groups while preserving input IDs, position inputs, video layout, and within-group spatial content. The paired states therefore differ in temporal arrangement under the same language context.

The construction in Figure 3 (left) uses candidates around the model's own predicted end e_i, rather than the annotated boundary used above. We form earlier/later pairs $(e_{i,r}^-, e_{i,r}^+)$ at three temporal radii, matching numeric-token counts within each pair. Let $h_{i,\text{full}}^{(\ell)}$ be the unmodified original-video residual at P_2, and let $Q_i(e)$ sum the numeric end-token log-probabilities under the original video and exact native prefix. We capture the model's native endpoint sensitivity through the score margin and its exact downstream Jacobian:

$$m_{i,r} = Q_i(e_{i,r}^+) - Q_i(e_{i,r}^-), \qquad g_{i,r}^{(\ell)} = \nabla_{h_{i,\text{full}}^{(\ell)}} m_{i,r}. \tag{2}$$

Stacking the individually normalized gradients gives a local sensitivity matrix. Its right singular vectors retained by the truncation rule in Appendix B.2 form the sample-local orthonormal basis $U_i^{(\ell)}$.

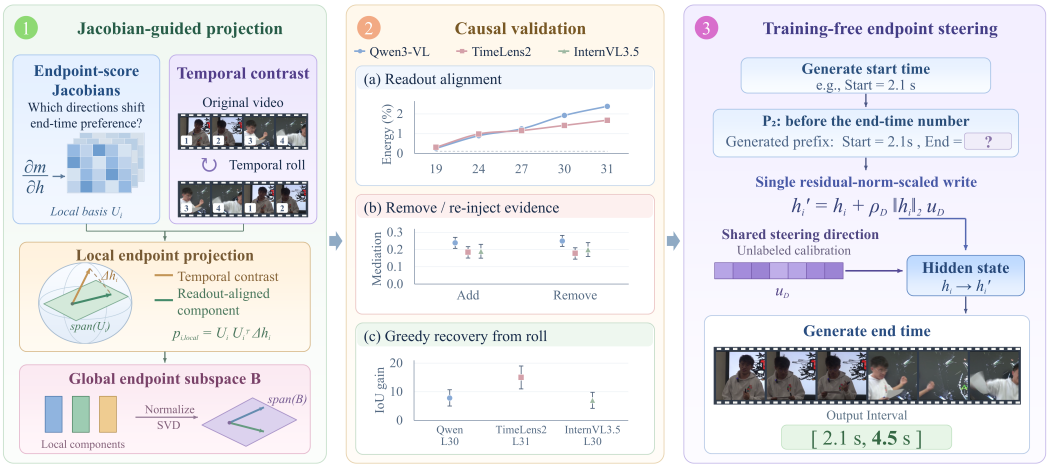


Figure 3: **Endpoint subspace construction and steering.** (1) Jacobian-guided projection and global basis. (2) Alignment in Qwen and TimeLens2 (a), mediation (b) and greedy recovery (c) in three models at rank 16: Qwen/InternVL L30, TimeLens2 L31. (3) One calibrated update. Timestamps are illustrative.

We then extract the temporal contrast and its component along these endpoint-sensitive directions:

$$\Delta h_i^{(\ell)} = h_{i,\mathrm{full}}^{(\ell)} - h_{i,\mathrm{roll}}^{(\ell)}, \qquad p_{i,\mathrm{local}}^{(\ell)} = U_i^{(\ell)} U_i^{(\ell)\top} \Delta h_i^{(\ell)}. \tag{3}$$

We aggregate these local components over unlabeled calibration queries to form a global endpoint subspace, independently for each model and layer. Each component is normalized before aggregation so that large activation magnitudes do not dominate the fit:

$$C_{i,:}^{(\ell)} = \frac{p_{i,\mathrm{local}}^{(\ell)\top}}{\|p_{i,\mathrm{local}}^{(\ell)}\|_2}, \qquad C^{(\ell)} = A^{(\ell)} \Sigma^{(\ell)} V^{(\ell)\top}, \qquad B^{(\ell)} = V_{:,1:R}^{(\ell)}. \tag{4}$$

The basis captures shared directions of readout-aligned temporal contrasts, with $B^{(\ell)\top} B^{(\ell)} = I$. Filtering before aggregation preserves the association between each query's temporal evidence and its endpoint sensitivity. Section 5.4 tests this association against temporal-only, gradient-only, and shuffled-pairing bases. Section 4 constructs a signed shared direction within the resulting subspace for deployment.

Unlabeled validation selects rank using split-half stability, causal-effect retention relative to rank 32, and separation from matched controls. Rank 16 is the smallest passing rank in Qwen and TimeLens2, retaining 90.30% and 96.30% of the rank-32 causal effect at L30. Figure 3, panels 2b,c, uses Qwen and InternVL L30 and TimeLens2 L31, which gives stronger diagnostic recovery than TimeLens2 L30. Appendices B.3 and B.4 give the selection evidence and site comparisons.

Causal validation. The fraction $\|p_{i,\mathrm{local}}^{(\ell)}\|_2^2 / \|\Delta h_i^{(\ell)}\|_2^2$ measures how much of the temporal contrast aligns with local endpoint sensitivity. This fraction is higher in later residual layers (Figure 3, panel 2a), reaching 2.39% in Qwen and 1.68% in TimeLens2 at L31, versus approximately 0.12% for a random projection. The aligned component is small in absolute energy, but its preferential alignment motivates a direct test of its effect on endpoint readout.

To test the reusable basis, we project each evaluation query's temporal contrast using the frozen global basis:

$$p_{i,\mathrm{global}}^{(\ell)} = B^{(\ell)} B^{(\ell)\top} \Delta h_i^{(\ell)}. \tag{5}$$

We remove this component from the original-video state or restore it to the temporally rolled state:

$$h_{i,\mathrm{full}}^{(\ell)\prime} = h_{i,\mathrm{full}}^{(\ell)} - \beta p_{i,\mathrm{global}}^{(\ell)}, \qquad h_{i,\mathrm{roll}}^{(\ell)\prime} = h_{i,\mathrm{roll}}^{(\ell)} + \beta p_{i,\mathrm{global}}^{(\ell)}. \tag{6}$$

Score mediation measures the fractional reduction in distance to the centered target candidate-score curve: the original-video curve for addition and the rolled curve for removal (Appendix B.4). At

$\beta = 1$, addition/removal mediation is 0.24/0.25 for Qwen L30, 0.19/0.18 for TimeLens2 L31, and 0.19/0.20 for InternVL L30 (Figure 3, panel 2b), with all six 95% intervals above zero. Both interventions move endpoint preferences toward their targets in all three models, exceeding random, orthogonal, and norm-matched raw-evidence controls.

All behavioral branches continue greedy decoding from the same retained P_2 prefix, fixing the start to the original-video prediction. Adding $p^{(\ell)}_{i,\text{global}}$ to the rolled state yields paired mean IoU gains of 7.80, 15.00, and 7.00 for Qwen L30, TimeLens2 L31, and InternVL L30, respectively (Figure 3, panel 2c). Recovery of candidate-score geometry is accompanied by improved greedy localization, identifying a causally effective subspace for endpoint control across the three models. Section 4 combines calibration contrasts into a shared direction for original-video states, avoiding per-query temporal contrasts at deployment. Section 5.2 evaluates the resulting gains.

4 Shared Endpoint Activation Steering

The construction in Section 3.3 provides each frozen model M with an orthonormal endpoint basis $B_M \in \mathbb{R}^{H_M \times R}$, where H_M is the residual dimension and R is the subspace rank. Its projector is $P_M = B_M B_M^\top$. The basis is fixed and reused across datasets.

For dataset D, let $\mathcal{C}_D$ be an unlabeled calibration set disjoint from the test set. Following the paired-state protocol in Section 3.3, we extract states from the original and temporally rolled videos under the same model-generated P_2 prefix. Their temporal contrast is $\Delta h_j = h_{j,\text{full}} - h_{j,\text{roll}}$, and its endpoint projection is $d_j = P\Delta h_j$. We combine the shared temporal components by averaging these projected differences and normalizing their mean:

$$\bar{d}_D = \frac{1}{|\mathcal{C}_D|} \sum_{j \in \mathcal{C}_D} d_j, \qquad u_D = \frac{\bar{d}_D}{\|\bar{d}_D\|_2}. \tag{7}$$

Basis construction normalizes each local component before aggregation (Eq. (4)). Shared-direction construction retains each d_j's original magnitude and normalizes only the final mean. The resulting u_D is fixed for each model–dataset pair.

We calibrate the update relative to the unmodified original-video residual using the same unlabeled samples:

$$\rho_D = \text{median}_{j \in \mathcal{C}_D} \left(\frac{\lambda \|d_j\|_2}{\|h_{j,\text{full}}\|_2} \right). \tag{8}$$

Here $\lambda > 0$ is a global strength coefficient fixed across datasets (Appendix D.2). The median captures a typical relative magnitude for the model–dataset pair, without temporal annotations.

At inference, the model follows native greedy generation with the original video to P_2 (Figure 3, right). At the selected decoder layer, we read the unmodified residual h_i and apply

$$h'_i = h_i + \rho_D \|h_i\|_2 u_D. \tag{9}$$

Since $\|u_D\|_2 = 1$, the update has fixed relative magnitude $\|h'_i - h_i\|_2 / \|h_i\|_2 = \rho_D$, while its absolute magnitude scales with the current residual norm. Greedy decoding continues without further writes to produce (s_i, e'_i), preserving the generated start s_i and all preceding tokens.

Basis construction and calibration are performed offline, with model weights frozen. Online steering uses the fixed u_D and ρ_D within one native generation call, adding only a residual-norm computation and a vector addition. Appendix C gives the pseudocode.

5 Experiments

5.1 Experimental Setup

We evaluate on Charades-TimeLens, QVHighlights-TimeLens, and ActivityNet-TimeLens from TimeLens-Bench (Zhang et al., 2026). Data splits and calibration protocols are provided in Appendix D.1. We report mIoU and R@0.5. Inference cost is measured by average online latency and generation-call count (Appendix F). We evaluate the 4B and 8B variants of Qwen3-VL-Instruct (Bai

Table 2: **Main results on TimeLens-Bench.** ΔmIoU is relative to the corresponding backbone's native decoding. Shaded rows denote ours. Bold marks column maxima.

	Charades-TimeLens			QVHighlights-TimeLens			ActivityNet-TimeLens		
Model	R@0.5	mIoU	ΔmIoU	R@0.5	mIoU	ΔmIoU	R@0.5	mIoU	ΔmIoU
LLaVA-NeXT-Video-7B	0.00	1.33	–	0.00	0.34	–	0.15	1.13	–
LLaVA-NeXT-Video-7B (ours)	0.00	1.83	+0.50	0.00	0.84	+0.50	0.15	1.13	+0.00
Qwen3-VL-4B	46.54	41.41	–	58.77	54.16	–	43.80	39.81	–
Qwen3-VL-4B (ours)	49.48	43.93	+2.52	60.96	56.53	+2.37	45.48	42.73	+2.92
Qwen3-VL-8B	34.59	34.48	–	47.81	47.24	–	41.04	38.43	–
Qwen3-VL-8B (ours)	35.64	36.52	+2.04	48.68	49.19	+1.95	42.27	39.92	+1.49
InternVL3.5-4B	26.83	31.08	–	39.04	38.97	–	25.42	27.06	–
InternVL3.5-4B (ours)	28.93	34.70	**+3.62**	41.23	42.36	**+3.39**	25.73	30.29	**+3.23**
InternVL3.5-8B	19.92	24.29	–	25.00	26.65	–	13.17	14.14	–
InternVL3.5-8B (ours)	22.43	26.45	+2.16	26.32	29.55	+2.90	15.47	16.43	+2.29
TimeLens2-4B	66.46	57.84	–	75.00	69.22	–	65.85	59.24	–
TimeLens2-4B (ours)	67.09	59.69	+1.85	75.88	71.20	+1.98	66.62	**60.22**	+0.98
TimeLens2-8B	68.13	58.13	–	75.88	70.51	–	65.85	58.32	–
TimeLens2-8B (ours)	**69.18**	**59.96**	+1.83	**77.19**	**71.59**	+1.08	**66.77**	59.16	+0.84

et al., 2025), InternVL3.5 (Wang et al., 2025b), and TimeLens2 (Zhu et al., 2026), together with LLaVA-NeXT-Video-7B (Zhang et al., 2024). All model weights remain frozen. Additional experimental settings are given in Appendix D.

5.2 Main Results

Table 2 compares shared endpoint steering with native decoding across seven frozen backbones. Steering improves mIoU in 20 of the 21 model–dataset settings. LLaVA-NeXT-Video-7B on ActivityNet-TimeLens is unchanged. Across the six Qwen3-VL, InternVL3.5, and TimeLens2 backbones, both mIoU and R@0.5 improve in all 18 settings, with average mIoU gains of +2.34, +2.28, and +1.96 on Charades-TimeLens, QVHighlights-TimeLens, and ActivityNet-TimeLens, respectively. InternVL3.5-4B gains +3.23 to +3.62, while the temporally specialized TimeLens2 models also gain +0.84 to +1.98. These results show that a single endpoint update benefits both general-purpose and temporally specialized VLMs across architectures and model sizes. Appendix E.5 reports results for additional video VLMs.

5.3 Component Ablations

Table 3 reports absolute mIoU and R@0.5 on Qwen3-VL-4B. The complete method performs best on both metrics across all three datasets. Removing either endpoint projection or temporal contrast lowers mIoU by 2.01–2.43, showing that both endpoint sensitivity and temporal evidence contribute to the steering direction. A fixed update magnitude loses 2.09–2.20 mIoU, supporting scaling each update to the current residual norm. The shared direction also outperforms sample-specific directions by 0.15–0.92 mIoU while avoiding online temporal contrasts. Sharing the direction therefore supports both localization quality and efficient deployment.

Direction and norm controls test whether the gains depend on the chosen direction or can be reproduced by changing activation magnitude alone. Reversed and random subspace directions fall below the complete method on both metrics. Preserving only the norm change or only the direction change also reduces performance. These controls support the complete signed update and its joint effect on residual direction and norm.

Write-position controls favor one update before the first end-time token over delayed or repeated writes. Even when the total update budget is matched, repeated writes trail the complete method by 1.59–1.84 mIoU. The advantage therefore persists after controlling the total dose, supporting intervention at the initial numerical decision. Section 5.4 further isolates endpoint sensitivity from generic low-rank filtering. Appendix D.4 defines the controls.

Table 3: **Component ablations on Qwen3-VL-4B.** Absolute mIoU and R@0.5. Native decoding and the complete method are included for reference.

	Charades-TimeLens		QVHighlights-TimeLens		ActivityNet-TimeLens	
Variant	mIoU	R@0.5	mIoU	R@0.5	mIoU	R@0.5
Qwen3-VL-4B (native)	41.41	46.54	54.16	58.77	39.81	43.80
Direction construction and scaling						
W/O endpoint projection	41.50	46.75	54.32	58.33	40.72	43.64
W/O temporal contrast	41.68	46.96	54.12	58.33	40.65	43.19
Fixed update magnitude	41.73	47.59	54.37	59.21	40.64	43.19
Sample-specific direction	43.78	49.06	55.61	59.65	42.44	45.02
Direction and norm controls						
Reversed direction	41.45	47.17	54.25	58.33	40.27	43.19
Random subspace direction	42.06	47.80	54.80	59.65	40.83	43.80
Norm-only update	42.20	47.59	54.78	58.77	40.94	43.49
Direction-only update	41.81	47.17	54.48	59.21	40.87	43.80
Write position and repetition						
Write at the final end token	42.14	47.38	54.69	58.77	41.17	44.10
Write at every end token	42.37	48.22	54.96	59.21	41.06	43.64
Repeated writes, matched total budget	42.34	47.59	54.91	59.65	40.89	43.80
Qwen3-VL-4B (ours)	**43.93**	**49.48**	**56.53**	**60.96**	**42.73**	**45.48**

5.4 Comparison with Training-Free Alternatives

On frozen Qwen3-VL-4B, Table 4 isolates basis construction using identical unlabeled data, L30, rank 16, P_2 writes, and update norms. Temporal-only SVD trails our method by 2.12–2.32 mIoU. Pooling endpoint-gradient axes narrows the gap to 0.67–0.72, showing the value of endpoint sensitivity. Shuffling query–Jacobian pairs loses 1.02–1.07, supporting filtering each contrast through its own sensitivity before aggregation. All nine paired comparisons favor our method.

FV-Action (Huang et al., 2026) scores local clips through binary questions and places a fixed-width interval around the selected peak. Our method exceeds this external search by 11.28–27.68 mIoU while retaining native timestamp generation. Appendix D.3 summarizes the protocols.

Table 4: **Training-free comparisons on Qwen3-VL-4B.**

	Charades-TimeLens		QVHighlights-TimeLens		ActivityNet-TimeLens	
Model / method	R@0.5	mIoU	R@0.5	mIoU	R@0.5	mIoU
Qwen3-VL-4B	46.54	41.41	58.77	54.16	43.80	39.81
+ FV-Action	27.46	32.65	21.49	28.85	12.40	19.29
Alternative subspace constructions						
Temporal-only basis	46.96	41.76	59.21	54.41	44.10	40.41
Endpoint-gradient basis	48.64	43.26	60.09	55.86	45.02	42.01
Shuffled pairing	48.22	42.91	59.65	55.51	44.72	41.66
Qwen3-VL-4B (ours)	**49.48**	**43.93**	**60.96**	**56.53**	**45.48**	**42.73**

5.5 Sensitivity and Calibration

Figure 4 varies one factor at a time on Qwen3-VL-4B. Scaling ρ_D by γ gives peak gains at the default $\gamma = 1$ across all three datasets. Every tested nonzero strength improves mIoU. Charades-TimeLens sweeps peak at rank 16 and L30, agreeing with the unlabeled selection in Section 3.3. Only 32 unlabeled calibration samples yield 1.05–1.38 mIoU gains. Gains increase with the tested

budget, reaching 2.37–2.92 with 256 samples, so additional calibration strengthens an already useful direction. Appendix D.4 gives the sweep settings.

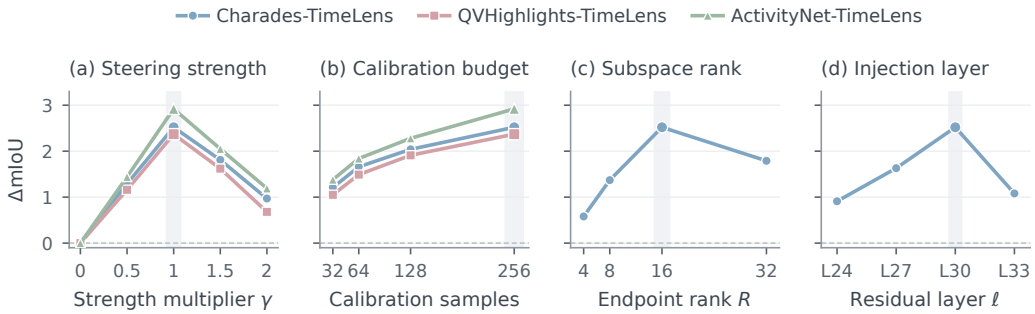


Figure 4: **Sensitivity and calibration on Qwen3-VL-4B.** (a) Steering strength and (b) calibration budget across three datasets. (c) Subspace rank and (d) injection layer on Charades-TimeLens. Shaded bands mark the defaults.

5.6 Endpoint-Only Evaluation

The main evaluation in Table 2 uses the model's own generated start. As a complementary assessment, we isolate end-time quality by re-evaluating the predicted intervals with the start fixed to the annotated boundary. Across three 8B backbones and three datasets (Table 5), steering improves both metrics in all nine settings, with mIoU gains of 1.02–1.75 and R@0.5 gains of 1.26–1.75. Protocol details and analysis by baseline IoU are provided in Appendix E.1.

Table 5: **Endpoint-only evaluation with a ground-truth start.** ΔmIoU is relative to native decoding. Shaded rows denote ours. Bold marks column maxima.

	Charades-TimeLens			QVHighlights-TimeLens			ActivityNet-TimeLens		
Model	R@0.5	mIoU	ΔmIoU	R@0.5	mIoU	ΔmIoU	R@0.5	mIoU	ΔmIoU
Qwen3-VL-8B	35.22	36.10	–	48.25	48.67	–	41.81	39.57	–
Qwen3-VL-8B (ours)	36.69	37.26	+1.16	50.00	50.19	**+1.52**	43.34	41.10	+1.53
InternVL3.5-8B	22.22	26.02	–	25.88	29.25	–	15.01	15.91	–
InternVL3.5-8B (ours)	23.69	27.30	**+1.28**	27.19	30.27	+1.02	16.54	17.61	+1.70
TimeLens2-8B	68.76	59.60	–	76.75	71.20	–	66.16	58.64	–
TimeLens2-8B (ours)	**70.02**	**60.69**	+1.09	**78.07**	**72.56**	+1.36	**67.53**	**60.39**	**+1.75**

6 Conclusion

Temporal information persists during generation, yet video supports termination less reliably than onset. Jacobian-guided projection identifies a low-dimensional endpoint subspace with a causal role in prediction. Direct controls support endpoint-sensitive filtering and correct query–Jacobian pairing beyond generic low-rank filtering. Our training-free method combines unlabeled calibration with one residual-norm-scaled update, preserving the generated start and native decoding. It improves localization across seven frozen backbones and three benchmarks, including temporally post-trained models, without parameter updates or additional video forward passes at inference. Extending this intervention to jointly refine start and end times is a natural direction for future work.

AI USE STATEMENT

We used generative AI tools to polish the writing and improve clarity of expression. The authors reviewed the revisions and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Sections 3.3 and 4 define subspace construction and shared steering. Appendices A–D provide the diagnostic protocols, algorithm, calibration procedure, and experimental settings. The implementation is available through the anonymous repository linked on the first page.

REFERENCES

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.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Dazhao Du, Liao Duan, Jian Liu, Tao Han, Yujia Zhang, Eric Liu, Xi Chen, and Song Guo. MLLMs know when before speaking: Revealing and recovering temporal grounding via attention cues. *arXiv preprint arXiv:2605.21954*, 2026.

Linfeng Fan, Yuan Tian, Ziwei Li, and Zhiwu Lu. STEAR: Layer-aware spatiotemporal evidence intervention for hallucination mitigation in video large language models. *arXiv preprint arXiv:2604.03045*, 2026.

Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. Trace: Temporal grounding video llm via causal event modeling. In *International Conference on Learning Representations*, volume 2025, pp. 89594–89618, 2025.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. VTimeLLM: Empower LLM to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14271–14280, 2024.

Ji Huang, Barry Devereux, and Hui Wang. Your VLM already knows when: Training-free temporal grounding by asking yes or no. *arXiv preprint arXiv:2608.08315*, 2026.

Wei Jia, Zhicong Lu, Yu Chen, Xiang Wang, Shuai Li, Wenqian Lv, Jiayue Cao, and Huaxing Liu. CAVE: Competence-aware visual boundary evidence alignment for video temporal grounding. *arXiv preprint arXiv:2608.02078*, 2026.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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*, 2025.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, volume 2026, pp. 109089–109117, 2026a.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, et al. Videochat3: Fully open video mllm for efficient and generalist video understanding. *arXiv preprint arXiv:2607.14935*, 2026b.

Sheng Liu, Haotian Ye, and James Y Zou. Reducing hallucinations in large vision-language models via latent space steering. In *International Conference on Learning Representations*, 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.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-VideoLLM: Sharpening fine-grained temporal grounding in video large language models. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 959–975, 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, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-R1: Post-training large vision language model for temporal video grounding. In *Advances in Neural Information Processing Systems*, volume 38, 2025c.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13754–13765, 2025.

Ziang Yan, Sheng Xia, Jiashuo Yu, Yue Wu, Tianxiang Jiang, Songze Li, Kanghui Tian, Yicheng Xu, Yinan He, Kai Chen, et al. Internvideo3: Agentify foundation models with multimodal contextual reasoning. *arXiv preprint arXiv:2606.12195*, 2026.

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 2025.

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.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, et al. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. *arXiv preprint arXiv:2601.23224*, 2026.

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.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: A strong zero-shot video understanding model, April 2024.

Yuhan Zhu, Changlian Ma, Xiangyu Zeng, Xinhao Li, Zhiqiu Zhang, Songze Li, Jun Zhang, Tianxiang Jiang, Yuandong Yang, Ziang Yan, et al. Timelens2: Generalist video temporal grounding with multimodal llms. *arXiv preprint arXiv:2607.17423*, 2026.

A Temporal Representation and Boundary-Readout Diagnostics

A.1 Temporal Representation Probes

Layerwise ridge probes at P_0 reveal a pronounced separation between full-video and query-only inputs (Figure 5). Temporal location becomes increasingly decodable with depth in both models, while query-only performance remains low. On the test set, Qwen's center MAE improves from 0.22 to 0.10 with video, and TimeLens2 reaches 0.10. To test query specificity, we pair two non-overlapping queried events in one video and verify whether the decoded event centers recover their temporal order. On the test set, Qwen and TimeLens2 achieve 92.66% and 93.58% accuracy, respectively, versus 66.06% and 71.56% under query-only inputs. This confirms the representation can distinguish the queried target event within a shared visual context.

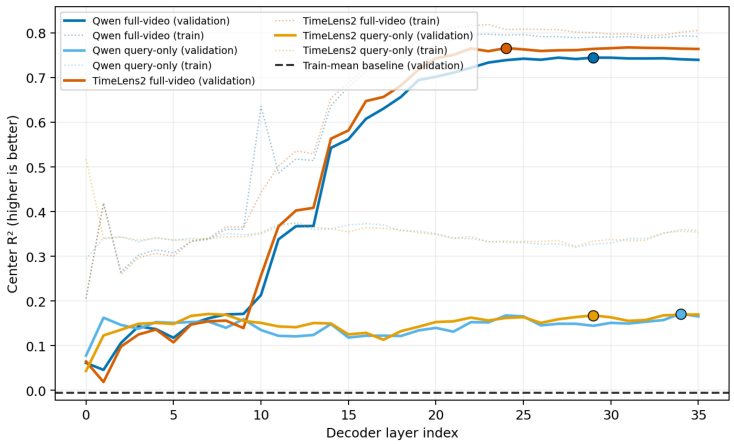


Figure 5: **Layerwise decoding of temporal event centers** from full-video and query-only representations.

Native generation outperforms probe-reconstructed intervals in mIoU: 41.41 versus 32.50 for Qwen, 57.84 versus 34.57 for TimeLens2, and 31.08 versus 23.86 for InternVL3.5.

Video information under fixed priors. For Qwen L30 and TimeLens2 L31, we predict e_i^*/T_i while providing the same query, video duration, and generated-start features to all four conditions. Prior uses these features alone. No-video, Roll, and Full additionally use the corresponding P_2 hidden state. All video conditions reuse the same canonical model-own-start prefix and are compared on the same eligible queries. The fixed-prior controls therefore isolate the additional hidden-state information. Table 6 reports the absolute errors underlying the paired reductions in Table 1. Full improves over Prior by 0.05 in both Qwen and TimeLens2. For InternVL, Full reduces end MAE from 0.27 without video to 0.15, with paired reductions of 0.12 against No-video and 0.06 against Roll.

The remaining-time target $(e_i^* - \hat{s}_i)/T_i$ uses the model's generated start $\hat{s}_i$. With this target, the video advantage persists: the No-video-minus-Full reduction is 0.03 in both Qwen and TimeLens2. Both end-coordinate and remaining-time readouts thus benefit from the hidden-state video information when the start and other priors are already available.

Table 6: **End and remaining-time probes with fixed priors.** Normalized MAE (↓).

Model	Target	Prior	No-video	Roll	Full
Qwen3-VL	End	0.15	0.13	0.12	**0.10**
	Remaining time	0.15	0.13	0.11	**0.10**
TimeLens2	End	0.15	0.13	0.11	**0.09**
	Remaining time	0.14	0.13	0.11	**0.09**

A.2 Boundary Candidate Scoring

Descriptive offset curves. Figure 6 shows Qwen's mean numeric-token costs around the annotated start and end on the test set. Zero offset denotes the annotated boundary. The eligible sample composition varies across offsets. These curves describe average score geometry. The within-query comparisons in Figure 2 assess boundary discrimination on the same query.

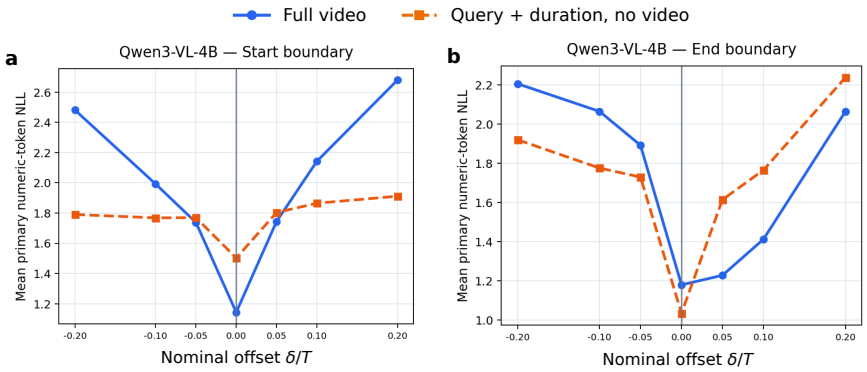


Figure 6: **Qwen's timestamp costs around annotated boundaries.** (a) Start-time offsets. (b) End-time offsets.

Within-query ranking and matched controls. Within each model, we retain queries eligible for both start and end scoring, compute each query's cost–distance Spearman coefficients, and average their paired difference $\rho_{i,\text{start}} - \rho_{i,\text{end}}$. Qwen and TimeLens2 additionally use the common eligible offset grid for the two boundaries. Full-video costs provide the overall ranking. C_{visual} isolates the incremental contribution of video. The visual comparison uses integer offsets and matches numeric-token counts to the annotated candidate. Figure 2 and Table 7 report the resulting means and paired gaps. Pairing is within each model, without requiring the eligible query sets to coincide across models.

Table 7: **Paired boundary discrimination on common queries.** $\Delta\rho$ is the mean within-query start-minus-end difference.

Model	Score	Start ρ	End ρ	$\Delta\rho$
Qwen3-VL	Full-video	0.62	0.40	+0.22
	Visual contribution	0.58	0.19	+0.39
TimeLens2	Full-video	0.69	0.57	+0.12
	Visual contribution	0.59	0.47	+0.12
InternVL3.5	Full-video	0.54	0.37	+0.17
	Visual contribution	0.50	0.21	+0.29

Before token-count matching, the start/end visual correlations are $0.53/-0.04$ for Qwen and 0.53/0.27 for TimeLens2. The matched analysis retains the start advantage. Qwen's end correlation also drops from 0.65 on correctly grounded queries to 0.09 on extent/phase errors.

B Endpoint Subspace Construction and Diagnostic Details

B.1 Native Endpoint Candidates and Scores

Let (s_i, e_i) be the model's native prediction for a video of duration T_i. We retain the predicted start, the exact native text before the end-number span, and the original tail after it. For radii $r \in \{0.05, 0.10, 0.20\}$, candidate pairs target $e_i \pm rT_i$ on a 0.1-second grid and satisfy

$$s_i \le e_{i,r}^{-} < e_i < e_{i,r}^{+} \le T_i. \tag{10}$$

Earlier and later candidates must contain the same number of numeric end tokens. Every scored token must lie entirely inside the end-number span, and tokenization must preserve the boundary between the retained prefix and continuation. Candidate selection prioritizes proximity to the target radius, followed by symmetry of the earlier/later distances. Duplicate pairs are removed. At least two valid radius pairs are required for basis calibration. No ground-truth endpoint is used.

For the original video v_i, query q_i, exact native prefix π_i, and candidate numeric-token sequence $z(e)$, the score used for Jacobian extraction is

$$Q_i(e) = \sum_{t \in \mathrm{numeric}(e)} \log p_\theta\big(z_t(e) \mid v_i, q_i, \pi_i, z_{<t}(e)\big). \tag{11}$$

Only end-number tokens contribute: delimiters, units, and the retained tail are excluded from the sum. Equal token counts make the earlier/later comparison length-matched. Both Jacobian extraction and score mediation use summed numeric-token log-probabilities. The mean numeric-token costs in Section 3.2 serve a separate descriptive diagnostic.

B.2 Jacobian Extraction and Global Aggregation

The exact downstream Jacobian in Eq. (2) is evaluated at the unmodified original-video P_2 residual, under the model's own generated start and exact native prefix. This ties the local basis directly to the model's native endpoint computation. The separately extracted temporally rolled state supplies the temporal contrast for projection. Autodifferentiation uses a numerically identical copy of the native residual, with exact agreement checked for the P_2 states across candidates and for candidate scores against an ordinary forward pass. Model weights stay frozen. For each query, the normalized gradient matrix and its SVD are

$$G^{(\ell)}_{i,r,:} = \frac{g^{(\ell)\top}_{i,r}}{\|g^{(\ell)}_{i,r}\|_2}, \qquad G^{(\ell)}_i = L^{(\ell)}_i S^{(\ell)}_i W^{(\ell)\top}_i. \tag{12}$$

The columns of $U^{(\ell)}_i$ are the right singular vectors with relative singular value at least 10^{-3}, measured against the largest singular value for that query. In the Qwen and TimeLens2 local diagnostics, there are an average of 2.96 valid candidate pairs. 96% of the model–query–site entries have rank 3 and the rest rank 2. No non-finite gradients were observed in these diagnostics.

We project each full–roll difference onto its own local basis, normalize the resulting component independently, and stack the normalized rows to obtain $C^{(\ell)}$ in Equation 4. The global fit first stores the leading 32 right singular vectors. The rank-16 basis consists of its first 16 columns. Equivalently, these are the leading eigenvectors of

$$C^{(\ell)\top} C^{(\ell)} = \sum_{i \in \mathcal{C}_B} \widehat{p}^{(\ell)}_{i,\mathrm{local}} \widehat{p}^{(\ell)\top}_{i,\mathrm{local}}, \qquad \widehat{p}^{(\ell)}_{i,\mathrm{local}} = \frac{p^{(\ell)}_{i,\mathrm{local}}}{\|p^{(\ell)}_{i,\mathrm{local}}\|_2}. \tag{13}$$

Here $\mathcal{C}_B$ is the unlabeled basis-calibration pool. This aggregation preserves the association between a query's endpoint sensitivity and its temporal contrast. The resulting basis is fitted separately for each model and site. In the subsequent shared-direction calibration, the projected differences are averaged at their original magnitudes and only the final mean is normalized, as in Equation 7.

B.3 Layer and Rank Selection

The rank rule selects the smallest rank with split-half subspace overlap at least 0.60, causal saturation relative to rank 32 at least 0.80, and positive paired separation from random, orthogonal, and norm-matched raw-evidence controls. Saturation measures retained causal mediation, not retained activation energy. Rank 16 is the minimum passing rank at the residual sites in Table 8. At L30, split-half stability is 0.76 for Qwen and 0.67 for TimeLens2, while causal saturation is 0.90 and 0.96. Both L30 and L31 pass the unlabeled criteria. The benchmark protocol fixes L30/rank 16 for both models. The layerwise diagnostics evaluate each site separately.

Table 8: **Unlabeled selection evidence for rank 16.** Stability is split-half overlap, and saturation is relative to the rank-32 causal effect. All listed residual sites pass the selection criteria at rank 16. Bold rows mark the deployment default. Layer indices are zero-based.

Model	Residual site	Rank	Stability	Causal saturation
Qwen3-VL	L24	16	0.71	0.84
	L27	16	0.70	0.90
	L30	**16**	**0.76**	**0.90**
	L31	16	0.77	0.94
TimeLens2	L24	16	0.62	0.90
	L27	16	0.64	0.90
	L30	**16**	**0.67**	**0.96**
	L31	16	0.67	0.95

The MLP transport audit gives complementary evidence for L30. Its actual/Jacobian-vector-product transport ratios are 0.62%/0.62% for Qwen and 0.43%/0.40% for TimeLens2. Both exceed random controls, and their associations with causal mediation remain positive on validation and test. L19 does not pass the complete transport criterion. This locates an evidence-transport contribution near the selected residual site. The deployed write uses the residual state.

B.4 Layerwise Causal and Behavioral Results

Controlled temporal contrast. All paired branches reuse the exact full-video-generated P_2 token prefix, including the model's own start. The half-roll permutation changes complete temporal tubelet groups while retaining input IDs, positional inputs, video grid, visual-token count, and within-tubelet spatial structure. Hooks that write back the original activation preserve logits, candidate scores, and greedy token IDs. Greedy recovery continues from the common prefix rather than regenerating it independently in the rolled branch.

Score mediation. For query i, let $\mathbf{s}_i^{(x)} = [S_{i1}^{(x)}, \dots, S_{iK_i}^{(x)}]^\top$ collect the summed numeric-token log-probabilities of its K_i distinct endpoint candidates under condition x. Scores follow Eq. (11) with the corresponding video and activation condition, so $S_{ic}^{(\text{full})} = Q_i(e_{ic})$. Every condition uses the same candidate set and native prefix. Candidate deduplication allows K_i to vary across queries. We center each vector within its own query:

$$\widetilde{\mathbf{s}}_i^{(x)} = \mathbf{s}_i^{(x)} - \left(\frac{1}{K_i} \sum_{c=1}^{K_i} S_{ic}^{(x)} \right) \mathbf{1}. \tag{14}$$

Centering removes score offsets shared by all candidates, focusing the comparison on their relative preferences. It preserves the original score scale, without variance or unit-norm normalization. Given source, target, and intervened score vectors, mediation measures the fraction of the source-to-target distance closed by the intervention:

$$\mathrm{Med}_i = 1 - \frac{\left\| \widetilde{\mathbf{s}}_i^{(\text{int})} - \widetilde{\mathbf{s}}_i^{(\text{target})} \right\|_2}{\left\| \widetilde{\mathbf{s}}_i^{(\text{source})} - \widetilde{\mathbf{s}}_i^{(\text{target})} \right\|_2}. \tag{15}$$

For addition, the source is the temporally rolled condition and the target is the original-video condition. Removal reverses these roles. The intervened states follow Eq. (6). A value of one denotes exact recovery of the centered target curve, zero denotes no reduction in distance, and a positive value denotes closer agreement with the target. Negative values indicate increased distance. Mediation is bounded above by one and is not clipped below zero. For numerical stability, queries with source-to-target distance below 10^{-6} are omitted from the mediation average.

Addition and removal are summarized separately. For the set $\mathcal{I}_a$ of queries with finite mediation under intervention a, the reported effect is

$$\widehat{M}^{(a)} = \frac{1}{|\mathcal{I}_a|} \sum_{i \in \mathcal{I}_a} \mathrm{Med}_i^{(a)}, \qquad a \in \{\text{add}, \text{remove}\}. \tag{16}$$

Each valid query receives equal weight, irrespective of its candidate count. For the diagnostic random-direction control, the ten repeats are averaged within each query, site, rank, intervention, and β before the query-level mean is computed. Combined causal-effect ratios use the mean of $t_i = \mathrm{Med}_i^{(\mathrm{add})} + \mathrm{Med}_i^{(\mathrm{remove})}$ in each compared condition. Confidence intervals use 1,000 video-cluster bootstrap resamples with seed 42: each sampled video contributes all its queries, including repeated copies when sampled multiple times, and the query-level mean is recomputed. The 2.5th and 97.5th percentiles give the 95% interval.

Table 9 reports interventions with $\beta = 1$. Global addition/removal intervals at the sites in Figure 3, panel 2b, are $[0.21, 0.27]$ / $[0.22, 0.28]$ for Qwen L30, $[0.15, 0.22]$ / $[0.14, 0.21]$ for TimeLens2 L31, and $[0.15, 0.23]$ / $[0.16, 0.24]$ for InternVL L30. All exclude zero. InternVL's matched random, orthogonal, and raw-evidence controls give addition/removal mediation of 0.03/0.04, 0.01/0.02, and 0.04/0.03, respectively.

Table 9: **Local and global score mediation on the local test set.** Entries are addition/removal. Global bases have rank 16.

	Qwen3-VL		TimeLens2		InternVL3.5	
Residual site	Local	Global	Local	Global	Local	Global
L24	0.07/0.10	0.15/0.16	0.02/0.06	0.09/0.10		
L27	0.07/0.11	0.16/0.18	0.03/0.07	0.08/0.10		
L30	0.11/0.16	0.24/0.25	0.05/0.12	0.18/0.18	0.08/0.12	0.19/0.20
L31	0.12/0.16	0.23/0.25	0.05/0.13	0.19/0.18		

On validation, the selected global interventions in Qwen and TimeLens2 separate positively from all three matched controls, and the effect direction agrees at $\beta = 0.5$ and $\beta = 1$. On the full test set, L31 global addition/removal remains positive: 0.25/0.25 for Qwen and 0.20/0.19 for TimeLens2.

Greedy boundary recovery. Table 10 shows frozen validation results. At L30, rank-16 addition raises IoU from 28.90 to 36.70 for Qwen, 22.70 to 36.00 for TimeLens2, and 21.50 to 28.50 for InternVL. TimeLens2 recovers more strongly at L31, reaching 37.70 IoU and 6.25 s end MAE. Figure 3, panel 2c, reports paired gains of 7.80 $[5.00, 10.70]$ for Qwen L30, 15.00 $[11.00, 19.00]$ for TimeLens2 L31, and 7.00 $[4.20, 9.80]$ for InternVL L30. These estimates use jointly valid queries. Behavioral evaluation follows configuration freezing and does not select the layer or rank.

Table 10: **Greedy recovery from the common P_2 prefix on validation.** The start remains fixed across full, roll, and intervened branches.

Model	Condition	End MAE (s)	IoU
Qwen3-VL	Full	3.38	42.60
	Roll	5.39	28.90
	Roll + L30 global	4.53	36.70
	Roll + L31 global	4.54	36.90
TimeLens2	Full	2.54	49.90
	Roll	10.02	22.70
	Roll + L30 global	6.76	36.00
	Roll + L31 global	6.25	37.70
InternVL3.5	Full	4.90	32.40
	Roll	7.10	21.50
	Roll + L30 global	5.90	28.50

C Shared Steering Procedure

The offline procedure combines the model-specific endpoint basis with dataset calibration to obtain a shared direction and relative strength. Both are fixed before evaluation. Online generation applies

one update to the residual at P_2, using its norm before the write. The generated start and preceding tokens are preserved, and decoding continues within the same generation call.

Algorithm 1 Shared Endpoint Activation Steering

Offline: frozen model, unlabeled calibration queries, and the configuration in Appendix D.2.

1. Construct the global endpoint basis B using Section 3.3.
2. Extract states from the original and temporally rolled videos under identical model-generated P_2 prefixes and compute projected differences d_j as defined in Section 4.
3. Construct the shared direction u_D and calibrate the relative magnitude ρ_D using Equations 7–8.

Online: a new video and query.

4. Generate with the original video up to P_2 and obtain h_i at layer ℓ.
5. Apply $h_i' = h_i + \rho_D \|h_i\|_2 u_D$ once, then continue greedy decoding.

D Additional Experimental Settings

D.1 Datasets, Splits, and Calibration

We split each dataset by video into training, validation, and test partitions with a 70/15/15 ratio and seed 42. For each model–dataset pair, 256 unlabeled video–query pairs from the training partition provide the calibration budget. Calibration and validation each use one query per video, while the test set retains all queries from its videos. Temporal annotations are not used to construct the shared direction or calibrate its strength. Only calibration samples with an identifiable P_2 prefix contribute to the shared direction and strength.

D.2 Steering and Input Settings

All models use rank $R = 16$ and $\lambda = 0.5$ in Equation 8. Table 11 lists the zero-based decoder layers used for the single residual write at P_2. Each model has an independent endpoint basis, fixed across datasets. Qwen3-VL-4B-Instruct and TimeLens2-4B reuse their diagnostic L30 bases. Other models follow the same construction in Section 3.3. Candidate and SVD settings are given in Appendices B.1 and B.2. Video inputs and decoding settings are summarized in Table 12.

Table 11: **Model dimensions and intervention sites.**

Model	Layers	H_M	Site
Qwen3-VL-4B-Instruct	36	2560	L30
Qwen3-VL-8B-Instruct	36	4096	L30
InternVL3.5-4B	36	2560	L30
InternVL3.5-8B	36	4096	L30
TimeLens2-4B	36	2560	L30
TimeLens2-8B	36	4096	L30
LLaVA-NeXT-Video-7B	32	4096	L27

Table 12: **Video inputs and decoding settings.**

Models	Setting	Value
Qwen3-VL / TimeLens2	Video sampling	2 FPS, 4–512 frames
	Per-frame pixel budget	16,384–230,400
	Total pixel budget	3,670,016
	Frame timestamps	Enabled
InternVL3.5	Frames	32
	Frame size	448×448 pixels
LLaVA-NeXT-Video	Frames	32
	Timestamp format	Absolute seconds: $[s, e]$
All models	Decoding	Greedy, sampling disabled
	Maximum new tokens	512
	Precision	bfloat16

D.3 Comparison Methods

Native and component controls. The baseline is the frozen model's native greedy output. Our method uses a shared projected direction with state-scaled magnitude. To isolate the role of projection, the raw-direction control uses

$$u_D^{\text{raw}} = \frac{\sum_{j \in \mathcal{C}_D} \Delta h_j}{\left\| \sum_{j \in \mathcal{C}_D} \Delta h_j \right\|_2}, \qquad h_i' = h_i + \rho_D \|h_i\|_2 u_D^{\text{raw}}, \tag{17}$$

with the same ρ_D as the projected method. This CAA-style raw-difference control tests the direction construction. To isolate sample-wise magnitude scaling, the fixed-amplitude control keeps the projected direction and uses

$$a_{\text{fixed}} = \rho_D \, \text{median}_{j \in \mathcal{C}_D} \, \|h_{j,\text{full}}\|_2, \qquad h_i' = h_i + a_{\text{fixed}} u_D. \tag{18}$$

These four conditions are evaluated for every model–dataset pair.

Matched low-rank basis controls. The three basis controls use the matched settings in Section 5.4. Let $\mathcal{N}(v) = v/\|v\|_2$ for nonzero v, and let $\mathcal{S}_R$ return the top R right singular vectors of vectors stacked as rows, without centering. Using the local basis U_i from Section 3.3, we compare

$$\begin{aligned} B_{\text{temp}} &= \mathcal{S}_R\big(\{\mathcal{N}(\Delta h_i)\}_i\big), \\ B_{\text{grad}} &= \mathcal{S}_R\big(\{(U_i)_{:,k}\}_{i,k}\big), \\ B_{\text{shuffle}} &= \mathcal{S}_R\big(\{\mathcal{N}(U_{\sigma(i)} U_{\sigma(i)}^\top \Delta h_i)\}_i\big). \end{aligned}$$

Here σ is a fixed random permutation with no query paired to itself. The gradient control pools local orthonormal sensitivity axes. The shuffled control breaks their query correspondence. For each basis, we project the same target-dataset mean contrast and normalize it as in Eq. (7), retaining the complete method's ρ_D to match update norms. Confidence intervals use paired bootstrap resampling by video. The anonymous repository linked on the first page provides the implementation.

Additional training-free comparisons. On Qwen3-VL-4B, we evaluate a numerical endpoint correction $e_i' = e_i + bT_i$, selecting b on validation from $\{-0.10, -0.05, 0, 0.05, 0.10\}$ and resolving ties in favor of the smaller absolute offset. A temporal logit-contrast comparison uses the original and temporally rolled video branches with coefficients in $\{-1, -0.5, 0, 0.5, 1\}$.

FV-Action implementation. We use the same frozen Qwen3-VL-4B as in native decoding. For each local clip, the prompt is *"Does this clip show {query}? Answer Yes or No."* From the next-token logits at the final prompt position, we compute

$$p(\text{Yes}) = \sigma(\text{logsumexp}\, z_{\text{Yes}} - \text{logsumexp}\, z_{\text{No}}), \tag{19}$$

where z_{Yes} and z_{No} collect the corresponding answer-token logits. Coarse probes are spaced uniformly over the video. Each probe supplies three distinct frames, padded to four model-input frames. Temporal non-maximum suppression retains up to three candidate peaks. Around each peak, eight uniformly spaced fine probes are scored with the same question. For a fine probe t around coarse peak c, we combine the scores as $S_{\text{joint}}(t, c) = S_{\text{coarse}}(c)^{0.5} S_{\text{fine}}(t)^{0.5}$. The highest-scoring time becomes the center of a fixed-width output interval. Table 13 lists the dataset-specific search settings. The search constructs the interval externally. It does not modify intermediate activations or autoregressively generate the two timestamps.

Table 13: **FV-Action search settings.**

Dataset	Coarse probes	Fine radius (s)	Output width (s)
Charades-TimeLens	12	± 8	8
QVHighlights-TimeLens	16	± 25	25
ActivityNet-TimeLens	16	± 25	80

D.4 Parameter and Mechanism Controls

Parameter sensitivity. Strength and calibration analyses use Qwen3-VL-4B on all three datasets. A multiplier γ scales the calibrated update as $h_i' = h_i + \gamma \rho_D \|h_i\|_2 u_D$, with $\gamma \in \{0, 0.5, 1, 1.5, 2\}$. The main method fixes $\gamma = 1$. Calibration budgets are $N \in \{32, 64, 128, 256\}$, formed as nested subsets with ordering seed 42. Calibration randomness is examined with seeds 42, 43, and 44. Layer and rank analyses use Qwen3-VL-4B on Charades-TimeLens, with $\ell \in \{24, 27, 30, 33\}$ and $R \in \{4, 8, 16, 32\}$. The default remains L30/rank 16 throughout these analyses. Each sweep in Figure 4 changes one factor while holding the others at their default values.

Direction and temporal contrast. The component ablations in Table 3 use Qwen3-VL-4B and report absolute mIoU and R@0.5, with native decoding and the complete shared-steering method as references. Direction selectivity compares $+u_D$, $-u_D$, and random directions inside the endpoint subspace using seeds 42, 43, and 44, at the same intervention magnitude. A control without temporal contrast replaces the shared difference direction by $\text{normalize}(P \, \text{mean}_j \, h_{j,\text{full}})$. A sample-specific control instead uses $u_i = \text{normalize}(P \Delta h_i)$ while retaining the target dataset's magnitude $\rho_D \|h_i\|_2$. It therefore requires a temporal contrast for each evaluated sample.

Norm and direction. Let $z_i = h_i + \rho_D \|h_i\|_2 u_D$ be the complete update. The norm-only control uses $(\|z_i\|_2 / \|h_i\|_2) h_i$, preserving the original direction. The direction-only control uses $(\|h_i\|_2 / \|z_i\|_2) z_i$, preserving the original norm. Together these controls separate a change in activation direction from a change in its magnitude.

Write position and repetition. The default is one write at P_2. Position controls move the write to the position predicting the final numeric end token. Repetition controls write at every numeric end token, with an additional total-dose-normalized variant that divides the per-write dose by the actual number of writes.

E Additional Quantitative Results

E.1 Endpoint-Only and Error-Group Details

Table 5 in the main text reports the endpoint-only comparison, where the start is fixed to the annotated boundary and only the predicted end determines the interval. Unparseable outputs count as zero in the full-sample denominator. Fixing the start raises native decoding on all nine model–dataset pairs. For example, Qwen3-VL-8B on Charades-TimeLens increases from 34.48 to 36.10 mIoU.

On the standard original-video protocol, Table 14 groups Qwen3-VL-4B queries by their native-decoding IoU. Partially correct intervals benefit most, gaining 3.68–6.08 mIoU, followed by low-overlap intervals at 3.50–4.60. Entirely missed intervals also improve by 2.10–2.60, while already

accurate intervals change little (-0.20 to 0.00). Weighting the group gains by their query shares recovers the overall improvements in Table 2.

Table 14: **Which queries benefit.** Qwen3-VL-4B under the standard original-video protocol, grouped by native-decoding IoU. Shares are query fractions. ΔmIoU is relative to native decoding. Share-weighted gains reproduce Table 2 to the reported precision.

	Charades-TimeLens		QVHighlights-TimeLens		ActivityNet-TimeLens	
Baseline IoU group	Share	ΔmIoU	Share	ΔmIoU	Share	ΔmIoU
$\mathrm{IoU} = 0$ (missed)	0.20	+2.10	0.19	+2.60	0.28	+2.30
$0 < \mathrm{IoU} < 0.3$ (low)	0.19	+3.50	0.14	+4.20	0.19	+4.60
$0.3 \leq \mathrm{IoU} < 0.7$ (partial)	0.39	+3.68	0.24	+5.37	0.24	+6.08
$\mathrm{IoU} \geq 0.7$ (accurate)	0.22	+0.00	0.43	+0.00	0.29	-0.20
All queries (weighted)	1.00	+2.52	1.00	+2.37	1.00	+2.92

E.2 Alternative Endpoint Construction

The deployed method converts the projected temporal contrast into a single signed direction (Section 4). As an alternative, we select one of five sample-specific endpoint actions using the same projected basis. Table 15 shows that the selector and the direct mapping agree within 0.50 mIoU on all three datasets, so the single fixed write used in the main results does not depend on this design choice.

Table 15: **Direct mapping versus five-action selection.** R@0.5 and mIoU for Qwen3-VL-4B. Both variants use the same projected endpoint basis and the target dataset's relative strength.

	Charades-TimeLens		QVHighlights-TimeLens		ActivityNet-TimeLens	
Method	R@0.5	mIoU	R@0.5	mIoU	R@0.5	mIoU
Direct mapping (ours)	49.48	43.93	60.96	56.53	45.48	42.73
Five-action selector	49.69	44.37	60.09	56.76	45.02	42.93

E.3 Additional Training-Free Comparisons

Table 16 lists three further training-free controls on Qwen3-VL-4B: a validation-selected numerical endpoint shift, a temporal logit-contrast combination of the original and temporally rolled branches, and FV-Action under its fixed protocol. Shared endpoint steering yields the largest gains over native decoding. Logit contrast also improves performance, but by a smaller margin. The numerical shift selects a zero offset on every validation set, while FV-Action falls clearly below native decoding. Appendix D.3 gives the corresponding protocols.

Table 16: **Additional training-free comparisons on Qwen3-VL-4B.** ΔmIoU relative to native decoding.

Method	Charades-TimeLens	QVHighlights-TimeLens	ActivityNet-TimeLens
Shared endpoint steering (ours)	+2.52	+2.37	+2.92
Full/roll logit contrast	+0.34	+0.41	+0.20
Validation-selected numeric end shift	+0.00	+0.00	+0.00
FV-Action, fixed protocol	-8.76	-25.31	-20.52

E.4 Cross-Dataset Calibration Transfer

Table 17 summarizes calibration transfer over three dataset pairs. For each transfer direction, we replace the target dataset's own calibration with its partner's calibration and measure the mIoU change relative to target-specific calibration. The table averages the two transfer directions. The pairwise losses range from 1.15 to 1.90 mIoU. Across these pairs, lower direction cosine accompanies a larger loss: QVHighlights–ActivityNet has the highest alignment (0.88) and smallest loss (1.15), whereas Charades–ActivityNet has the lowest alignment (0.64) and largest loss (1.90). The directions remain aligned but are not interchangeable, supporting calibration on the target dataset.

Table 17: **Cross-dataset calibration transfer on Qwen3-VL-4B.** Each row averages the mIoU changes in the two transfer directions, each relative to the target dataset's own calibration. The same-dataset reference is zero. Cosine compares the two calibrated directions.

Dataset pair	ΔmIoU	Direction cosine
Charades–QVHighlights	−1.65	0.75
Charades–ActivityNet	−1.90	0.64
QVHighlights–ActivityNet	−1.15	0.88

E.5 Additional Video VLM Results

Table 18 reports additional video VLM results on the three TimeLens benchmarks.

Table 18: **Additional video VLM results on TimeLens-Bench.**

	Charades-TimeLens		QVHighlights-TimeLens		ActivityNet-TimeLens	
Model	R@0.5	mIoU	R@0.5	mIoU	R@0.5	mIoU
Molmo2-4B (Clark et al., 2026)	31.10	34.70	62.60	60.80	40.30	40.90
VideoChat3-4B (Li et al., 2026b)	64.90	56.10	72.50	67.00	60.50	54.60
InternVideo3-8B (Yan et al., 2026)	61.70	53.20	62.60	59.20	51.40	46.30
Video-o3-7B (Zeng et al., 2026)	34.20	38.60	49.50	47.40	41.00	38.90
VideoChat-Flash-7B (Li et al., 2026a)	37.90	39.70	30.60	32.70	21.80	24.80
MiMo-VL-7B (Yue et al., 2025)	42.60	39.60	42.60	41.50	38.70	35.50
TimeSuite-7B (Zeng et al., 2025)	35.50	38.10	16.90	21.70	17.50	19.80
VideoChat-R1-7B (Li et al., 2025)	30.80	33.70	19.10	21.50	23.90	25.00
Time-R1-7B (Wang et al., 2025c)	32.00	36.60	51.50	49.20	31.00	33.10
TRACE (Guo et al., 2025)	21.80	27.10	39.10	39.00	33.90	32.70
TRACE-uni	22.90	28.10	40.00	39.80	35.10	33.60
Grounded-VideoLLM (Wang et al., 2025a)	28.70	30.00	33.80	33.40	29.60	30.00

F Computational Cost

F.1 Offline Calibration

Each model–dataset pair uses 256 unlabeled calibration samples, which require approximately 510–511 generation or scoring calls. Table 19 reports the eligible samples and the total wall-clock time. This cost is paid once per model–dataset pair and reused for every evaluation query.

Table 19: **Offline calibration cost.** Two hundred fifty-six unlabeled samples per dataset, requiring approximately 510–511 generation or scoring calls.

Dataset	Eligible samples	Total time (s)
Charades-TimeLens	255/256	985.60
QVHighlights-TimeLens	254/256	1125.50
ActivityNet-TimeLens	254/256	1724.10

F.2 Online Deployment

Table 20 compares the deployed single-pass write with a replay implementation that recomputes the intervention in a separate generation call. Both variants produce identical prefixes, token IDs, raw outputs, and endpoint values, but the single-pass write needs one generation call per query instead of about two and reduces the wall-clock time per query by 1.48–1.81 s. Table 21 reports the resulting latency against native decoding.

Table 20: **Online single-pass write versus per-sample replay.** Both variants match on prefixes, tokens, raw outputs, and endpoint values. Replay calls and times are per query.

Dataset	Parity	Replay calls	Online calls	Replay time (s)	Online time (s)
Charades-TimeLens	100%	1.99	1	3.68	2.20
QVHighlights-TimeLens	100%	2.00	1	4.87	3.21
ActivityNet-TimeLens	100%	2.00	1	4.72	2.91

Table 21: **Online latency per query on Qwen3-VL-4B (s).**

Dataset	Native decoding	Shared endpoint steering	Overhead
Charades-TimeLens	2.03	2.20	+0.17
QVHighlights-TimeLens	2.97	3.21	+0.24
ActivityNet-TimeLens	2.86	2.91	+0.05

G Endpoint Subspace Visualization

Figure 7 visualizes endpoint-margin Jacobians inside the rank-16 readout subspace. In the PCA view (panel a), PC1 and PC2 explain 20.20% and 14.90% of the variance, respectively, or approximately 35.10% together. The PCA and t-SNE views reveal structure across candidate radii (panels a and b), while coloring by the native end position shows how endpoint location varies across the embedding (panel c). Panel d compares the Jacobian directions with the calibrated shared direction. Median cosine decreases as the candidate radius grows in all three datasets, showing that alignment with the shared update depends on the temporal scale of the endpoint comparison.

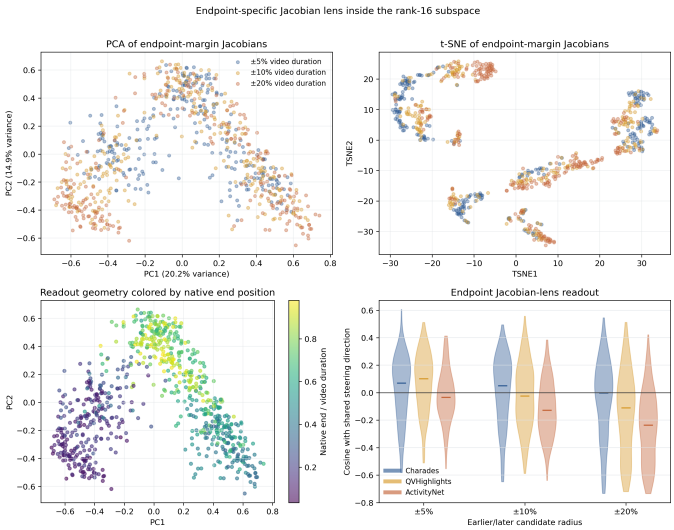


Figure 7: **Endpoint Jacobians in the rank-16 subspace** (Qwen3-VL-4B). (a,b) PCA/t-SNE by radius. (c) Native end position. (d) Cosine with the shared direction.