MEMORY-GROUNDED VIDEO LATENT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-video understanding and complex video reasoning require vision–language models (VLMs) to identify sparse visual cues amid redundant content and integrate evidence across distant moments through complex reasoning. Textual chain-of-thought (CoT) and agent-based approaches incur high inference costs; extended textual reasoning can lose focus on visual evidence, while repeated local inspection can reinforce errors. Although image-based visual latent reasoning reduces reliance on lengthy textual descriptions, we observed that directly transferring it to videos degrades performance, revealing **visual forgetting** and **representation shift**. We propose **Memory-Grounded Video Latent Reasoning (MVLR)**, which interleaves textual reasoning with continuous visual latent reasoning grounded in adaptive visual memory. Guided by textual CoT, the memory preserves keyframe detail, compresses complementary context, and supplies relevant evidence at every latent step. Two-stage supervised training combines CoT and answer supervision with retrieval diversity regularization, learning latent representations without explicit visual-feature targets. Counterfactual-constrained reinforcement learning further assigns contrastive credit by comparing the model's support for correct answers under factual memory with a frozen reference model's support for the same answers under counterfactual memory. MVLR achieves state-of-the-art performance across 5 benchmarks spanning long-video understanding and complex video reasoning, while delivering $2.63\times$ the inference throughput of strong baseline.

1 INTRODUCTION

VLMs have made substantial progress across visual understanding tasks by connecting visual encoders with large language models (LLMs) (Maaz et al., 2024; Bai et al., 2025). However, long-video understanding and complex video reasoning remain challenging: sparse, informative cues are often obscured by redundant content, while evidence distributed across distant temporal positions must be integrated through complex reasoning steps. Recent approaches address these challenges by strengthening textual CoT reasoning (Feng et al., 2025) or introducing agent-based, multi-turn interactions with video environments (He et al., 2026; Yang et al., 2026b; Zeng et al., 2026). Both approaches, however, incur substantial inference costs and face additional limitations. First, as textual reasoning chains grow longer, models become progressively less focused on visual evidence. Second, repeated inspection of local video segments reinforces local errors. Recent advances in visual latent reasoning for images offer a promising direction. By aligning LLM hidden states with visual embeddings, some methods interleave textual CoT with continuous visual latent representations, enabling intermediate computation in a continuous space and reducing reliance on lengthy textual descriptions of visual content (Li et al., 2026; Dong et al., 2025). Yet videos involve longer temporal horizons, greater redundancy, and more complex dependencies across frames, making the extension of this paradigm nontrivial. This raises the central question of our work: *How can we develop effective video latent reasoning that remains focused on sparse visual evidence distributed across the full video and within keyframes throughout the reasoning process?*

To investigate this question, we construct a baseline that directly extends image-based latent reasoning to videos by aligning latent states with keyframe visual features, which we refer to *vanilla video latent reasoning*. This direct transfer degrades performance, and our analysis identifies two key limitations. First, as the reasoning context grows, visual information becomes increasingly diluted: latent states lose visual specificity and contribute less to subsequent reasoning. We term this phenomenon *visual forgetting*. Second, under direct visual feature matching supervision, the distribution of latent

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.

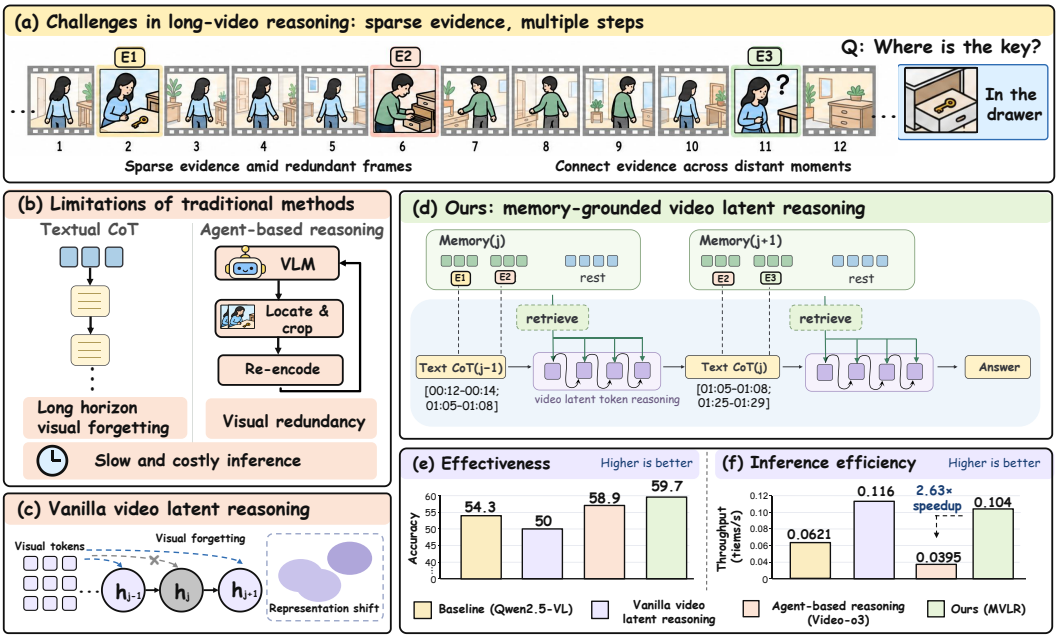


Figure 1: **Overview of challenges and our MVLR**. (a) Illustration of core challenges. (b) Limitations of textual CoT and agent-based video reasoning. (c) Visual forgetting and representation shift in vanilla video latent reasoning. (d) Our MVLR framework pipeline. (e,f) MVLR delivers the best performance while achieving faster inference.

representations shifts toward the training domain, limiting generalization to out-of-domain reasoning tasks. We refer to this limitation as *representation shift*. Together, these limitations compromise the ability of visual latent states to retain useful evidence and support subsequent reasoning.

Building on these insights, we propose **MVLR**, which interleaves textual reasoning with continuous visual latent reasoning and grounds each latent segment in an adaptive visual memory. As illustrated in Figure 1, the preceding textual CoT performs explicit reasoning and identifies relevant temporal intervals. The memory preserves fine-grained frame features within these intervals while compressing the remaining video content, balancing local detail with global coverage. It is updated between latent segments as the reasoning focus changes. At each latent step, the current hidden state queries memory for relevant evidence, which is incorporated into the next step's input. This continual access to visual evidence mitigates visual forgetting during extended reasoning. To improve efficiency and encourage reliance on memory-grounded latents, we use a low-resolution initial video input and mask direct attention to its tokens after the initial textual planning stage.

We train MVLR through two-stage supervised fine-tuning (SFT) followed by reinforcement learning. SFT first initializes the model for textual video reasoning and then trains memory-grounded latent reasoning, with CoT and answer supervision in both stages. Latent representations are learned through the reasoning task without explicit visual feature matching, while a retrieval diversity objective encourages successive latent steps to access complementary evidence. Finally, we introduce *counterfactual-constrained reinforcement learning*, which compares the model's support for correct answers under factual video memory with a frozen reference model's support for the same answers under counterfactual video memory. This comparison provides additional training credit that encourages reliance on informative visual evidence during latent reasoning.

We implement MVLR on Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and evaluate it on seven benchmarks spanning long-video understanding and complex video reasoning. MVLR achieves substantial gains over the strong baselines, with particular strengths on very long videos and challenging reasoning tasks. **Our main contributions are as follows**: 1) We analyze the challenges of long-video understanding and complex video reasoning, and systematically investigate the direct transfer of image-based visual latent reasoning to videos, identifying **visual forgetting** and **representation shift** as two key limitations. 2) We propose MVLR, which continually grounds visual latent reasoning in adaptive visual memory. Its training framework combines supervised fine-tuning without explicit visual feature matching with counterfactual-constrained reinforcement learning. Together, these

designs support effective and efficient video latent reasoning. 3) Experiments on seven benchmarks demonstrate the effectiveness of MVLR for long-video understanding and complex video reasoning, together with 2.63× the inference throughput of strong baseline.

2 Related Work

Long-Video Understanding and Complex Video Reasoning. Video-ChatGPT enables instruction-based video understanding (Maaz et al., 2024), LongVU applies query-guided spatiotemporal compression (Shen et al., 2025), and Video-R1 reinforces temporal reasoning (Feng et al., 2025). FrameThinker, LongVT, and Video-o3 interleave reasoning with repeated frame or clip inspection (He et al., 2026; Yang et al., 2026b; Zeng et al., 2026). Such inspection causes local errors while adding interaction rounds. MVLR introduces a CoT-guided memory of full keyframe features and compressed complementary context. Updated between latent segments and queried at every latent step, this memory supports efficient evidence access beyond the current local focus.

Visual Latent Reasoning. LVR reconstructs query-relevant image features (Li et al., 2026), while Mirage and ILVR interleave text and visual latents with initial intermediate-image feature supervision (Yang et al., 2026a; Dong et al., 2025). For videos, Future-L1 explores latent imagination for event prediction (Jiang et al., 2026), while STORM combines hidden-state feedback with initial thought-video alignment and subsequent answer-only training (Liang et al., 2026). We observe *visual forgetting* and *representation shift* in our direct-alignment video baseline. MVLR therefore retrieves evidence from adaptive memory at every latent step and learns latent states through CoT and answer supervision without explicit visual-feature targets. Counterfactual-constrained reinforcement learning further encourages reliance on retrieved evidence.

3 Is vanilla video latent reasoning effective?

Image-based visual latent reasoning uses vision-encoder features from relevant image regions or intermediate helper images to supervise latent token hidden states during training. At inference, these states are generated through hidden-state feedback. We transfer this paradigm to video by aligning latent token hidden states with vision-encoder features from keyframes, and refer to this baseline as *vanilla video latent reasoning*. However, as shown in Figure 2, we find that this direct transfer degrades performance. Moreover, replacing the learned latent representations with random vectors or zeros leaves performance nearly unchanged, indicating that the model's capabilities depend little on the learned latent content. Further analysis identifies two key limitations of this baseline.

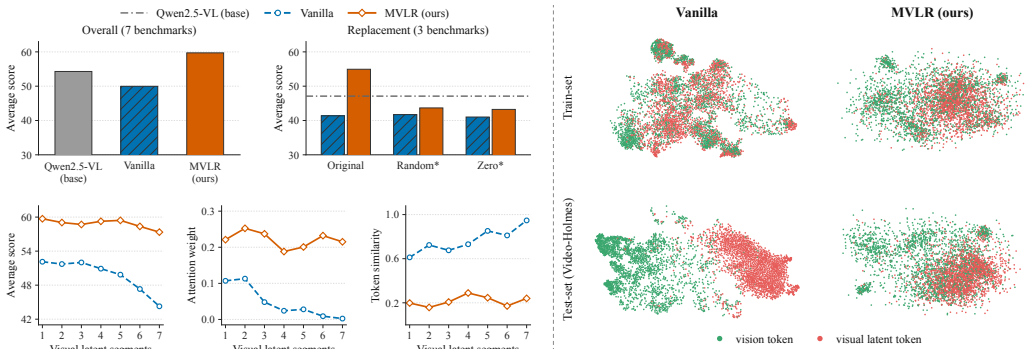


Figure 2: **Diagnostics of video latent reasoning. Top left:** average performance across benchmarks. **Top middle:** average performance with original, random, and zero latent representations. **Bottom:** average performance, attention weight, and latent-token similarity as the number of visual latent segments increases. **Right:** t-SNE visualizations of vision tokens and visual latent tokens.

Visual forgetting. As illustrated in Figure 2 left, Performance degrades as the number of visual latent reasoning segments increases. Attention from generated CoT tokens shifts toward earlier text and away from visual latents, while latent representations become increasingly homogeneous. These trends suggest that visual information is progressively diluted as the reasoning context grows: latent

states lose visual specificity and contribute less to subsequent reasoning, a phenomenon we term *visual forgetting*. **Representation shift.** As illustrated in Figure 2 right, we use t-SNE to visualize visual latent tokens and their corresponding input vision tokens from the training and test sets. Their distributions overlap closely on training examples, whereas latent tokens exhibit a clear, localized shift relative to input vision tokens on test examples. This indicates that directly aligning latent hidden states with prescribed keyframe features induces *representation shift*, limiting their ability to generalize out-of-domain reasoning.

These findings motivate **MVLR**. To counter *visual forgetting*, MVLR interleaves textual and memory-grounded latent reasoning. Textual segments perform explicit reasoning and identify relevant temporal intervals, guiding an adaptive memory that preserves full keyframe features alongside compressed complementary context. Each latent step combines hidden-state feedback with retrieved visual evidence, while subsequent textual segments update the memory as the reasoning focus shifts, continuing until the final answer is produced. We downsample the initial video input and mask direct attention to its tokens after the first textual CoT segment to reduce visual redundancy, encourage reliance on visual latent tokens, and improve inference efficiency. To mitigate *representation shift*, MVLR learns latent representations without explicit visual-feature matching through two-stage SFT followed by counterfactual-constrained RL. Stage I initializes textual video reasoning with CoT and answer supervision; Stage II trains memory-grounded latent reasoning under the same supervision, with visual retrieval diversity encouraging complementary evidence access. RL further encourages the use of visual latent tokens through contrastive credit, comparing the model's support for a correct answer under factual memory with a frozen reference's support under counterfactual memory.

4 METHOD

4.1 ADAPTIVE VISUAL MEMORY

We first describe the construction of adaptive visual memory, which preserves local detail in selected frames while maintaining context from the full video. Given sampled frames $V = \{(I_f, \tau_f)\}_{f=1}^{N_f}$, we encode each frame at its original spatial resolution and cache its projected features $\{\mathbf{x}_{f,p} \in \mathbb{R}^d\}_{p=1}^{P_f}$, where d is the decoder hidden dimension. We augment each feature with temporal and spatial information to form $w_i = [\mathbf{x}_{f,p}; \rho\xi_i]$, indexed by $i = (f, p)$. Here, $\xi_i = \psi(\tau_f, \mathbf{r}_{f,p}) \in \mathbb{R}^{d_p}$ encodes the frame timestamp and spatial position, and ρ controls the relative scale of this positional information. Before each latent reasoning segment Z_j, the preceding textual segment T_{j-1} predicts temporal ranges $\mathcal{G}_j$. We retain all visual tokens from frames within these ranges and use K-means to compress the remaining descriptors into at most B_g context entries. The memory thus preserves detailed features from selected frames alongside a compact representation of the remaining video.

Let $\mathcal{U}_{j,m}$ denote the source tokens represented by memory entry m, comprising either a single retained token or a cluster of context tokens. Each entry stores their mean descriptor and the fraction of source tokens it represents:

$$\mathbf{a}_{j,m} = \frac{1}{|\mathcal{U}_{j,m}|} \sum_{i \in \mathcal{U}_{j,m}} w_i, \qquad \mu_{j,m} = \frac{|\mathcal{U}_{j,m}|}{N}, \qquad \mathcal{M}^{(j)} = \big((\mathbf{a}_{j,m}, \mu_{j,m})\big)_{m=1}^{S_j}. \tag{1}$$

Here, $N = \sum_f P_f$ is the total number of cached tokens, and S_j is the number of memory entries. The normalized masses $\mu_{j,m}$ account for differences in the number of tokens represented by each entry during retrieval. Before each latent segment, we update the selected frames using the latest textual ranges and recluster the remaining tokens. This adapts the visual memory to successive reasoning segments while reusing cached features, without re-encoding the video. Appendix A.1 details the token partition and clustering objective.

4.2 MEMORY-GROUNDED LATENT REASONING

We next describe the MVLR inference pipeline, which alternates textual CoT with memory-grounded latent reasoning. Given an initial prefix $\mathbf{X}_0$ containing low-resolution video tokens and a question q, the decoder generates textual CoT T_0, whose predicted temporal ranges guide the construction of the initial memory. The subsequent `<vis_start>` token activates the *visual seal*, which blocks direct attention to the initial video tokens in subsequent generation. Subsequent positions can still attend

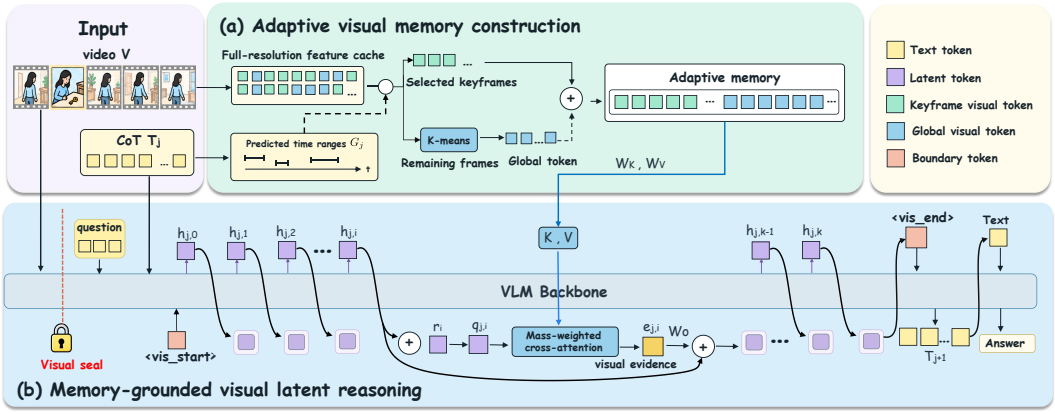


Figure 3: **Overview of MVLR.** (a) CoT-guided memory preserves full-resolution keyframe features and compresses the remaining video context. (b) Each latent step combines hidden-state feedback with retrieved visual evidence.

to preceding text and latent states, which retain information from the video. This routes subsequent visual retrieval through memory and encourages reliance on visual latent tokens. Appendix A.2 provides the attention mask. For latent segment j, we initialize $\mathbf{h}_{j,0} = f_\theta(\mathbf{P}_j; \mathbf{M}_{\text{seal}})$, where $\mathbf{P}_j$ is the prefix through the segment's <vis_start> token and f_θ returns the decoder's last-position hidden state. At step k, we form a memory query from the current hidden state and a learnable register $\mathbf{r}_k \in \mathbb{R}^{d_k}$, which provides an explicit signal of the step's position within the segment:

$$\mathbf{q}_{j,k} = W_Q \,\mathrm{LN}(\mathbf{h}_{j,k-1}) + \mathbf{r}_k, \qquad \mathbf{k}_{j,m} = W_K \mathbf{a}_{j,m}, \qquad \mathbf{v}_{j,m} = W_V \mathbf{a}_{j,m}. \tag{2}$$

The registers are shared across segments and videos, while the memory keys and values are reused within each segment. To account for the number of source tokens represented by each memory entry, we incorporate its mass into the attention weights:

$$\alpha_{j,k,m} = \frac{\mu_{j,m} \exp(\mathbf{q}_{j,k}^\top \mathbf{k}_{j,m}/\sqrt{d_k})}{\sum_{n=1}^{S_j} \mu_{j,n} \exp(\mathbf{q}_{j,k}^\top \mathbf{k}_{j,n}/\sqrt{d_k})}, \qquad \mathbf{e}_{j,k} = \sum_{m=1}^{S_j} \alpha_{j,k,m} \mathbf{v}_{j,m}. \tag{3}$$

Each centroid appears once for every source token it represents. Appendix E establishes this equivalence and bounds the retrieval error relative to attention over the original token cache, for a fixed query, in terms of the clustering distortion. We combine the retrieved evidence with the preceding hidden state to form the next visual latent token:

$$\mathbf{z}_{j,k} = \mathbf{h}_{j,k-1} + W_O \mathbf{e}_{j,k}, \qquad \mathbf{h}_{j,k} = f_\theta([\mathbf{P}_j; \mathbf{z}_{j,1}; \ldots; \mathbf{z}_{j,k}]; \mathbf{M}_{\text{seal}})\,. \tag{4}$$

This residual update preserves hidden-state feedback while incorporating visual evidence selected by the current query at each step. The projections and step registers are learned jointly with the decoder. Emitting <vis_end> closes the latent segment and resumes textual reasoning. The decoder alternates between textual and memory-grounded latent reasoning, updating the memory from the preceding text at the start of each latent segment, until it generates the final answer.

4.3 Learning visual latent reasoning from textual supervision

We next describe the two-stage SFT procedure shown in Figure 4. To address the representation shift associated with direct visual-feature alignment, we train latent states through their contribution to CoT and answer prediction, without explicit feature-matching targets. Stage I trains textual video reasoning using teacher-forced cross-entropy on CoT and answers, with a low-resolution video prefix and ordinary causal attention. Stage II introduces memory-grounded latent segments and the visual seal. Following Video-o3's data-construction pipeline (Zeng et al., 2026), we localize and verify answer-supporting clips and filter reasoning trajectories for logical consistency and agreement with the visual evidence. Each example contains textual CoT segments, annotated evidence ranges, and a final answer, with at least one latent segment.

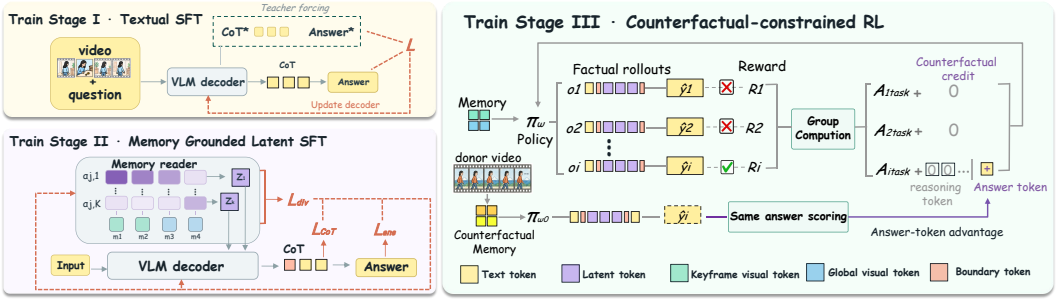


Figure 4: **Training process of MVLR.** Training consists of textual SFT, memory-grounded latent SFT, and counterfactual-constrained RL.

The annotated ranges determine each segment's memory. For segment j, we set the latent length according to the number n_j of distinct sampled frames within these ranges:

$$K_j^\star = \min\{64, \max\{16, n_j\}\}. \tag{5}$$

We generate the $K_j^\star$ latent states using Eq. (4) and apply teacher-forced supervision to the CoT, answer, and boundary tokens. To encourage different latent steps to retrieve complementary evidence, we penalize overlap between their attention distributions. We first rescale attention by the square root of each entry's mass to account for differences in the number of represented source tokens, then normalize the resulting vectors. The penalty averages squared cosine similarity over all step pairs within each segment and across the J segments in a trajectory:

$$\begin{aligned} \mathbf{b}_{j,k} &= \left(\frac{\alpha_{j,k,m}}{\sqrt{\mu_{j,m}}}\right)_{m=1}^{S_j}, \qquad \widehat{\mathbf{b}}_{j,k} = \frac{\mathbf{b}_{j,k}}{\|\mathbf{b}_{j,k}\|_2}, \\ \mathcal{L}_{\text{div}} &= \frac{1}{J}\sum_{j=1}^{J} \frac{2}{K_j^\star(K_j^\star-1)} \sum_{1\le k<\ell\le K_j^\star} \left(\widehat{\mathbf{b}}_{j,k}^\top \widehat{\mathbf{b}}_{j,\ell}\right)^2. \end{aligned} \tag{6}$$

This yields the same squared cosine similarity as uniformly distributing each entry's attention over its represented source tokens, allowing retained tokens and context clusters to be compared on a common basis. Appendix E provides the derivation. The Stage II objective combines token prediction with this diversity penalty:

$$\mathcal{L}_{\text{II}}(\omega) = \mathbb{E}_{x\sim\mathcal{D}_{\text{II}}}\left[-\sum_{t\in\mathcal{C}\cup\mathcal{A}} \log \pi_\omega\left(s_t^\star \mid \mathcal{H}_t^\omega; \mathbf{M}_{\text{seal}}\right) + \lambda_{\text{div}}\mathcal{L}_{\text{div}}(\omega; x)\right]. \tag{7}$$

Here, π_ω is the model's token distribution, and $s_t^\star$ is the target token. The supervised positions include CoT tokens, temporal ranges, and both segment boundaries ($\mathcal{C}$), as well as answer tokens and termination ($\mathcal{A}$). The conditioning history $\mathcal{H}_t^\omega$ contains teacher-forced text and model-generated latent states at the prescribed lengths, while λ_{div} controls the strength of the diversity penalty.

Gradients propagate through all memory reads and the full latent recurrence. The visual encoder and projector remain frozen in both stages. Stage II jointly updates the decoder, retrieval projections, and step registers, collectively denoted by ω. The resulting checkpoint ω_0 initializes RL.

4.4 Counterfactual-constrained reinforcement learning

To encourage greater reliance on visual latent tokens and improve generalization, we augment GRPO (Shao et al., 2024) with a contrastive signal. This signal assigns additional credit to correct answers whose likelihood under the factual trajectory is low relative to a frozen reference using counterfactual visual memory. For each question n, the behavior policy ω_{old} samples $G \geq 2$ trajectories using the source video V_n. Accuracy and format rewards determine the group-normalized task advantage A_{ni}^{task} for trajectory i. The frozen Stage II model ω_0 also generates one reference trajectory per question, using the same question and low-resolution source-video prefix but counterfactual memory features drawn from another training video. We sample this donor video uniformly without using answer labels. The reference generates its own text and latent states, and its predicted temporal ranges determine a memory partition over the source video. We then replace each entry's visual

component with mapped and aggregated donor features, preserving the source partition, positional information, and masses. The substituted features enter both memory keys and values, without reclustering. Appendix A.5 details the mapping and aggregation.

For a correct, well-formed factual trajectory and a well-formed reference prefix, we compare the log probabilities of the *same verified answer* $\widehat{y}_{ni}$. Let $\mathcal{D}_{ni}$ index its sampled text and end/continue decisions, with realized actions a_{nit} and count $N_{ni} = |\mathcal{D}_{ni}|$, and let $\mathcal{A}_{ni} \subseteq \mathcal{D}_{ni}$ index final-answer tokens, including termination. With $L_{ni} = \max(1, |\mathcal{A}_{ni}|)$, the scores are

$$\begin{aligned} s^{+}_{ni}(\omega) &= \frac{1}{L_{ni}} \sum_{t \in \mathcal{A}_{ni}} \log \pi_{\omega}(a_{nit} \mid \mathcal{H}^{\omega}_{nit}), \\ b^{-}_{ni} &= \frac{1}{L_{ni}} \sum_{u=1}^{L_{ni}} \log \pi_{\omega_0}\left(\widehat{y}_{ni,u} \mid \mathcal{H}^{-,\omega_0}_{n}, \widehat{y}_{ni,<u}\right). \end{aligned} \tag{8}$$

The reference prefix $\mathcal{H}^{-,\omega_0}_{n}$ ends before its own answer; we teacher-force each factual answer after this shared prefix. Its score b^{-}_{ni} stays fixed during the update. These scores define a two-candidate contrastive preference p^{cf}_{ni} (van den Oord et al., 2018) and loss $\ell^{\mathrm{cf}}_{ni} = -\log p^{\mathrm{cf}}_{ni}$. Differentiating at a fixed trajectory yields

$$\begin{aligned} p^{\mathrm{cf}}_{ni}(\omega) &= \frac{\exp(s^{+}_{ni}(\omega)/\tau_{\mathrm{cf}})}{\exp(s^{+}_{ni}(\omega)/\tau_{\mathrm{cf}}) + \exp(b^{-}_{ni}/\tau_{\mathrm{cf}})}, \qquad \tau_{\mathrm{cf}} > 0, \\ -\nabla_{\omega} \ell^{\mathrm{cf}}_{ni} &= \frac{1 - p^{\mathrm{cf}}_{ni}(\omega)}{\tau_{\mathrm{cf}}} \nabla_{\omega} s^{+}_{ni}(\omega), \qquad w_{ni} = \frac{1 - p^{\mathrm{cf}}_{ni}(\omega_{\mathrm{old}})}{\tau_{\mathrm{cf}}}. \end{aligned} \tag{9}$$

We freeze w_{ni} for the update and set it to zero for ineligible trajectories. Larger weights correspond to weaker factual support relative to the reference. We add this credit only at final-answer decisions:

$$\widetilde{A}_{nit} = \mathrm{sg}\left(A^{\mathrm{task}}_{ni} + \lambda_{\mathrm{cf}} \frac{N_{ni}}{L_{ni}} w_{ni} \mathbf{1}\{t \in \mathcal{A}_{ni}\} \right), \qquad \lambda_{\mathrm{cf}} > 0. \tag{10}$$

The factor N_{ni}/L_{ni} converts GRPO's per-trajectory average to the answer score's per-token average. For each eligible trajectory, holding the sampled actions and reference score fixed, the extra ascent gradient at $\omega = \omega_{\mathrm{old}}$ is exactly $-\lambda_{\mathrm{cf}} \nabla_{\omega} \ell^{\mathrm{cf}}_{ni}$ (Appendix G). No further centering or normalization is applied, so eligible correct answers can receive credit even when equal group rewards make task advantages zero. Using these augmented advantages, we minimize the clipped GRPO objective:

$$\begin{aligned} \mathcal{L}_{\mathrm{RL}}(\omega) = -\frac{1}{|\mathcal{B}|G} \sum_{n \in \mathcal{B}} \sum_{i=1}^{G} \frac{1}{N_{ni}} \sum_{t \in \mathcal{D}_{ni}} \Big[& \min\{\varrho_{nit} \widetilde{A}_{nit}, \mathrm{clip}(\varrho_{nit}, 1-\epsilon, 1+\epsilon) \widetilde{A}_{nit}\} \\ & - \beta \mathcal{K}_{nit} \Big], \end{aligned} \tag{11}$$

where $\mathcal{B}$ is the question batch, $\epsilon > 0$, and $\beta > 0$. Action distributions are categorical for text and binary for end/continue decisions. The ratio ϱ_{nit} compares current and behavior policies; $\mathcal{K}_{nit}$ is the conditional KL to ω_0 replaying the factual trajectory with *factual* memory (Appendix A.6). Each policy replays the same discrete actions and fixed memory memberships using its own parameters to recompute latent states. Only sampled text and end/continue decisions contribute action probabilities: continuous inputs and forced closure carry none, and a sampled end is counted once. Appendix F derives this likelihood and the stopping law. Answer gradients still pass through retrieval and the entire latent recurrence, jointly training the decoder and memory reader.

5 Experiments

5.1 Implementation Details

MVLR uses Qwen2.5-VL-7B-Instruct with a 32k-token visual context, context-memory budget $B_g = 512$, and positional scale $\rho = 0.5$. The budget B_g caps compressed context entries; retained keyframe tokens are additional. The initial video input preserves aspect ratio, uses spatial dimensions divisible by 28, and is padded to an even N_f frames. Capping per-frame pixels at $28^2 \min\{768, 32768/N_f\}$

Table 1: Comparison on video understanding benchmarks. Bold numbers indicate the best result for each benchmark, and "–" denotes an unreported result.

Methods	Sizes	Long Video Understanding				Video Reasoning		
		Video MME	MLVU	LVBench	LongVideo Bench	Video MMMU	MMVU	Video-Holmes
Native Video MLLMs								
Qwen2.5-VL	7B	65.1	70.2	45.3	56.0	47.4	61.3	34.7
LLaVA-Video	7B	63.3	70.8	–	58.2	–	–	–
Video-R1	7B	61.4	–	–	–	52.4	63.8	38.5
VideoChat-R1	7B	63.6	68.7	–	58.2	–	–	37.5
Rewatch-R1	7B	65.6	–	43.3	–	51.9	–	44.3
Video-Thinker	7B	–	–	37.0	–	–	–	43.2
Agent-Based Video MLLMs								
Video-RTS	7B	63.0	–	–	56.6	52.7	66.4	–
Video-MTR	7B	59.0	59.7	38.6	56.4	–	–	35.7
LOVE-R1	7B	64.3	69.9	46.7	58.0	50.6	65.0	39.5
Conan	7B	60.5	63.4	39.2	56.6	–	–	44.6
LongVT	7B	64.3	–	41.3	–	45.4	–	–
Video-Zoomer	7B	64.6	69.9	44.0	55.9	51.4	65.6	43.8
Video-o3	7B	**66.5**	**72.1**	47.6	60.5	51.7	67.2	46.5
Latent-Based Video MLLMs								
STORM	7B	61.0	–	–	–	–	65.9	37.8
Vanilla video latent reasoning	7B	62.4	65.6	42.0	54.2	43.5	52.7	29.4
MVLR (ours)	7B	66.3	71.9	**48.8**	**61.9**	**53.1**	**68.0**	**48.0**

limits this input to 16k visual tokens after two-fold temporal compression. **SFT.** Using Video-o3 and OneThinker data (Zeng et al., 2026; Feng et al., 2026), Stage I trains textual reasoning for one epoch on 420k cold-start trajectories; Stage II trains for one epoch on 300k tool-use examples adapted for timestamp-guided memory and interleaved latent reasoning. Both use a learning rate of 10^{-5}, global batch size 128, and frozen visual encoder and modality projector. Stage II sets $\lambda_{\text{div}} = 1$ and assigns latent lengths via Eq. (5). We use $K_{\max} = 64$ throughout training and inference, with adaptive `<vis_end>` stopping during RL and inference. **RL.** Starting from Stage II, Stage III trains on data from Video-o3 and OneThinker, using a global batch of 32 questions, $G = 16$ factual trajectories per question, a constant learning rate of 10^{-6}, and $\lambda_{\text{cf}} = 1$. We set the KL coefficient $\beta = 0.01$, clipping threshold $\epsilon = 0.2$, contrastive temperature $\tau_{\text{cf}} = 1$, and format-reward weight $\lambda_{\text{fmt}} = 0.1$. **Evaluation.** We sample at 2 fps, uniformly subsampling over the full duration when videos exceed 768 frames. We follow the official protocols of the seven benchmarks detailed in Appendix B.

5.2 Quantitative Results

Table 1 compares MVLR with native, agent-based, and latent-based video MLLMs across seven benchmarks covering long-video understanding and complex video reasoning. Two main observations emerge. **1) Comparison with latent-based video MLLMs.** MVLR consistently outperforms latent-based baselines across all benchmarks. Relative to vanilla video latent reasoning, MVLR improves MMVU and Video-Holmes by 15.3 and 18.6 percentage points, respectively. These gains demonstrate the effectiveness of memory-grounded video latent reasoning. **2) Comparison with native and agent-based video MLLMs.** MVLR achieves competitive performance on long-video understanding, with particularly strong results on LVBench, which evaluates hour-scale video understanding, and LongVideoBench. It also achieves the best results on all three complex video reasoning benchmarks. Together, these results suggest that MVLR can effectively retrieve and integrate sparse visual evidence across time to support long-video understanding and complex video reasoning.

5.3 Ablation Studies

Component ablation. Table 2 reports component ablations. The "w/o CCRL" variant uses standard GRPO without the counterfactual credit term. Removing AVMC causes the largest degradation, highlighting the importance of adaptive access to visual evidence. Removing the diversity loss

Table 2: Component ablation of MVLR.

Variant	LVBench	MMVU	Video-Holmes
MVLR	**48.8**	**68.0**	**48.0**
w/o AVMC	43.3	61.5	35.2
w/o Visual Seal	45.7	65.2	44.6
w/o div loss	44.1	63.4	40.2
w/o CCRL	44.2	62.7	38.1

Table 3: Latent-token ablation of MVLR.

Variant	LVBench	MMVU	Video-Holmes
MVLR	**48.8**	**68.0**	**48.0**
random	43.4	53.7	33.9
zero	42.9	53.1	33.7

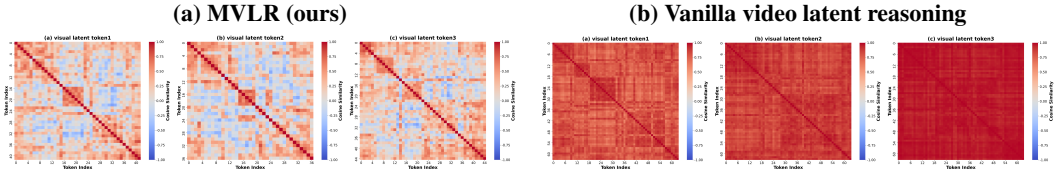


Figure 5: Visual latent token similarity patterns for MVLR (left) and vanilla video latent reasoning (right). MVLR exhibits heterogeneous off-diagonal structure, while the vanilla baseline shows increasingly uniform similarity across the displayed panels.

or CCRL lowers Video-Holmes accuracy by 7.8 and 9.9 points, respectively, underscoring the importance of learning to reason effectively with visual latents. The visual seal yields consistent gains of 2.8–3.4 points, supporting its role in directing visual evidence access through memory. Together, these results show that every component contributes to MVLR's overall effectiveness.

Analysis of latent tokens. Table 3 shows that replacing learned latent representations with random vectors or zeros substantially degrades performance across benchmarks. Both replacements yield similarly low performance, suggesting that arbitrary activations cannot substitute for learned latent content. These findings support the effectiveness of learned latent representations in video reasoning.

5.4 Further Analysis

Analysis of Visual Latent Token Consistency. Figure 5 compares visual latent-token similarity patterns. MVLR exhibits more varied off-diagonal structure than vanilla video latent reasoning, suggesting reduced token homogenization. Together with the latent-token ablations in Table 3, these patterns support the effectiveness of learned visual latents in video reasoning.

Table 4: Accuracy and inference throughput on Video-Holmes.

Model	Accuracy	Items/s
Qwen2.5-VL	34.7	0.0621
LOVE-R1	39.5	0.0318
Video-o3	46.5	0.0395
MVLR	**48.0**	**0.104**

Efficiency Analysis. Table 4 compares accuracy and inference throughput on Video-Holmes under the same hardware and inference configuration for all methods. We measure end-to-end throughput. MVLR achieves 48.0% accuracy at 0.104 items/s, outperforming Video-o3 by 1.5 percentage points while delivering 2.63× its throughput. These results demonstrate that MVLR combines effective video latent reasoning with efficient inference.

6 Conclusion

We presented MVLR to address models that identify sparse visual cues amid redundant content and integrate evidence across distant moments for complex reasoning by addressing visual forgetting and representation shift when extending image-based latent reasoning to videos. MVLR interleaves textual reasoning with continuous visual latent reasoning grounded in adaptive visual memory. Its training combines two-stage supervised learning without explicit visual feature matching, retrieval diversity regularization, and counterfactual-constrained reinforcement learning. MVLR achieves state-of-the-art performance across 5 benchmarks spanning long-video understanding and complex video reasoning, while delivering 2.63× the inference throughput of strong baseline.

AI USE STATEMENT

LLMs were utilized solely for grammatical correction and manuscript polishing to improve the clarity and readability of the writing. All content and findings in this paper are independently designed, verified, and finalized by the authors.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, 2025. URL https://arxiv.org/abs/2502.13923.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-Holmes: Can MLLM Think Like Holmes for Complex Video Reasoning?, 2025. URL https://arxiv.org/abs/2505.21374.

Shuai Dong, Siyuan Wang, Xingyu Liu, Chenglin Li, Haowen Hou, and Zhongyu Wei. Interleaved latent visual reasoning with selective perceptual modeling, 2025. URL https://arxiv.org/abs/2512.05665.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8eb3976840e08b80dda9667562574246-Abstract-Conference.html.

Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, Yan Feng, Peng Pei, Xunliang Cai, and Xiangyu Yue. OneThinker: All-in-one reasoning model for image and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5432–5443, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Feng_OneThinker_All-in-one_Reasoning_Model_for_Image_and_Video_CVPR_2026_paper.html.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. FrameThinker: Learning to think with long videos via multi-turn frame spotlighting. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/79ba1b827d3fc58e129d1cbfc8ff69f2-Abstract-Conference.html.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos, 2025. URL https://arxiv.org/abs/2501.13826.

Tianxiang Jiang, Linquan Wu, Sheng Xia, Songze Li, Ziang Yan, Haoyu Yang, Yu Qiao, and Yi Wang. Imagine before you predict: Interleaved latent visual reasoning for video event prediction, 2026. URL https://arxiv.org/abs/2606.05769.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/efa0939caa8a906fd0a61a9c60e46e8b-Paper-Conference.pdf.

Yiming Liang, Yixiao Chen, Yiyang Zhou, Yixuan Wang, Shoubin Yu, Andong Deng, Fuxiao Liu, Qin Zhang, Chen Chen, Mohit Bansal, and Huaxiu Yao. STORM: Internalized modeling for spatial-temporal reasoning in video-language models, 2026. URL https://arxiv.org/abs/2605.26014.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024. doi: 10.18653/v1/2024.acl-long.679. URL https://aclanthology.org/2024.acl-long.679/.

John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. Gradient estimation using stochastic computation graphs. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/de03beffeed9da5f3639a621bcab5dd4-Abstract.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, 2024. URL https://arxiv.org/abs/2402.03300.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://proceedings.mlr.press/v267/shen25j.html.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. LVBench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_LVBench_An_Extreme_Long_Video_Understanding_Benchmark_ICCV_2025_paper.html.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding, 2024. URL https://arxiv.org/abs/2407.15754.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yang_Machine_Mental_Imagery_Empower_Multimodal_Reasoning_with_Latent_Visual_Tokens_CVPR_2026_paper.html.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, and Lidong Bing. LongVT: Incentivizing "thinking with long videos" via native tool calling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yang_LongVT_Incentivizing_Thinking_with_Long_Videos_via_Native_Tool_Calling_CVPR_2026_paper.html.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, Ziang Yan, Yi Wang, Hongjie Zhang, Yali Wang, and Limin Wang. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning, 2026. URL https://arxiv.org/abs/2601.23224.

Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, Zhijian Xu, Chengye Wang, Weifeng Pan, Ziyao Shangguan, Xiangru Tang, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. MMVU: Measuring Expert-Level Multi-Discipline Video Understanding, 2025. URL https://arxiv.org/abs/2501.12380.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhou_MLVU_Benchmarking_Multi-task_Long_Video_Understanding_CVPR_2025_paper.html.

A METHOD IMPLEMENTATION DETAILS

We describe memory construction, latent decoding, and the training protocols used in Section 4. Algorithm 1 summarizes one latent segment, and Table 5 gives the stopping rules for SFT, RL, and inference. Supporting derivations appear in Appendices E–G.

A.1 MEMORY CONSTRUCTION

For each sampled frame, the frozen encoder and modality projector produce the full-resolution cache

$$\mathbf{X}_f = \mathrm{Proj}(\mathrm{Enc}(I_f)) = [\mathbf{x}_{f,1}; \ldots; \mathbf{x}_{f,P_f}] \in \mathbb{R}^{P_f \times d}, \qquad N = \sum_{f=1}^{N_f} P_f. \tag{A.1}$$

Let $\mathcal{I} = \{(f,p) : 1 \le f \le N_f, 1 \le p \le P_f\}$. Before segment j, the textual ranges $\mathcal{G}_j = \{[a_{j,\ell}, b_{j,\ell}]\}_{\ell=1}^{L_j}$ partition these tokens into keyframe and context sets:

$$\begin{gathered} \mathcal{K}_j = \left\{ f : \tau_f \in \bigcup_{\ell=1}^{L_j} [a_{j,\ell}, b_{j,\ell}] \right\}, \\ \mathcal{I}^{(j)}_{\mathrm{key}} = \{(f,p) \in \mathcal{I} : f \in \mathcal{K}_j\}, \qquad \mathcal{I}^{(j)}_{\mathrm{ctx}} = \mathcal{I} \setminus \mathcal{I}^{(j)}_{\mathrm{key}}. \end{gathered} \tag{A.2}$$

The context budget $B_g \in \mathbb{N}_{>0}$ gives $M_j = \min\{B_g, |\mathcal{I}^{(j)}_{\mathrm{ctx}}|\}$ clusters. K-means on the joint descriptors $w_i = [\mathbf{x}_{f,p}; \rho\xi_i]$ approximately minimizes

$$\min_{\{S_{j,m}, c_{j,m}\}_{m=1}^{M_j}} \sum_{m=1}^{M_j} \sum_{i \in S_{j,m}} \|w_i - c_{j,m}\|_2^2, \quad \bigsqcup_{m=1}^{M_j} S_{j,m} = \mathcal{I}^{(j)}_{\mathrm{ctx}}, \quad S_{j,m} \neq \varnothing. \tag{A.3}$$

We use K-means++ initialization with a single initialization run and a maximum of five iterations for each memory construction. Each context entry is the arithmetic mean of its assigned descriptors. Together with the singleton keyframe entries, these clusters give the cells $\mathcal{U}_{j,m}$ in Eq. (1). For an empty context set, $M_j = 0$ and the memory consists of keyframe tokens. When no frame is selected, clustering covers the full cache.

A.2 VISUAL SEAL AND READER DIMENSIONS

Let $\mathcal{I}_v$ index the low-resolution visual tokens in $\mathbf{X}_0$, and let $\mathcal{I}_{\mathrm{post}}$ contain every position from the first `<vis_start>` onward, including latent, text, control, and answer positions. For query position a and key position b, every decoder layer and attention head applies

$$[\mathbf{M}_{\mathrm{seal}}]_{a,b} = \begin{cases} -\infty, & b > a \text{ or } \quad (a \in \mathcal{I}_{\mathrm{post}} \text{ and } b \in \mathcal{I}_v), \\ 0, & \text{otherwise.} \end{cases} \tag{A.4}$$

Before this boundary, ordinary causal attention applies. The decoder's multimodal positional encoding supplies spatial and temporal information for its input tokens; the memory uses the joint descriptors defined in Section 4.1.

The reader projections have dimensions $W_Q \in \mathbb{R}^{d_k \times d}$, $W_K \in \mathbb{R}^{d_k \times (d+d_p)}$, $W_V \in \mathbb{R}^{d \times (d+d_p)}$, and $W_O \in \mathbb{R}^{d \times d}$. There are $K_{\max} = 64$ registers $\mathbf{r}_k \in \mathbb{R}^{d_k}$. We write $\omega = (\theta, \phi)$, where θ contains decoder parameters and ϕ contains these projections and registers. Text-mode answer generation is autoregressive over token embeddings; each continuous input is instead computed by Eq. (4).

A.3 LATENT-SEGMENT EXECUTION AND MEMORY COST

Algorithm 1 applies after a `<vis_start>` boundary. The ranges come from target text during SFT and generated text during RL or inference. The visual seal remains active after the segment closes. RL replay follows the recorded end/continue decisions.

Algorithm 1 One memory-grounded latent segment

Require: Cached descriptors $\{w_i\}$, ranges $\mathcal{G}_j$, prefix $\mathbf{P}_j$
Require: Mode $\in \{\mathrm{SFT}, \mathrm{RL}, \mathrm{inference}\}$; $K_j^\star$ for SFT
Build cells $\mathcal{U}_{j,m}$ using Eqs. (A.2)–(A.3), and compute their means and masses using Eq. (1)
Project the memory keys and values using Eq. (2)
$\mathbf{h}_{j,0} \leftarrow f_\theta(\mathbf{P}_j; \mathbf{M}_{\mathrm{seal}})$
for $k = 1, \ldots, K_{\max}$ **do**
    Compute $\mathbf{q}_{j,k}$ and mass-weighted evidence $\mathbf{e}_{j,k}$ using Eqs. (2)–(3)
    Form $\mathbf{z}_{j,k}$ and compute $\mathbf{h}_{j,k}$ using Eq. (4)
    if mode is SFT **then**
        if $k = K_j^\star$ **then**
            break
        end if
    else if $k = K_{\max}$ **then**
        break
    else if mode is RL **then**
        Sample $B_k \sim \mathrm{Ber}(p_{j,k}^{\mathrm{end}})$ and record this decision
        if $B_k = 1$ **then**
            break
        end if
    else
        if $p_{j,k}^{\mathrm{end}} \geq \frac{1}{2}$ **then**
            break
        end if
    end if
end for
Append the <vis_end> embedding and return to text mode

Table 5: Segment lengths and stopping rules during training and inference.

Mode	Length rule	Probability or training treatment
SFT	Prescribed $K_j^\star \in [16, 64]$	Cross-entropy on the closing <vis_end> after $K_j^\star$ latent steps.
RL	Sample end/continue; force closure at 64	Binary likelihood and KL for each sampled decision before the cap.
Inference	$p_{j,k}^{\mathrm{end}} \geq \frac{1}{2}$, or the cap is reached	Greedy end/continue selection with a maximum of 64 latent steps.

Memory size and read cost. The memory combines all selected keyframe tokens with at most B_g context entries:

$$S_j = |\mathcal{I}_{\mathrm{key}}^{(j)}| + \min\{B_g, |\mathcal{I}_{\mathrm{ctx}}^{(j)}|\} \leq N. \tag{A.5}$$

The cache stores N visual descriptors of width d together with their positional information and is shared across segments. Each segment builds its partition and projects the memory keys and values once. At each latent step, attention scoring and value aggregation take $O(S_j(d_k + d))$ arithmetic, followed by the output projection and decoder update in Eq. (4). The full computation comprises cache encoding, memory construction, and the interleaved latent and text decoding steps.

A.4 Supervised targets

Stage I uses examples $(V, q, T^\star, y^\star) \in \mathcal{D}_{\mathrm{I}}$. With the entirely textual target $s^\star = [T^\star; y^\star]$, including termination, its standard teacher-forced objective is

$$\mathcal{L}_{\mathrm{I}}(\theta) = -\mathbb{E}_{(V,q,T^\star,y^\star)\sim\mathcal{D}_{\mathrm{I}}} \sum_{t=1}^{|s^\star|} \log \pi_\theta(s_t^\star \mid \mathbf{X}_0, s_{<t}^\star). \tag{A.6}$$

Each Stage II example provides CoT segments $(T_0^\star, \ldots, T_J^\star)$, evidence ranges $(\mathcal{G}_1^\star, \ldots, \mathcal{G}_J^\star)$, and an answer $y^\star$, with $J \geq 1$. The preceding text $T_{j-1}^\star$ specifies $\mathcal{G}_j^\star$, and the number of selected frames is

$$n_j = \left| \left\{ f : \tau_f \in \bigcup_{[a,b] \in \mathcal{G}_j^\star} [a, b] \right\} \right|. \tag{A.7}$$

The timestamps correspond to frames in the sampled cache. Eq. (5) sets each segment's length to $K_j^\star$. Training runs this many latent steps, supervises the closing <vis_end>, and resumes teacher forcing on the target text. Appendix F derives the corresponding training and rollout likelihoods.

A.5 Counterfactual memory construction

For each question, sample a donor $\widetilde{V}_n$ uniformly from training videos other than V_n, without using answer labels. The frozen reference's own predicted ranges determine the source-video partition. Construct one source-to-donor mapping for the question as follows. Normalize source frame times as $t_f = (\tau_f - \tau_1)/(\tau_{N_f} - \tau_1)$ when the denominator is positive, and set $t_f = 0$ otherwise; normalize donor times $\widetilde{t}_g$ analogously. Let $\mathbf{u}_{f,p}, \widetilde{\mathbf{u}}_{g,s} \in [0,1]^2$ be patch centers normalized by their respective frame dimensions. Map each source token to the nearest donor frame in normalized time, then to the nearest spatial patch in that frame:

$$\begin{aligned} g^*(f) &= \arg\min_g |t_f - \widetilde{t}_g|, \\ s^*(f,p) &= \arg\min_s \|\mathbf{u}_{f,p} - \widetilde{\mathbf{u}}_{g^*(f),s}\|_2^2, \qquad \mathcal{T}_n(f,p) = (g^*(f), s^*(f,p)). \end{aligned} \tag{A.8}$$

Both ties are resolved by the smallest index. Donor features use the same frozen encoder and projector as source features. For a memory cell $\mathcal{U}_{j,m}$, replace the visual component by

$$\widetilde{\mathbf{f}}_{j,m} = \frac{1}{|\mathcal{U}_{j,m}|} \sum_{i \in \mathcal{U}_{j,m}} \widetilde{\mathbf{x}}_{\mathcal{T}_n(i)}, \qquad \widetilde{\mathbf{a}}_{j,m} = [\widetilde{\mathbf{f}}_{j,m}; \rho \boldsymbol{\xi}_{j,m}], \tag{A.9}$$

where $\boldsymbol{\xi}_{j,m}$ is the source cell's mean positional descriptor. Both key and value projections use the donor descriptors, while the source partition, positions, masses, and visual seal are retained.

The donor and mapping are shared across a question's reference segments. Each rollout caches its source-token memberships and random seeds for policy replay. The frozen reference generates one reasoning prefix, which is reused to teacher-force every eligible factual answer in the group.

A.6 Rewards, ratios, and policy replay

For factual trajectory τ_{ni}^+ and answer $\widehat{y}_{ni}$, define

$$\begin{aligned} r_{\mathrm{acc},ni} = \mathbf{1}\{\mathrm{Verify}(\widehat{y}_{ni}, y_n^\star) = 1\}, \qquad r_{\mathrm{fmt},ni} = \mathbf{1}\{\mathrm{Valid}(\tau_{ni}^+) = 1\}, \\ R_{ni} = r_{\mathrm{acc},ni} + \lambda_{\mathrm{fmt}} r_{\mathrm{fmt},ni}, \qquad 0 < \lambda_{\mathrm{fmt}} < 1. \end{aligned} \tag{A.10}$$

Valid traces contain at least one non-nested span <vis_start> <vis>^{K_j} <vis_end>, with $1 \leq K_j \leq K_{\max}$, and a final answer in the prescribed format after the last closed span. Here <vis> denotes one continuous latent input. The task advantage is

$$A_{ni}^{\mathrm{task}} = \frac{R_{ni} - \overline{R}_n}{\mathrm{Std}_{i=1}^G(R_{ni}) + \epsilon_A}, \qquad \overline{R}_n = \frac{1}{G} \sum_{i=1}^G R_{ni}, \qquad \epsilon_A > 0. \tag{A.11}$$

Trajectories with $r_{\mathrm{acc},ni} = r_{\mathrm{fmt},ni} = 1$ and a well-formed reference prefix receive contrastive credit. For these trajectories, $L_{ni} = |\mathcal{A}_{ni}| \geq 1$. All other trajectories use $w_{ni} = 0$ and $L_{ni} = \max(1, |\mathcal{A}_{ni}|)$, and skip contrastive scoring.

For every sampled factual decision, the ratio and conditional KL in Eq. (11) are

$$\begin{aligned} \varrho_{nit}(\omega) &= \frac{\pi_\omega(a_{nit} \mid \mathcal{H}_{nit}^\omega)}{\pi_{\omega_{\mathrm{old}}}(a_{nit} \mid \mathcal{H}_{nit}^{\omega_{\mathrm{old}}})}, \\ \mathcal{K}_{nit}(\omega) &= \mathrm{KL}\big(\pi_\omega(\cdot \mid \mathcal{H}_{nit}^\omega) \,\big\|\, \pi_{\omega_0}(\cdot \mid \mathcal{H}_{nit}^{\omega_0,+})\big). \end{aligned} \tag{A.12}$$

Here π is categorical for text and binary for latent end/continue decisions. These sampled decisions determine the ratio and KL terms; appending the closing embedding completes the segment. Appendix F derives their probabilities and the treatment of forced closure.

All policies in Eq. (A.12) replay the same factual discrete history with factual memories. Each independently recomputes its trainable projections, registers, hidden states, and continuous inputs; the current policy retains gradients through these computations. The KL reference uses factual memory along this shared history. The contrastive reference score b_{ni}^{-} is computed separately with donor memory and its generated reasoning prefix, as described in Appendix A.5.

B Evaluation Benchmarks

Our evaluation covers broad video understanding, retrieval over long temporal contexts, and reasoning that combines visual observations with domain knowledge. Table 6 summarizes the published benchmark collections and question formats. For MVLR, we sample videos at 2 fps and uniformly subsample across the full duration when the resulting sequence exceeds 768 frames, as described in Section 5.1.

Table 6: Overview of the seven benchmarks. Coverage describes the published collections; the MLVU entry summarizes its development set. M-Avg and G-Avg summarize multiple-choice and generation tasks, respectively.

Benchmark	Published coverage	Question format	Official metric
Video-MME	900 videos; 2,700 questions	Four-option choice	Accuracy
MLVU	2,593 development questions	Choice and generation	M-Avg; G-Avg
LVBench	103 videos; 1,549 questions	Four-option choice	Accuracy
LongVideoBench	3,763 videos; 6,678 questions	Multiple choice	Accuracy
Video-MMMU	300 videos; 900 questions	Choice and open-ended	Accuracy
MMVU	1,529 videos; 3,000 questions	Five-option choice and open-ended	Accuracy
Video-Holmes	270 videos; 1,837 test questions	Six-option choice	Accuracy

B.1 General and long-video understanding

Video-MME. Video-MME (Fu et al., 2025) evaluates video understanding across diverse content and temporal scales. Its 900 videos each have three human-annotated, four-option questions, covering six domains and 30 subcategories. The collection is balanced across short (under 2 minutes), medium (4–15 minutes), and long (30–60 minutes) videos. Twelve question types assess perceptual recognition, reasoning, and information synthesis. Accuracy is reported overall and by duration, domain, and task, enabling analysis of performance as the temporal context grows. Official evaluation provides settings with and without subtitles; the subtitle-enabled setting selects subtitles aligned with sampled frame timestamps. The original study also evaluates audio as an additional input.

MLVU. MLVU (Zhou et al., 2025) separates long-video understanding into global interpretation, retrieval of individual details, and integration of multiple details. Its videos span movies, surveillance footage, egocentric recordings, animation, and games, with durations from roughly three minutes to over two hours. The development set contains 2,593 questions across seven multiple-choice tasks and two generation tasks. The choice tasks cover topic reasoning, anomaly recognition, needle question answering, egocentric reasoning, plot understanding, action ordering, and action counting. Video summarization and sub-scene captioning assess free-form generation. M-Avg is the mean accuracy across the choice tasks, while G-Avg separately aggregates the generation scores. This task structure distinguishes broad video comprehension from precise retrieval and temporal aggregation.

LVBench. LVBench (Wang et al., 2025) focuses on extended video contexts. Its 103 videos last at least 30 minutes, extend to approximately two hours, and average 4,101 seconds. They cover sports, documentaries, event recordings, everyday activities, television programs, and animation. The 1,549 four-option questions combine six capabilities: temporal grounding, summarization, reasoning, entity

recognition, event understanding, and key information retrieval. Annotations identify the temporal clues supporting each answer. The benchmark reports overall accuracy and capability-specific results over the collection. Questions are based on visually available evidence, and the published evaluation excludes audio. The long sequences and distributed clues emphasize retaining video context while locating the details needed for a particular question.

LongVideoBench. LongVideoBench (Wu et al., 2024) evaluates referring reasoning over video-language inputs. It contains 3,763 videos and 6,678 human-annotated multiple-choice questions across 17 categories, with 1,337 validation and 5,341 test questions. A question uses a visual or textual reference to identify a relevant context, then asks about its details or relations. Videos range from brief clips to one-hour sequences and cover everyday life, movie recaps, news, and knowledge-oriented content. The standard input interleaves video frames with temporally aligned subtitles, followed by the question and answer options. Evaluation reports answer accuracy, with breakdowns by duration and question category, assessing both the retrieval of relevant moments and reasoning over their content.

B.2 Knowledge-intensive and multi-step video reasoning

Video-MMMU. Video-MMMU (Hu et al., 2025) evaluates knowledge acquisition from college-level educational videos. It includes 300 videos and 900 questions covering 30 subjects across art, business, science, medicine, humanities, and engineering. Each video has one question for each of three tracks: perception, comprehension, and adaptation. These progress from extracting information to understanding concepts and applying the demonstrated knowledge to new problems. Questions include choice-based and open-ended formats, with some allowing multiple correct choices. Standard evaluation uses rule-based answer parsing and micro-averaged accuracy. Adaptation questions can also include a question image. The official suite provides transcript variants and a separate knowledge-gain analysis for adaptation.

MMVU. MMVU (Zhao et al., 2025) assesses expert-level understanding across science, healthcare, humanities and social sciences, and engineering. Its 3,000 expert-annotated questions cover 27 subjects and 1,529 videos, requiring visual observations to be combined with domain knowledge. The collection contains 1,858 five-option questions and 1,142 open-ended questions, divided into 1,000 validation and 2,000 hidden test examples. Annotations include reasoning rationales and relevant background knowledge. Video selection emphasizes visual content: audio tracks are removed and excessive on-screen text is avoided. The original accuracy evaluation uses GPT-4o to extract final answers and assess correctness for both formats. Validation data are publicly available, while the benchmark authors manage hidden-test evaluation.

Video-Holmes. Video-Holmes (Cheng et al., 2025) evaluates reasoning over suspense short films. Its test set contains 1,837 six-option single-choice questions from 270 videos lasting one to five minutes. Seven tasks address social relationships, physical anomalies, multimodal hints, character intentions, temporal causality, event chronology, and core themes. Answering requires locating and connecting clues distributed across video segments. Dataset construction uses structured human annotations of plots, relationships, reasoning scenes, and themes to develop questions and answer explanations. The official evaluator extracts the selected option and reports overall and task-specific accuracy. A separate analysis compares generated reasoning with annotated explanations, and the original study also examines the contribution of audio.

C Additional Ablations

C.1 Temporal-range perturbations

Table 7 compares MVLR with two temporal-range perturbations applied at every latent segment while preserving the selection budget. Random replaces the predicted event intervals with randomly selected intervals; Shift uses reversed event intervals. Both variants change the temporal ranges used to construct visual memory. They are distinct from the random/zero latent-vector replacements in Table 3.

Table 7: Accuracy (%) under temporal-range perturbations at every latent segment.

Variant	LVBench	MMVU	Video-Holmes
MVLR (original)	**48.8**	**68.0**	**48.0**
Random ranges	44.4	64.2	37.1
Shift (reversed ranges)	43.8	63.0	36.4

Compared with MVLR, Random reduces accuracy by 4.4, 3.8, and 10.9 percentage points on LVBench, MMVU, and Video-Holmes, respectively. Shift yields drops of 5.0, 5.0, and 11.6 points. Both interventions reduce accuracy across all three benchmarks, with the largest observed decrease on Video-Holmes.

C.2 Definition of the memory ablation

The "w/o AVMC" variant in Table 2 performs visual latent reasoning without the visual memory mechanism. The visual seal is also disabled, retaining direct attention to the initial video tokens. This variant therefore removes both visual memory and the visual seal from the full MVLR configuration.

D Qualitative Results

We examine the structure of visual latent representations through the similarity matrices in Figures 6–8 and the training/test projections in Figure 2. The first pair enlarges the matrices in Figure 5; the remaining pairs provide additional visualizations.

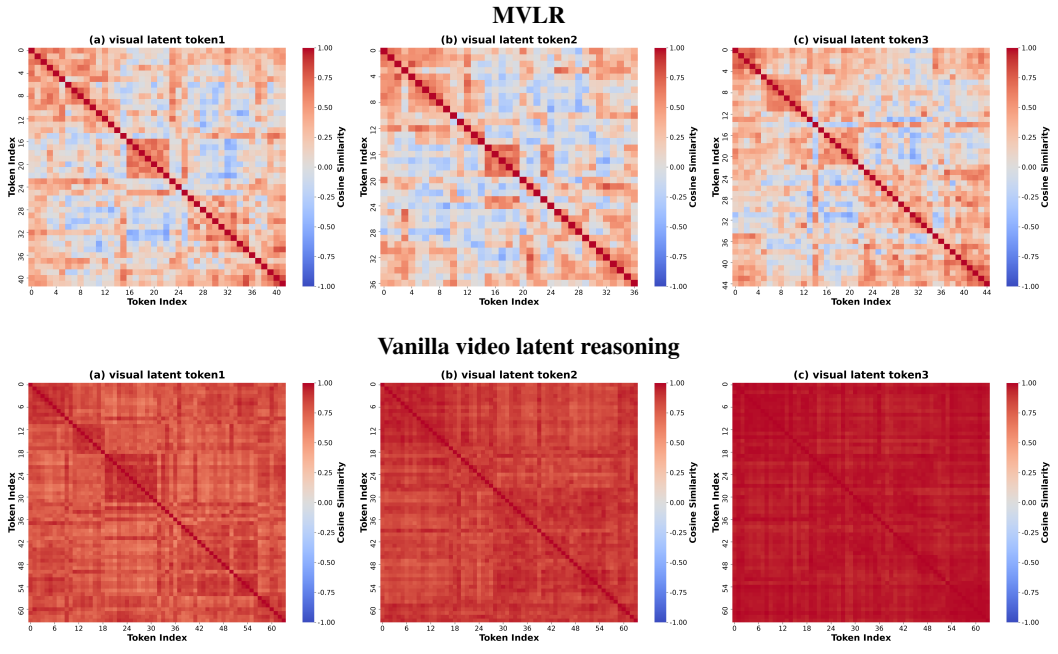


Figure 6: Enlarged visual latent-token cosine similarities from Figure 5: MVLR above, vanilla video latent reasoning below. Both axes index tokens; colors span $[-1, 1]$.

Structure of visual latent representations. The MVLR matrices show varied off-diagonal relationships across the displayed visualizations. The vanilla matrices show broader, more uniform regions, with less separation between groups of tokens. These patterns illustrate the greater differentiation of MVLR's latent representations and the more homogeneous structure of the vanilla representations.

Training and test representations. The t-SNE plots in Figure 2 compare vanilla visual latent tokens with their vision-encoder counterparts. The training-set representations are drawn from 1,000 samples, while the test-set representations use the full Video-Holmes test set. The two point clouds overlap broadly in the training-set projection, whereas the Video-Holmes projection shows a clearer separation. This contrast visualizes the train–test representation mismatch discussed in Section 3. Together, the similarity matrices and projections provide complementary views of representation structure: the former show relationships among latent tokens, and the latter show their relationship to the original visual features.

D.1 Case Studies

Figures 9–11 present qualitative comparisons between Qwen2.5-VL-7B-Instruct and MVLR on three Video-Holmes examples. Each figure includes the question, selected video frames, and the corresponding model responses. The MVLR trajectories show how evidence from different moments informs successive reasoning steps.

For questions Figures 9 and 11, Qwen2.5-VL-7B-Instruct selects an incorrect option, whereas MVLR connects evidence across frames and selects the annotated answer. For question Figure 10, the displayed Qwen2.5-VL-7B-Instruct response repeats observations and ends without a final answer; MVLR follows the sequence of events and selects the annotated answer. These cases illustrate temporal evidence integration for interpreting a changing visual detail, following a causal sequence, and inferring a character's intent.

E Mass-preserving memory analysis

The memory design combines token-count weighting, attention diversity, and centroid compression. We connect these choices to the source-token representation and derive a readout error bound for the joint visual–positional descriptors used in Section 4.1.

Notation. Fix a latent segment j and suppress its index. Its memory induces a partition $\mathcal{I} = \bigsqcup_{m=1}^{S} \mathcal{U}_m$ of the $N \geq 1$ cached tokens. A retained key-frame token is a singleton cell; a context entry corresponds to one nonempty clustering cell. Thus

$$n_m = |\mathcal{U}_m| \geq 1, \qquad \mu_m = \frac{n_m}{N}, \qquad \sum_{m=1}^{S} \mu_m = 1, \\ \mathbf{a}_m = \frac{1}{n_m} \sum_{i \in \mathcal{U}_m} w_i. \tag{E.1}$$

The symbols $w_i \in \mathbb{R}^{d+d_p}$ denote the descriptors from Section 4.1. For a query $\mathbf{q} \in \mathbb{R}^{d_k}$, let $s_m = \mathbf{q}^\top W_K \mathbf{a}_m / \sqrt{d_k}$, $\mathbf{v}_m = W_V \mathbf{a}_m$, and $Z = \sum_m \mu_m e^{s_m} > 0$. The memory attention and readout are $\alpha_m = \mu_m e^{s_m}/Z$ and $\mathbf{e} = \sum_m \alpha_m \mathbf{v}_m$.

E.1 Mass-weighted retrieval

Proposition E.1 (Multiplicity of memory entries)**.** *With the integer counts in Eq.* (E.1)*, the mass-weighted memory readout equals ordinary softmax attention over a sequence containing n_m identical copies of each $\mathbf{a}_m$. Equivalently,*

$$\alpha_m = \frac{n_m e^{s_m}}{\sum_r n_r e^{s_r}} = \left[\mathrm{softmax}((s_r + \log n_r)_{r=1}^{S})\right]_m. \tag{E.2}$$

Proof. Each copy of $\mathbf{a}_m$ receives softmax probability $e^{s_m}/\sum_r n_r e^{s_r}$. Summing over its n_m copies gives Eq. (E.2), since the factor $1/N$ in $\mu_m = n_m/N$ cancels. All copies have value $\mathbf{v}_m$, so their total readout contribution is $\alpha_m \mathbf{v}_m$.

Mass weighting therefore assigns each centroid its source-token multiplicity. It can be implemented by adding $\log n_m$ to the attention logit of entry m.

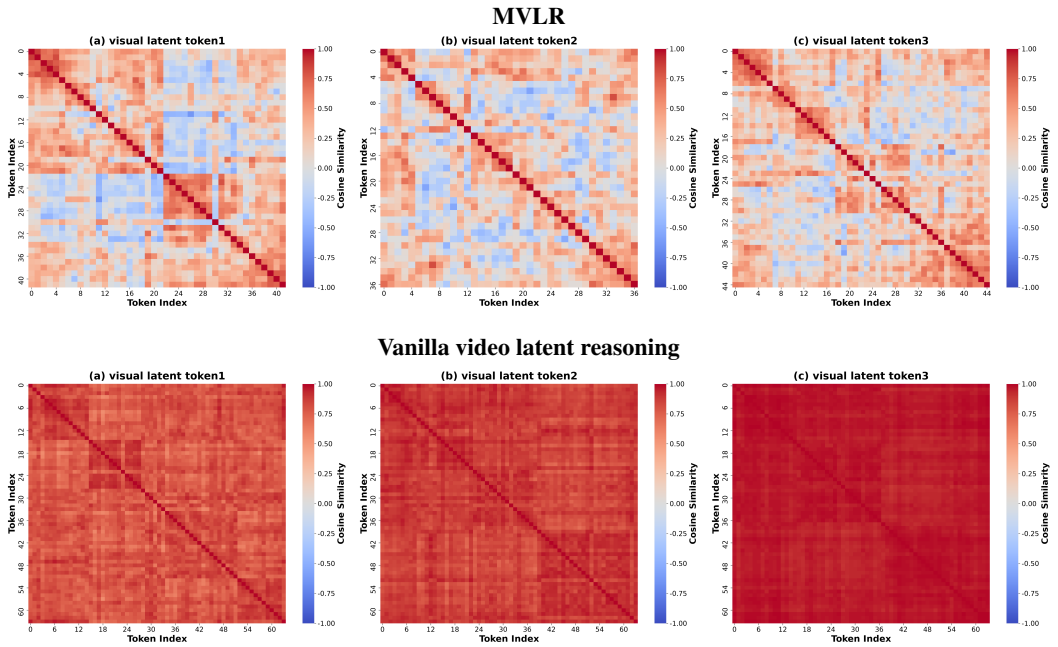


Figure 7: Visual latent-token cosine similarities (visualization 2): MVLR above, vanilla video latent reasoning below. Both axes index tokens; colors span $[-1, 1]$.

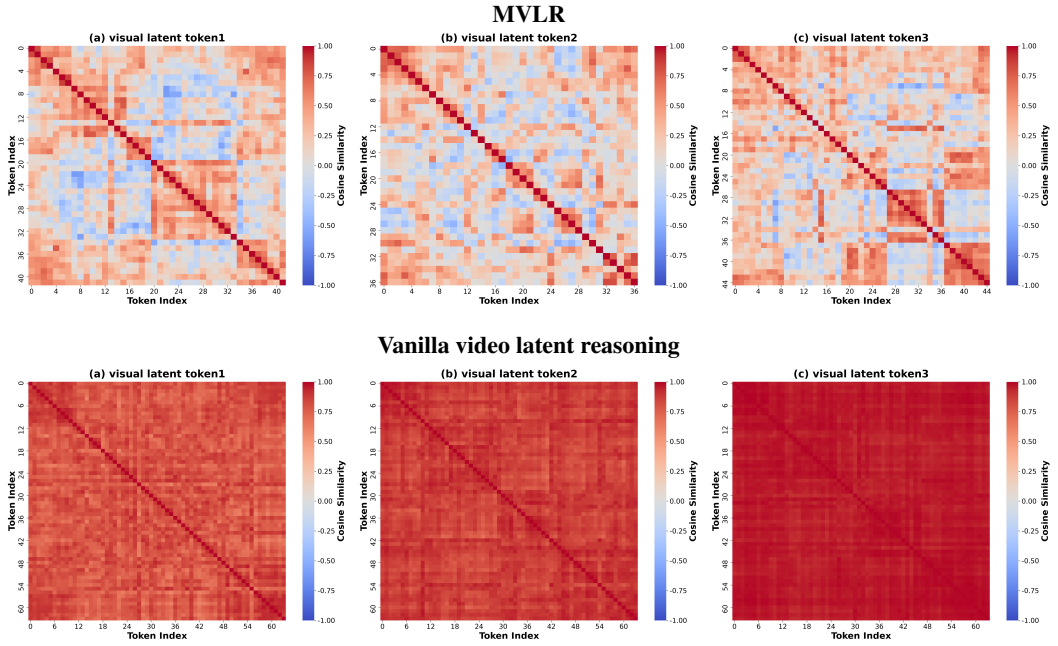
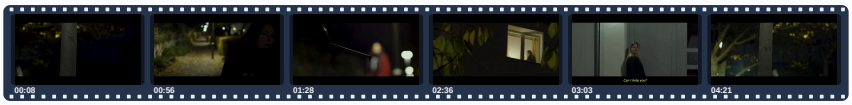


Figure 8: Visual latent-token cosine similarities (visualization 3): MVLR above, vanilla video latent reasoning below. Both axes index tokens; colors span $[-1, 1]$.

Video:



Question: What is the direct reason for the addition of two marks on the street lamp post?

A. Two new victims were killed.
B. Man marks patrol route
C. The brown-haired woman carved out a distress signal.
D. Statistics of street light maintenance times
E. Community event countdown
F. The coincidence of animal scratches

Qwen2.5-VL-7B-Instruct

<think>Let's analyze the video step by step to determine the reason for the addition of two marks on the street lamp post:

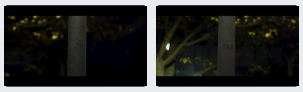
1. The video starts with a dark scene with the text "Alone at Night," setting a mysterious and possibly ominous tone.
2. The scene transitions to a woman walking down a dimly lit path at night, which suggests a sense of isolation and unease.
3. The woman encounters a man who appears to be in distress, as indicated by his body language and the fact that he is leaning against a pole.
4. The man's actions, such as leaning against the pole and the subsequent scene where he is seen with a knife, suggest a violent or threatening situation.
5. The woman's reaction to the man's actions, including her running away and the subsequent scene where she is seen with a knife, further indicates that she is in danger.
6. The scene where the woman is seen with a knife and the marks on the street lamp post suggest that the marks might be related to the violent events that have occurred.
7. The marks on the street lamp post could be interpreted as a distress signal or a warning, which aligns with the woman's actions and the overall tone of the video.

Given this analysis, the most logical reason for the addition of two marks on the street lamp post is that the brown-haired woman carved out a distress signal, as this aligns with the violent events and the woman's actions in the video.

Therefore, the correct answer is C. The brown-haired woman carved out a distress signal.
</think>
<answer>C. The brown-haired woman carved out a distress signal.</answer>

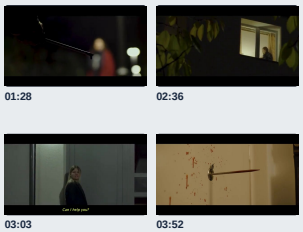
MVLR (Ours)

The question links two new marks to an event in the story, but their meaning is still open. They might record something the man has done, or have been made for another purpose. I will compare key frames from **[[6,10],[259,263]]** seconds to establish how the markings change between the opening and closing views. That comparison should identify the count before I connect it to the people and events between those views.



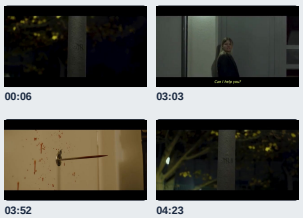
<vis_start>
<vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis>
<vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis>
<vis><vis><vis><vis><vis><vis>
<vis_end>

The opening view has five tally strokes, and the closing view has seven. Their arrangement is consistent with an updated tally, although the marks alone do not identify what is being counted. A record of patrols and a record of victims could both take that form. To distinguish them, I will retrieve key frames from **[[84,92],[152,160],[177,185],[220,232]]** seconds, examining the man's object, the view of the brunette's window, the encounter at the entrance, and the later incident at her door. The explanation should connect those events to an increase of two.



<vis_start>
<vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis>
<vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis>
<vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis>
<vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis>
<vis_end>

A gloved hand holds a slender metal implement, and the later views show a rod protruding from the brunette's door with blood marks around it. Another woman turns toward someone off screen at the building entrance, but what happens to her is not shown. This leaves one attack supported by its aftermath and a tally that has increased by two. I will revisit key frames from **[[6,10],[177,185],[220,232],[259,263]]** seconds to check how the entrance scene and the attack's aftermath relate to the added marks.



<vis_start>
<vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis>
<vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis>
<vis><vis><vis><vis><vis><vis><vis>
<vis_end>

A count of victims best connects the added marks with the violent events in the story. The attack at the brunette's door supplies direct evidence of violence; the blonde's encounter with someone off screen suggests another encounter with the attacker. Her death is implied by that scene and the increase of two, rather than shown directly. Patrols, maintenance, a community countdown, and a distress signal do not explain this sequence as well. The best answer is A, with the second killing understood as an inference from the ending.

Ground truth: A **MVLR output: <answer>A</answer>**

Figure 9: Qualitative comparison on Video-Holmes question 833.

Video:



Question: What is the direct reason for the young man and woman to finally meet?

A. Both people make a wish at the same time.
B. The elf throws gold coins to trigger gas release.
C. The bearded man secretly helps.
D. The bus is broken.
E. The guidance of dandelions
F. Coins drop naturally.

Qwen2.5-VL-7B-Instruct

<think>Let's analyze the sequence of events in the video to determine the direct reason for the young man and woman to finally meet.

1. The video starts with the young man and woman separately, with the man making a wish at the fountain.
2. The man's wish is granted, and he is seen with a surprised expression.
3. The woman is then shown in a flower shop, seemingly unaware of the man's wish.
4. The man is then seen in a café, still looking surprised.
5. The woman is shown again in the flower shop, and the man is seen outside the flower shop.
6. The man is then seen in a bus, and the woman is shown again in the flower shop.
7. The man is seen outside the flower shop, and the woman is shown again in the flower shop.
8. The man is then seen in a bus, and the woman is shown again in the flower shop.

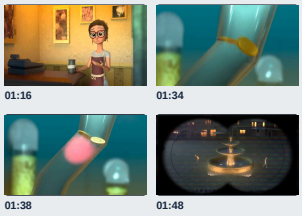
[Items 9-86 omitted here because of repetition]

87. The man is seen outside the flower shop, and the woman is shown again in the flower shop.
88. The man is then seen in a bus, and the woman is shown again in the flower shop.
89. The man is seen outside the flower shop, and the woman is shown

The supplied output ends mid-sentence at item 89, without a final answer.

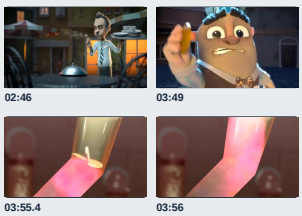
MVLR (Ours)

The question asks for the event that finally brings the young man and woman together, so I should separate their wishes from the mechanism that produces the meeting. The options mention coins, gas, dandelions, and several incidental events. I will first revisit **[[74,101],[104,113]]** seconds, covering the two wishers at the fountain and what happens inside the apparatus immediately afterward. This should show whether the original wishes succeed on their own or leave an unresolved obstacle that requires a later intervention.



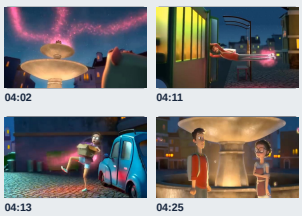
<vis_start>
<vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis>
<vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis>
<vis><vis><vis><vis><vis><vis><vis><vis><vis><vis>
<vis_end>

Two coins lodge across a bend in the transparent pipe, and pink gas collects beneath them. The man and woman remain apart afterward, so making their wishes is not sufficient to explain the eventual meeting. I need to identify which later action changes that blocked state. I will inspect **[[146,169],[214,237]]** seconds to compare the flower-based attempt with the elf's later handling of a new coin, including its path into the fountain and the pipe. The decisive difference should be a visible change in the stalled mechanism.



<vis_start>
<vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis>
<vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis>
<vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis>
<vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis>
<vis_end>

The dandelion reaches another man rather than the intended young man. Later, the elf deliberately throws an additional coin into the fountain; it strikes the lodged coins, and the pink gas moves through the cleared bend. That links the elf's action to a release, but the question concerns the couple's meeting. I will now retrieve **[[237,254],[255,275]]** seconds to follow where the released gas goes and how the two people respond. This final check should distinguish the actual cause of their convergence from events that merely occur nearby.



<vis_start>
<vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis>
<vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis><vis>
<vis><vis><vis>
<vis_end>

The pink gas bursts from the fountain, reaches the young man as he is boarding the bus, and also pulls the woman away from her previous position. Both end up together at the fountain. This completes the causal sequence: the elf throws the extra coin, the blocked coins are displaced, the gas escapes, and the two wishers are brought together. The dandelion attempt had reached someone else, while the initial wishes had left the pipe jammed. The direct cause therefore matches option B.

Ground truth: B **MVLR output: <answer>B</answer>**

Figure 10: Qualitative comparison on Video-Holmes question 1705.

Video:



Question: What was the real intention of the man in the shirt asking Bran to register and enter his phone number at 1:31?

A. Convenient for contact in the future for meal delivery.
B. Test mobile phone signal strength
C. Share your own social account.
D. Confirm Brandon's identity
E. Show spelling difficulty
F. Make opportunities to take away the mobile phone.

Qwen2.5-VL-7B-Instruct

<think>
The video shows a man in a shirt asking Bran to register and enter his phone number at 1:31. This action is typically associated with confirming identity or establishing a connection for future purposes. The man's intention is likely to verify Bran's identity, which is a common practice in various contexts, such as security checks or verifying someone's presence in a specific location. Therefore, the most plausible intention is to confirm Brandon's identity.

The other options do not fit the context of the scene:
- "Convenient for contact in the future for meal delivery" is not a typical reason for asking for a phone number.
- "Test mobile phone signal strength" is not a common reason for asking for a phone number.
- "Share your own social account" is not a reason for asking for a phone number.
- "Make opportunities to take away the mobile phone" is not a reason for asking for a phone number.
- "Show spelling difficulty" is not a reason for asking for a phone number.

</think>
<answer>D. Confirm Brandons identity</answer>

MVLR (Ours)

The question asks for the purpose behind the request, so I need to distinguish an ordinary exchange of contact details from an attempt to gain control of the phone. The choices also mention identity, signal strength, a social account, and spelling, which would require different evidence. I will inspect **[[91,125]]** seconds, covering the transition into the coffee conversation and its immediate context. First I should establish how the two men interact, without treating the request itself as proof of either a helpful or a deceptive intention.

<vis_start><vis><vis><vis

E.2 MASS-NORMALIZED ATTENTION DIVERSITY

Proposition E.2 (Token-level interpretation of diversity)**.** *For each latent step k within a fixed segment, uniformly expand its memory attention to a distribution over the original token indices by setting*

$$p_{k,i} = \frac{\alpha_{k,m}}{n_m} \quad (i \in \mathcal{U}_m), \qquad \mathbf{p}_k = (p_{k,i})_{i\in\mathcal{I}}. \tag{E.3}$$

Let $\mathbf{b}_k = (\alpha_{k,m}/\sqrt{\mu_m})_{m=1}^S$, as in Eq. (6)*. Then, for every pair k, ℓ,*

$$\frac{\mathbf{b}_k^\top \mathbf{b}_\ell}{\|\mathbf{b}_k\|_2\|\mathbf{b}_\ell\|_2} = \frac{\mathbf{p}_k^\top \mathbf{p}_\ell}{\|\mathbf{p}_k\|_2\|\mathbf{p}_\ell\|_2}. \tag{E.4}$$

Thus, for any segment with at least two steps, Eq. (6) *averages squared cosine overlaps of these uniformly expanded token distributions.*

Proof. The entries of $\mathbf{p}_k$ are nonnegative and sum to one. Using $\mu_m = n_m/N$, we obtain

$$\mathbf{p}_k^\top \mathbf{p}_\ell = \sum_{m=1}^S \frac{\alpha_{k,m}\alpha_{\ell,m}}{n_m} = \frac{1}{N}\mathbf{b}_k^\top \mathbf{b}_\ell. \tag{E.5}$$

The same factor applies to the nonzero squared norms, proving Eq. (E.4).

The factor $1/\sqrt{\mu_m}$ thus measures overlap in source-token space while computing the loss over only S memory entries. Each cell's attention is shared uniformly among its represented tokens before their overlap is evaluated.

E.3 COMPRESSION DISTORTION AND READOUT ERROR

Proposition E.3 (Fixed-query compression error)**.** *Fix the cached descriptors, a nonempty-cell partition, linear projections W_K, W_V, and a query $\mathbf{q}$. Assume $\|w_i\|_2 \le R$ for every i, with $R \ge 0$, and use the exact arithmetic means and masses in Eq.* (E.1)*. Define*

$$\begin{gathered}\boldsymbol{\beta} = \frac{W_K^\top \mathbf{q}}{\sqrt{d_k}}, \qquad a = \|\boldsymbol{\beta}\|_2, \qquad v = \|W_V\|_{\mathrm{op}}, \\ D = \frac{1}{N}\sum_{m=1}^S \sum_{i\in\mathcal{U}_m} \|w_i - \mathbf{a}_m\|_2^2.\end{gathered} \tag{E.6}$$

Let $\mathbf{e}_{\mathrm{full}}$ be ordinary attention over all cached descriptors with this same query, and let $\mathbf{e}_{\mathrm{mem}}$ be the mass-weighted compressed readout:

$$\mathbf{e}_{\mathrm{full}} = \frac{\frac{1}{N}\sum_{i\in\mathcal{I}} e^{\boldsymbol{\beta}^\top w_i} W_V w_i}{\frac{1}{N}\sum_{i\in\mathcal{I}} e^{\boldsymbol{\beta}^\top w_i}}, \qquad \mathbf{e}_{\mathrm{mem}} = \frac{\sum_{m=1}^S \mu_m e^{\boldsymbol{\beta}^\top \mathbf{a}_m} W_V \mathbf{a}_m}{\sum_{m=1}^S \mu_m e^{\boldsymbol{\beta}^\top \mathbf{a}_m}}. \tag{E.7}$$

Then

$$\boxed{\|\mathbf{e}_{\mathrm{full}} - \mathbf{e}_{\mathrm{mem}}\|_2 \le e^{2aR}\, v\, a(1+aR)\, D.} \tag{E.8}$$

Proof. Let $g(x) = e^{\boldsymbol{\beta}^\top x}$ and $F(x) = e^{\boldsymbol{\beta}^\top x} W_V x$. Their second directional derivatives are

$$\begin{aligned}\mathrm{D}^2 g(x)[h,h] &= e^{\boldsymbol{\beta}^\top x}(\boldsymbol{\beta}^\top h)^2, \\ \mathrm{D}^2 F(x)[h,h] &= e^{\boldsymbol{\beta}^\top x}\big[(\boldsymbol{\beta}^\top h)^2 W_V x + 2(\boldsymbol{\beta}^\top h) W_V h\big].\end{aligned}$$

By convexity, every centroid and its connecting segments to cell members lie in the radius-R ball. On this ball,

$$\begin{aligned}|\mathrm{D}^2 g(x)[h,h]| &\le e^{aR} a^2 \|h\|_2^2, \\ \|\mathrm{D}^2 F(x)[h,h]\|_2 &\le e^{aR} v(a^2 R + 2a)\|h\|_2^2.\end{aligned}$$

For $i \in \mathcal{U}_m$, set $\delta_i = w_i - \mathbf{a}_m$. Taylor's integral formula gives, componentwise when $H = F$,

$$H(\mathbf{a}_m + \delta_i) = H(\mathbf{a}_m) + \mathrm{D}H(\mathbf{a}_m)[\delta_i] + \int_0^1 (1-t)\mathrm{D}^2 H(\mathbf{a}_m + t\delta_i)[\delta_i, \delta_i]\, dt,$$

for $H \in \{g, F\}$. Exact centroids imply $\sum_{i \in \mathcal{U}_m} \delta_i = 0$, cancelling the first-order terms within each cell. Write

$$Z_{\mathrm{f}} = \frac{1}{N}\sum_i g(w_i), \quad Z_{\mathrm{c}} = \sum_m \mu_m g(\mathbf{a}_m),$$
$$A_{\mathrm{f}} = \frac{1}{N}\sum_i F(w_i), \quad A_{\mathrm{c}} = \sum_m \mu_m F(\mathbf{a}_m).$$

Summing the remainders and using $\int_0^1 (1-t)\, dt = 1/2$ yields

$$|Z_{\mathrm{f}} - Z_{\mathrm{c}}| \le \tfrac{1}{2} e^{aR} a^2 D, \tag{E.9}$$
$$\|A_{\mathrm{f}} - A_{\mathrm{c}}\|_2 \le \tfrac{1}{2} e^{aR} v(a^2 R + 2a) D. \tag{E.10}$$

Since $Z_{\mathrm{f}} \ge e^{-aR}$ and the convex-combination readout satisfies $\|\mathbf{e}_{\mathrm{mem}}\|_2 \le vR$, the identity

$$\mathbf{e}_{\mathrm{full}} - \mathbf{e}_{\mathrm{mem}} = \frac{A_{\mathrm{f}} - A_{\mathrm{c}} + \mathbf{e}_{\mathrm{mem}}(Z_{\mathrm{c}} - Z_{\mathrm{f}})}{Z_{\mathrm{f}}}$$

gives

$$\begin{aligned}\|\mathbf{e}_{\mathrm{full}} - \mathbf{e}_{\mathrm{mem}}\|_2 &\le e^{aR}\left[\tfrac{1}{2} e^{aR} v(a^2 R + 2a) D + vR\, \tfrac{1}{2} e^{aR} a^2 D\right] \\ &= e^{2aR} v\, a(1 + aR) D.\end{aligned}$$

Connection to memory construction. Retained key-frame tokens are singleton cells and contribute zero distortion. Hence D equals the context clustering objective in Eq. (A.3) divided by N. Proposition E.3 connects this objective to readout approximation: its upper bound scales linearly with D, with the coefficient determined by the query, descriptor, and projection scales. The same proof applies to the exact centroids of the partition returned by K-means. Multiplying the bound by $\|W_O\|_{\mathrm{op}}$ gives the corresponding error bound for the injected evidence $W_O\mathbf{e}$.

F Likelihood and stopping in latent trajectories

We derive the action likelihood used in RL and the corresponding stopping law. Throughout, $x = (V, q)$ and the parameter-independent randomness ξ used for memory construction are fixed. The frozen visual features, discrete prefix, and ξ determine the memory memberships. Policy replay retains these memberships and discrete actions, and recomputes trainable projections, registers, hidden states, and latent inputs under each policy's own parameters.

F.1 Discrete-action likelihood and policy replay

Let $\mathbf{a} = (a_1, \ldots, a_N)$ be the sampled decisions in causal order: categorical text tokens and binary end/continue events. Given $x, \xi, a_{<t}$, the continuous recurrence determines the history $\mathcal{H}_t^\omega$, yielding the conditional action law

$$q_t^\omega(a \mid a_{<t}) := \pi_\omega(a \mid \mathcal{H}_t^\omega(x, a_{<t}, \xi)). \tag{F.1}$$

For a complete finite trajectory, the probability chain rule gives

$$\begin{aligned}P_\omega(\mathbf{a} \mid x, \xi) &= \prod_{t=1}^N q_t^\omega(a_t \mid a_{<t}), \\ \nabla_\omega \log P_\omega(\mathbf{a} \mid x, \xi) &= \sum_{t=1}^N \nabla_\omega \log q_t^\omega(a_t \mid a_{<t}),\end{aligned} \tag{F.2}$$

where the gradient is evaluated at fixed actions, with positive realized probabilities and differentiable conditional policies. Continuous latent states enter through the history in each factor.

Writing the continuous prefix as $S_t(\omega)$ and its conditional log probability as $g_t(\omega, S_t(\omega))$, differentiation gives

$$\nabla_\omega \log q_t^\omega(a_t \mid a_{<t}) = \partial_\omega g_t + (D_\omega S_t)^\top \nabla_{S_t} g_t. \tag{F.3}$$

Thus answer gradients propagate through the full latent recurrence and memory reader, following the usual treatment of deterministic nodes in stochastic computation graphs (Schulman et al., 2015).

Policy ratios and KL. The GRPO surrogate uses the per-decision ratio $\varrho_t(\omega) = q_t^\omega(a_t \mid a_{<t})/q_t^{\omega_{\text{old}}}(a_t \mid a_{<t})$ (Shao et al., 2024). The current, old, and reference policies replay the same factual discrete prefix with factual memories, each using its own continuous states. The conditional KL in Eq. (A.12) is categorical for text decisions and Bernoulli for stopping decisions, with end probabilities p, p_0:

$$\mathrm{KL}(\mathrm{Ber}(p) \,\|\, \mathrm{Ber}(p_0)) = p \log \frac{p}{p_0} + (1-p) \log \frac{1-p}{1-p_0}. \tag{F.4}$$

F.2 Stopping and prescribed segment lengths

The truncated RL stopping law. For an RL rollout, condition on the prefix entering segment j. Define $p_k = p_{j,k}^{\text{end}}$ along the prefix that has continued through all earlier steps, and let $S_k = \prod_{r=1}^{k-1}(1-p_r)$, with $S_1 = 1$. These are the survival probabilities for reaching latent step k.

Under the sampled stopping protocol used in RL, the conditional segment length has distribution

$$\Pr_\omega(K_j = k \mid \text{entry prefix}) = \begin{cases} S_k p_k, & 1 \le k < K_{\max}, \\ S_{K_{\max}}, & k = K_{\max}. \end{cases} \tag{F.5}$$

An unforced stop at $k < K_{\max}$ requires $k-1$ continue actions followed by one end action, giving $S_k p_k$. Reaching the cap requires only the preceding continue actions; closure then occurs with probability one. Since $S_k p_k = S_k - S_{k+1}$, normalization follows by telescoping:

$$\sum_{k=1}^{K_{\max}-1} S_k p_k + S_{K_{\max}} = \sum_{k=1}^{K_{\max}-1} (S_k - S_{k+1}) + S_{K_{\max}} = 1.$$

The end probability is the vocabulary-softmax probability of <vis_end>. All other vocabulary outcomes share the same continuous update and are grouped into continue, with probability $1-p^{\text{end}}$. A sampled end contributes one binary decision to the likelihood; the appended <vis_end> embedding and forced closure contribute no additional probability or KL term. At inference, we choose greedily between the same end/continue probabilities, ending the segment when $p^{\text{end}} \ge 1 - p^{\text{end}}$, equivalently $p^{\text{end}} \ge \frac{1}{2}$, with ties resolved in favor of ending. The same cap $K_{\max} = 64$ applies.

Prescribed-length SFT objective. Each SFT example supplies $K_j^\star$ through Eq. (5), with $16 \le K_j^\star \le K_{\max} = 64$. The recurrence runs for exactly this many steps and the diversity term uses the same length. With lengths externally prescribed, the cross-entropy in Eq. (7) supervises $\mathcal{C} \cup \mathcal{A}$, including <vis_end> once at the final latent step. Intermediate continue decisions are omitted from this SFT objective. Lengths and teacher-forced actions stay fixed during differentiation, while gradients pass through the full latent recurrence and memory reader.

G LOCAL GRADIENT EQUIVALENCE OF CONTRASTIVE CREDIT

Fix one eligible factual trajectory and a well-formed frozen reference prefix. Its discrete actions, answer set $\mathcal{A}$, lengths N, L, task advantage, and reference score b^- remain fixed during the update, with $L = |\mathcal{A}| \geq 1$. Factual continuous states are recomputed under ω using the replay protocol in Appendix F. For $\tau_{\mathrm{cf}} > 0$, define

$$\begin{aligned} s(\omega) &= \frac{1}{L} \sum_{t \in \mathcal{A}} \log q_t^{\omega}(a_t \mid a_{<t}), \\ \ell^{\mathrm{cf}}(\omega) &= \log\left(1 + \exp \frac{b^- - s(\omega)}{\tau_{\mathrm{cf}}}\right). \end{aligned} \tag{G.1}$$

Thus $\ell^{\mathrm{cf}} = -\log p^{\mathrm{cf}}$, and direct differentiation gives the exact fixed-trajectory identity

$$-\nabla_\omega \ell^{\mathrm{cf}}(\omega) = \frac{1 - p^{\mathrm{cf}}(\omega)}{\tau_{\mathrm{cf}}} \nabla_\omega s(\omega). \tag{G.2}$$

The coefficient decreases as factual support increases relative to the frozen reference, assigning more credit to answers with a smaller factual–reference score gap.

Proposition G.1 (Extra credit equals contrastive descent locally)**.** *For the fixed trajectory above, assume the conditional policies are differentiable at ω_{old}, their realized old-policy probabilities are positive, and $\epsilon > 0$. Let $w = (1 - p^{\mathrm{cf}}(\omega_{\mathrm{old}}))/\tau_{\mathrm{cf}}$ be held fixed, and define two fixed advantage arrays by*

$$A_t^{\mathrm{base}} = A^{\mathrm{task}}, \qquad A_t^{\mathrm{aug}} = A^{\mathrm{task}} + \lambda_{\mathrm{cf}} \frac{N}{L} w \, \mathbf{1}\{t \in \mathcal{A}\}. \tag{G.3}$$

For either array, let the per-trajectory clipped ascent surrogate be

$$J(A; \omega) = \frac{1}{N} \sum_{t=1}^{N} \min\{\varrho_t(\omega) A_t, \mathrm{clip}(\varrho_t(\omega), 1 - \epsilon, 1 + \epsilon) A_t\}. \tag{G.4}$$

Then

$$\nabla_\omega [J(A^{\mathrm{aug}}; \omega) - J(A^{\mathrm{base}}; \omega)]\big|_{\omega_{\mathrm{old}}} = -\lambda_{\mathrm{cf}} \, \nabla_\omega \ell^{\mathrm{cf}}(\omega)\big|_{\omega_{\mathrm{old}}}. \tag{G.5}$$

The equality also holds when both surrogates include the same reference KL regularizer.

Proof. Because each policy replays the same discrete trajectory with its own parameters, $\varrho_t(\omega_{\mathrm{old}}) = 1$. By continuity and the finite number of decisions, there is a neighborhood of ω_{old} in which every ratio is strictly between $1 - \epsilon$ and $1 + \epsilon$. There clipping is the identity and the two arguments of the minimum coincide, irrespective of the sign of either advantage. Holding the denominator fixed gives

$$\nabla_\omega \varrho_t(\omega)|_{\omega_{\mathrm{old}}} = \nabla_\omega \log q_t^{\omega}(a_t \mid a_{<t})|_{\omega_{\mathrm{old}}}. \tag{G.6}$$

Consequently, with all derivatives below evaluated at ω_{old},

$$\begin{aligned} \nabla_\omega (J_{\mathrm{aug}} - J_{\mathrm{base}}) &= \frac{1}{N} \sum_{t \in \mathcal{A}} \lambda_{\mathrm{cf}} \frac{N}{L} w \, \nabla_\omega \log q_t^{\omega}(a_t \mid a_{<t}) \\ &= \lambda_{\mathrm{cf}} w \, \nabla_\omega \left[\frac{1}{L} \sum_{t \in \mathcal{A}} \log q_t^{\omega}(a_t \mid a_{<t})\right] = -\lambda_{\mathrm{cf}} \nabla_\omega \ell^{\mathrm{cf}}. \end{aligned} \tag{G.7}$$

The factor N/L converts the surrogate's per-trajectory average into the answer score's per-answer average. The last equality is Eq. (G.2) at the old parameters. Identical KL terms cancel when subtracting the surrogates.

For an ineligible trajectory $w = 0$, so its extra update is zero. Linearity extends the proposition to the batch and group averages in Eq. (11). Because that equation is a minimization loss, its added loss gradient at ω_{old} is $+\lambda_{\mathrm{cf}} \nabla_\omega \ell^{\mathrm{cf}}$, averaged over eligible samples using the original batch denominator.

This identity describes the update at ω_{old}. Subsequent GRPO steps use the frozen weights together with updated policy ratios and clipping, as specified in Eq. (11).