STORM: Internalized Modeling for Spatial-Temporal Reasoning in Video-Language Models

Anonymous authors
Paper under double-blind review

ABSTRACT

Many video reasoning tasks require tracking motion, temporal order, and evolving visual states across frames. Existing methods built on large vision–language models (LVLMs) often address this challenge by externalizing reasoning through textual chain-of-thought (CoT), keyframe selection, repeated frame reinsertion, or external tool use. While effective, such pipelines increase inference-time latency and engineering complexity, and they force temporal-visual evidence to be serialized into text or repeatedly re-encoded from frames. Inspired by the intuition that visual reasoning can occur implicitly before verbalization, we propose STORM (**S**patial-**T**emporal reas**O**ning via inte**R**nalized **M**odeling), a two-stage framework that teaches LVLMs to reason through bounded continuous latent trajectories instead of explicit textual CoT. In Stage I, STORM aligns latent tokens with thought-video representations derived from generated videos, grounding the latent states in dynamic visual evidence. In Stage II, the model is further trained with answer-only supervision, encouraging the reasoning process to be internalized without step-by-step annotations. Generated thought videos are used only during training; at inference, STORM performs a bounded latent rollout without regenerating videos, reinserting frames, or invoking external visual tools. Experiments on VideoMME, MVBench, TempCompass, and MMVU show that STORM improves video reasoning accuracy while substantially reducing inference overhead compared with tool or video-generation-based reasoning pipelines.

1 INTRODUCTION

Large vision–language models (LVLMs) have achieved strong performance on image understanding by aligning visual representations with large language models (LLMs) (Bai et al., 2025; Li et al., 2024a). However, extending this success to video reasoning remains challenging. Unlike static images, video understanding requires tracking motion, temporal order, and evolving scene states across frames. Although textual chain-of-thought (CoT) reasoning can help structure the reasoning process, it often struggles to faithfully represent fine-grained temporal and visual evidence. Video reasoning inherently involves multiple interacting factors, such as object dynamics, spatial relationships, and scene transitions, that evolve simultaneously over time, whereas textual rationales must linearize these entangled signals into a discrete token sequence.

To address this limitation, existing approaches often externalize reasoning by repeatedly revisiting visual inputs through frame reinsertion, temporal decomposition, or external tool-based pipelines (Arnab et al., 2025; Li et al., 2025b; Wang et al., 2025d; Zhou et al., 2025). These methods improve reasoning performance by exposing intermediate visual evidence during inference, but they also introduce substantial latency, engineering complexity, and sensitivity to retrieval or tool failures. This reveals a fundamental trade-off in video reasoning: strong reasoning performance benefits from preserving rich intermediate visual-temporal information, yet explicitly generating, retrieving, or reprocessing such evidence at inference time can significantly reduce efficiency and robustness.

These limitations motivate reasoning directly in continuous latent space, where hidden representations can compactly preserve visual-temporal information without repeatedly materializing it into explicit textual or visual intermediates. Unlike discrete textual reasoning, latent representations naturally support superposed and distributed features, allowing multiple reasoning factors to coexist and interact within the same hidden states (Zhu et al., 2026). Building on this insight, for complex video reasoning,

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.

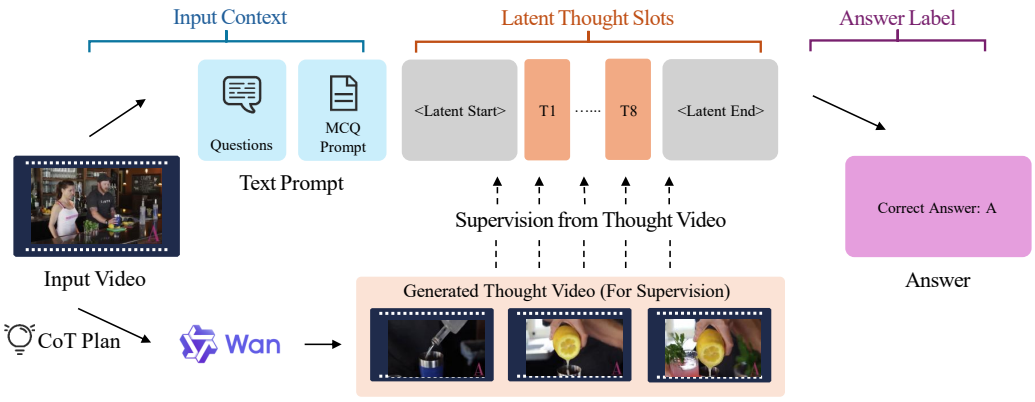

Figure 1: Interleaved STORM training sequence. By introducing latent tokens into autoregressive generation and applying thought-video supervision, STORM performs implicit reasoning in latent space and generates the final answer at the end of the sequence.

we ask whether intermediate visual evidence can be used during training to induce an internal latent reasoning process, while keeping inference as a single bounded decoding procedure. Rather than explicitly revisiting visual evidence at test time, the model learns to internalize temporal-visual reasoning within hidden states.

In this paper, we introduce STORM, a two-stage framework that teaches a video-language model to reason through continuous latent rather than explicit textual CoT or tools. The key idea is to expose the model to rich dynamic evidence during training, but to require it to use only compact hidden-state rollouts at inference time. Specifically, as shown in Figure 1, each training instance contains three aligned signals: (1) sampled keyframes from the source video, (2) a generated thought video that visualizes question-relevant motion and state changes, and (3) the target answer sequence. We first construct thought-video supervision by combining selected keyframes with a teacher reasoning plan, producing a short dynamic trace that externalizes the visual evidence needed for the answer. We then insert a small number of latent tokens between the video-question context and the final answer, and supervise these latent states to encode thought-video features before answer generation.

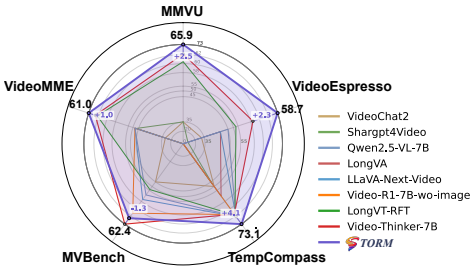

Figure 2: Comparison between STORM and representative video reasoning methods across benchmarks.

The training procedure is organized in two stages. In **Stage I**, we jointly optimize answer-token loss and latent-state alignment to thought-video representations, grounding the latent slots in temporally relevant visual evidence without requiring textual CoT supervision. In **Stage II**, we adopt a Coconut-like setting (Hao et al., 2026): the latent slots remain in the sequence, but supervision is applied only to the final answer. This encourages the model to consolidate the grounded latent behavior into an answer-focused internal reasoning process. At inference time, STORM does not generate thought videos, revisit frames, or call external visual tools. Once the model enters the latent segment, it performs a bounded continuous rollout through hidden states and then resumes normal text decoding for the final answer. In other words, dynamic visual evidence is used only to teach the latent states during training; at inference, the model reuses this learned behavior through a compact, fixed-budget latent trajectory without generating new videos via heavy external tools.

This design differs from recent "thinking with images" and latent-reasoning methods (Xu et al., 2026; Wu et al., 2024; Fu et al., 2024; Ray et al., 2025; Yang et al., 2025b; Li et al., 2025a). Prior approaches are often designed for static images, continuous text-space reasoning, explicit visual generation, or inference-time tool use. In contrast, STORM targets video reasoning directly by using generated videos only as training-time supervision and by internalizing spatial-temporal reasoning

into bounded latent states. Empirically, Figure 2 and our benchmark results show that this formulation improves video reasoning accuracy while substantially reducing the inference overhead associated with external reasoning pipelines. In summary, our main contributions are:

- We introduce ***STORM***, a video reasoning framework that performs internal dynamic simulation in continuous latent space instead of relying on explicit textual CoT or tool-heavy inference.
- We propose a two-stage training strategy: Stage I aligns latent tokens with thought-video supervision plus answer-token loss, and Stage II uses answer-only supervision to strengthen internalized reasoning.
- We build a thought-video supervision pipeline that aligns generated dynamic traces, keyframes, and QA annotations, enabling scalable training for latent video reasoning.
- We show that ***STORM*** improves video reasoning performance across benchmarks while reducing dependence on computationally expensive and operationally cumbersome test-time pipelines.

2 Related Work

2.1 Latent Visual and Implicit Reasoning

Recent work explores replacing explicit reasoning traces with compact continuous or latent representations to improve efficiency. Latent-visual methods augment VLM decoding with internal visual tokens or reconstruct task-relevant visual features, enabling multi-step visual inference without generating full-resolution images (Yang et al., 2025b; Li et al., 2025a). Related multimodal latent-thinking methods further introduce modality-agnostic latent tokens or active visual imagination for spatial reasoning (Ray et al., 2026; Cao et al., 2025). In parallel, continuous-thought methods train language models to reason by feeding back hidden states or distilling textual CoT into shorter latent trajectories that preserve reasoning ability while reducing token overhead (Hao et al., 2026; Deng et al., 2024; Shen et al., 2025b), while perception-token methods encode structured visual cues as dedicated latent inputs for fine-grained visual reasoning (Bigverdi et al., 2024; Yang et al., 2025a). However, these methods are mostly designed for text reasoning, static images, or spatial imagination, and do not explicitly address long-horizon video reasoning with unified, temporally grounded, task-aware latent representations.

2.2 Video–Language Models

Recent video LVLMs extend image-based multimodal models to temporal understanding by improving video representation, video instruction tuning, and token-efficient temporal modeling. Generalist models such as Qwen2.5-VL (Bai et al., 2025) and LLaVA-OneVision (Li et al., 2024a) unify image, multi-image, and video inputs to support broad multimodal instruction following. Video-specialized models such as LLaVA-Video (Zhang et al., 2025d) and VideoLLaMA3 (Zhang et al., 2025a) further strengthen video understanding through large-scale video instruction tuning and vision-centric multimodal training. In parallel, InternVL3 (Zhu et al., 2025), InternVL3.5 (Wang et al., 2025b), SlowFast-LLaVA-1.5 (Xu et al., 2025), and Eagle 2.5 (Chen et al., 2025) explore stronger multimodal pretraining, long-context post-training, and token-efficient long-form video comprehension. These advances improve how videos are encoded and consumed by LVLMs, but they typically still perform reasoning over sampled or compressed visual tokens at inference time. In contrast, ***STORM*** focuses on internalizing dynamic spatial-temporal reasoning into bounded latent trajectories.

2.3 Test-time Visual Reasoning

A line of work improves multimodal reasoning by making intermediate visual evidence explicit during reasoning. At the image or multi-image level, methods such as CMMCoT and ReFocus use multimodal CoT, memory augmentation, or visual editing to support cross-image comparison and structured image understanding (Zhang et al., 2025b; Fu et al., 2025b). For video understanding, frame-aware reasoning methods construct or train on reasoning traces that explicitly identify question-relevant frames, while temporal CoT methods iteratively select useful video context at inference time (Ghazanfari et al., 2026; Arnab et al., 2025). In parallel, tool-driven and active-perception frameworks call vision experts, zooming modules, selective revisitation mechanisms, or world

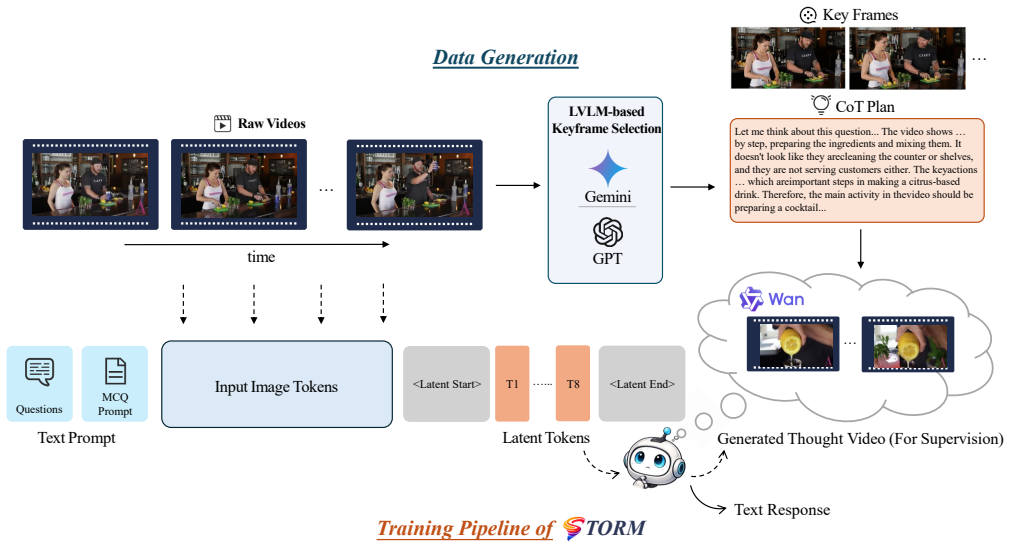

Figure 3: System workflow of the refined **STORM** framework, including keyframe input, CoT planning, thought-video generation, latent simulation, and final answer prediction.

models to gather additional visual evidence during reasoning (Yang et al., 2023; Shen et al., 2025a; Chung et al., 2026; Zhang et al., 2025c; Yu et al., 2026). Related training-time prompting methods such as VisionCoach show that visual guidance can be internalized without requiring prompts at inference (Lee et al., 2026). In contrast, **STORM** uses generated thought videos as training-time supervision for bounded latent trajectories, avoiding repeated frame selection, visual editing, and external tool calls at inference while still capturing temporal structure and critical visual cues.

3 STORM

We introduce **STORM**, a two-stage training framework that teaches an LVLM to reason through internal latent trajectories instead of explicit textual CoT. **STORM** uses three aligned signals for each QA pair: (1) sampled keyframes from the source video, (2) a generated thought video that externalizes dynamic reasoning evidence, and (3) the target answer token sequence.

3.1 Overview

Figure 3 summarizes the overall **STORM** system workflow used in our method, given a video $\mathcal{V} = \{f_1, \ldots, f_T\}$ and question token sequence $\mathcal{Q} = \{q_1, \ldots, q_J\}$, the model predicts answer tokens $\mathcal{Y} = \{y_1, \ldots, y_N\}$:

$$P_\theta(\mathcal{Y} \mid \mathcal{V}, \mathcal{Q}) = \prod_{t=1}^{N} P_\theta(y_t \mid y_{<t}, \mathcal{V}, \mathcal{Q}). \tag{1}$$

STORM augments this process with a short latent simulation segment containing K latent tokens (we use $K = 8$). During this segment, the decoder hidden states are treated as simulation states rather than decoded text. Let $\mathbf{z}_i$ be the hidden state at latent position i. Answer generation is then conditioned on both the original video-question context and the latent simulation states:

$$P_\theta(\mathcal{Y} \mid \mathcal{V}, \mathcal{Q}, \mathcal{Z}) = \prod_{t=1}^{N} P_\theta(y_t \mid y_{<t}, \mathcal{V}, \mathcal{Q}, \mathcal{Z}), \quad \mathcal{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_K\}. \tag{2}$$

Intuitively, $\mathcal{Z}$ acts as an internal "thought video" summary: it stores dynamic visual evidence in compact latent form before answer emission.

During inference, the model follows the same latent-slot pattern autoregressively. Once the special token `<|latent_start|>` is generated, the decoder enters a latent simulation mode. In this

mode, subsequent discrete token predictions are replaced by <|latent_pad|> placeholders in the token sequence, while the corresponding hidden states are fed back as continuous input embeddings for the next decoding step. The model remains in this mode until it generates <|latent_end|>, after which decoding resumes in the normal text space for final answer generation. To bound the computation, we impose a maximum latent budget of K slots, with $K = 8$ in all main experiments. If the model exceeds this budget without producing <|latent_end|>, we force the latent segment to terminate by inserting <|latent_end|>, ensuring a fixed upper simulation length and stable inference latency.

3.2 Data Pipeline

As illustrated in Figure 3, starting from video-QA samples, we construct interleaved training sequences for latent simulation supervision.

Step 1: Keyframe extraction. For each sample, we use GPT-4o (OpenAI et al., 2024) to extract question-relevant keyframes from the source video and denote them as $\mathcal{K} = \{k_1, \ldots, k_M\}$, where M is the number of selected keyframes.

Step 2: Thought-video construction. We denote the teacher-generated CoT sequence as $\mathcal{C}^{\text{teach}} = \{c_1, \ldots, c_L\}$, where L is the CoT length. Conditioned on the first keyframe $k_1 \in \mathcal{K}$ and teacher-generated CoT $\mathcal{C}^{\text{teach}}$, we generate thought videos using Wan 2.1 (Wan et al., 2025) with a conditional diffusion process.

Step 3: Token alignment. As shown in figure 1, we align three spans in one sequence: keyframe visual tokens, latent simulation slots, and answer tokens. This alignment yields direct supervision for latent simulation (on latent spans) and answer generation (on text spans), while avoiding explicit textual CoT supervision.

3.3 Two-Stage Training

STORM uses two training stages to first ground latent simulation, then internalize reasoning behavior.

Stage I: Latent-grounded supervised training. In Stage I, latent slots are supervised to match thought-video features, while answer tokens are trained with standard language modeling loss.

Let $\mathbf{z}_i$ denote the predicted latent state at slot i, and let $\mathbf{g}_i$ denote the corresponding thought-video target feature. For each training sample, we first encode the generated thought video with the Qwen visual encoder and obtain frame-level visual tokens. We concatenate the tokens from the sampled thought-video frames along the token dimension, producing a sequence $\mathbf{H} \in \mathbb{R}^{M \times d}$, where M is the total number of visual tokens and d is the hidden dimension. Since M may vary across samples, we apply one-dimensional adaptive average pooling over the token dimension to compress $\mathbf{H}$ into a fixed-length sequence of K target vectors, $\mathbf{G} = \{\mathbf{g}_i\}_{i=1}^{K} \in \mathbb{R}^{K \times d}$. Each $\mathbf{g}_i$ is therefore the average representation of one adaptive segment of the thought-video token sequence, providing compact temporal-visual supervision for the corresponding latent slot $\mathbf{z}_i$. We then use an MSE objective to align each predicted latent state $\mathbf{z}_i$ with its pooled target $\mathbf{g}_i$. Stage I losses are:

$$\mathcal{L}_{\text{latent}} = \frac{1}{K} \sum_{i=1}^{K} \|\mathbf{z}_i - \mathbf{g}_i\|_2^2, \qquad \mathcal{L}_{\text{ans}} = -\sum_{t=1}^{N} \log P_\theta(y_t \mid y_{<t}, \mathcal{V}, \mathcal{Q}, \mathcal{Z}). \tag{3}$$

The Stage I objective is

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{ans}} + \lambda \mathcal{L}_{\text{latent}}, \tag{4}$$

where λ controls latent alignment strength. In other words, Eq. 3 trains the model to compress the pooled visual dynamics of the thought video into a short latent simulation trajectory, while the answer-token loss ensures that these simulated states remain useful for downstream prediction.

Stage II: Answer-only consolidation. In Stage II, we remove direct latent supervision and optimize only $\mathcal{L}_{\text{ans}}$. Latent slots are still present in the input pattern, but the model receives no explicit loss on latent states. This forces latent reasoning to become self-organized and robust, while preserving the answer-focused objective used at inference.

Table 1: General video benchmark results. The orange/blue bars indicate think-with-tools (LongVT) and think-with-video (Video-Thinker) with 32 max frames for a single round. Except for those two methods, all experiments use uniform sampling with 32 frames.

Model	VideoMME	MVBench	TempCompass	Frames
Proprietary Models				
GPT-4o (OpenAI et al., 2024)	71.9	-	73.75	32
Gemini-1.5-Pro (Team et al., 2024)	75.0	-	67.1	32
Open-Source Models				
VideoChat2 (Li et al., 2024b)	39.5	51.1	48.8	32
LLaMA-VID (Li et al., 2023)	-	41.9	45.6	32
VideoLLaVA (Lin et al., 2024)	-	43.0	49.77	32
ST-LLM (Liu et al., 2024b)	37.9	54.85	-	32
ShareGPT4Video (Chen et al., 2024)	39.9	51.2	-	32
LLaVA-Next-Video (Zhang et al., 2024b)	60.2	-	66.74	32
PLLaVA (Xu et al., 2024)	-	46.6	-	32
LongVA (Zhang et al., 2024a)	52.6	-	56.9	32
VideoLLaMA2 (Cheng et al., 2024)	47.9	54.6	-	32
Video-CCAM (Fei et al., 2024)	-	**64.6**	-	32
Long-LLaVA (Wang et al., 2025c)	52.9	-	-	32
Kangaroo (Liu et al., 2024a)	56.0	61.1	62.5	32
LongVT-RFT (LongVT, think with tools) (Yang et al., 2026)	59.7	53.9	68.3	32
Video-Thinker-7B (think with video) (Wang et al., 2025a)	60.0	62.4	67.8	32
Qwen2.5-VL-7B-Instruct(CoT) (Bai et al., 2025)	60.8	42.4	72.8	32
Qwen2.5-VL-7B-SFT	55.4	60.5	69.9	32
Video-R1-7B-wo-image (Feng et al., 2025)	53.8	60.9	69.8	32
Video-R1-7B (Feng et al., 2025)	57.4	62.7	72.6	32
STORM	**61.0**	61.1	**74.3**	32

Table 2: Results on reasoning-heavy video benchmarks. The orange/blue bars indicate think-with-tools (LongVT) and think-with-video (Video-Thinker) with 32 max frames for a single round. Except for those two methods, all experiments use uniform sampling with 32 frames.

Model	VideoEspresso	Video-Holmes	MMVU	Frames
Proprietary Models				
GPT-4o (OpenAI et al., 2024)	-	42.0	75.4	32
Gemini-1.5-Pro (Team et al., 2024)	44.2	45.7	-	32
Open-Source Models				
LongVT-RFT (LongVT, think with tools) (Yang et al., 2026)	52.7	35.5	61.4	32
Video-Thinker-7B (think with video) (Wang et al., 2025a)	56.4	43.1	63.4	32
Qwen2.5-VL-7B-SFT	49.6	31.4	62.1	32
Qwen2.5-VL-7B-Instruct(CoT) (Bai et al., 2025)	45	27.8	60.0	32
Video-R1-7B (Feng et al., 2025)	-	36.5	64.2	32
STORM	**58.7**	**37.8**	**65.9**	32

Overall, this two-stage design trains the model to first learn meaningful thought-video-aligned latent dynamics, then rely on those internal dynamics without handcrafted latent targets.

4 EXPERIMENT

This section provides a complete empirical evaluation of ***STORM***. We first summarize the benchmark suite used to assess general video understanding and reasoning-heavy video QA, then describe the training data, backbone, optimization setup, latent-token budget, and hardware configuration. We next report the main results across general and reasoning-heavy benchmarks, including a smaller-backbone scalability study and an inference-latency comparison against tool- and video-generation-based reasoning pipelines. Finally, we analyze the design choices behind ***STORM*** through ablations on latent size and two-stage training, visualize latent-token alignment, and probe whether the learned latent states preserve video-specific information through same-video retrieval diagnostics.

Table 3: Results on a smaller backbone, showing that STORM remains effective across model sizes.

Model	VideoMME	TempCompass	MMVU
Qwen2.5-VL-3B-Instruct	50.9	55.0	42.0
Qwen2.5-VL-3B + STORM	**55.4**	**67.5**	**58.5**

4.1 Benchmarks

We evaluate our method on six benchmarks that cover both general video understanding and video reasoning: **VideoMME** (Fu et al., 2025a) contains 900 videos from six domains and approximately 2,700 multiple-choice questions over videos ranging from 11 seconds to 1 hour; **MVBench** (Li et al., 2024b) spans around 20 challenging tasks requiring temporal understanding and multimodal reasoning; **TempCompass** (Liu et al., 2024c) focuses on temporal perception through questions about speed, direction, order, and change; **Video-Holmes** (Cheng et al., 2025) comprises 1,837 questions from 270 suspense short films and emphasizes clue-based multi-step reasoning; **VideoEspresso** (Han et al., 2024) is a large-scale chain-of-thought-based VideoQA benchmark built with semantic core-frame selection across 14 task types; and **MMVU** (Zhao et al., 2025) evaluates expert-level multi-discipline video understanding, requiring models to combine detailed visual evidence with higher-level reasoning across diverse domains.

4.2 Experiment Setup

Implementation details. STORM is built upon Qwen2.5-VL-7B-Instruct and trained in two stages. In Stage I, we jointly optimize latent-state alignment and answer-token cross-entropy, with a balancing coefficient set to 0.1. In Stage II, we keep the latent simulation slots but remove direct latent supervision, and optimize only answer-token cross-entropy. Each stage is trained for one epoch with a learning rate of 1×10^{-5} and a batch size of 1. To mitigate instability from variable latent lengths, we set the latent token budget to 8. All experiments are conducted on four NVIDIA GB200 GPUs.

Datasets. For training, we use the video-based portion of the dataset introduced in (Feng et al., 2025), which focuses on everyday scenarios and temporal reasoning. We exclude image-only samples for efficiency. Following the data construction pipeline in Section 3.2, we extract question-relevant keyframes with GPT-4o and generate 5K short thought videos conditioned on the keyframes and teacher CoT prompts. The resulting keyframe, thought-video, and answer annotations are used for the two-stage training.

4.3 Results and Analysis

Performance on general video benchmarks. We evaluate STORM on a suite of general video benchmarks, including VideoMME, MVBench, and TempCompass (Table 1). Crucially, STORM consistently outperforms its standard supervised fine-tuned (SFT) counterpart, Qwen2.5-VL-7B-SFT, which is trained on the same dataset as STORM but without thought-video supervision. This consistent improvement across spatial-temporal tasks suggests that the latent simulation objective provides benefits beyond standard supervised fine-tuning. Beyond the SFT baseline, STORM also matches or exceeds specialized models like LongVT and Video-Thinker that rely on expensive test-time compute pipelines. This result shows that internalizing reasoning in latent space can match the efficacy of explicit external reasoning without incurring the same inference bottlenecks.

Generalizability to smaller backbones. To ensure these benefits are not tied to a single model scale, we also validate STORM on the smaller Qwen2.5-VL-3B backbone. As shown in Table 3, STORM consistently improves baseline performance across VideoMME, TempCompass, and MMVU, demonstrating the framework's generalizability.

Performance on reasoning-focused video benchmarks. Similar improvements emerge on reasoning-focused benchmarks, including VideoEspresso, Video-Holmes, and MMVU (Table 2). STORM consistently outperforms standard SFT baselines, proving that internalized visual reasoning effectively handles complex, multi-step logic. Even when evaluated against state-of-the-art models that rely on explicit video generation or complex external toolchains, STORM exhibits competitive

Table 4: Latency analysis on MMVU using provided runtime statistics. STORM latency is converted from case-study throughput (2.14 it/s), corresponding to about 0.47 s/item. Lower is better.

Model	Throughput	Latency (s/item)
LongVT	0.065 it/s	15.44
Video-Thinker	0.058 it/s	17.20
STORM	2.14 it/s	**0.47**

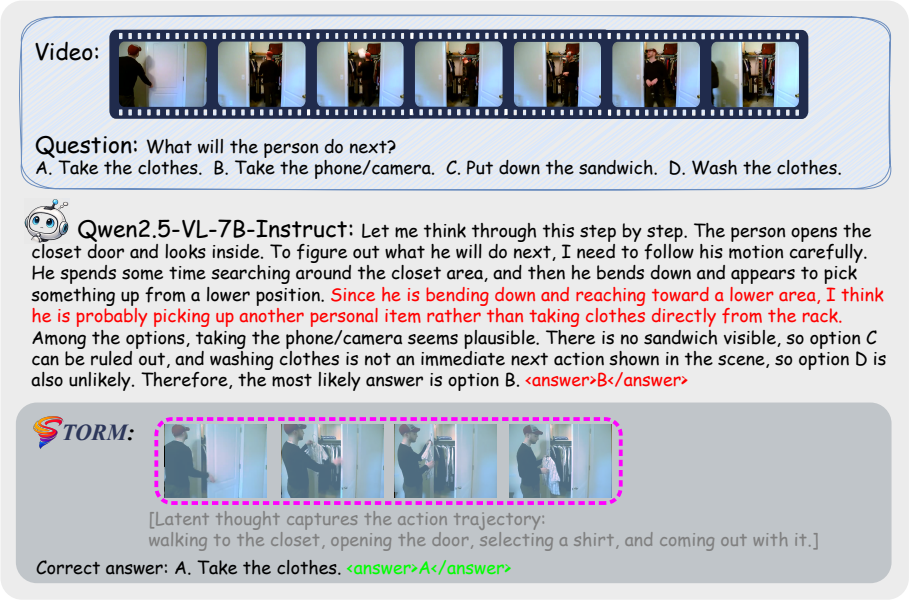

Figure 4: Case study comparing textual reasoning with STORM latent reasoning on future-action prediction. STORM captures the action trajectory in latent space and recovers the correct answer.

performance. Ultimately, implicit latent simulation proves to be a highly efficient alternative to resource-heavy pipelines, delivering deep video understanding without the operational bloat.

Latency analysis. We further compare runtime efficiency on MMVU in Table 4. LongVT and Video-Thinker require 15.44 and 17.20 s/item, corresponding to only 0.065 and 0.058 items/s, respectively. In contrast, STORM processes 2.14 items/s, or approximately 0.47 s/item. This efficiency gap suggests that internalizing the reasoning process makes STORM substantially more efficient than methods that rely on costly test-time tool calls or repeated frame reinsertion. During inference, STORM carries out intermediate reasoning through a compact set of latent tokens, keeping the additional reasoning cost small and predictable. With this training design, the expensive visual reasoning signal is introduced through thought-video supervision during training, while inference remains a compact single-model decoding process. This substantially reduces test-time overhead compared with methods that explicitly revisit visual inputs or invoke external reasoning modules.

4.4 Ablations

Effects of different latent sizes. To study how the latent-slot budget affects performance, we vary the number of latent tokens used to compress thought-video features. As shown in Table 5, very small latent budgets provide limited capacity for representing temporal-visual evidence, while overly large budgets can dilute the supervision signal and introduce redundant visual tokens. The best performance is achieved with $K = 8$, which provides a compact but sufficiently expressive latent trajectory. In contrast, increasing the latent size to 16 or removing adaptive pooling entirely does not

Table 5: Evaluation across latent-token sizes. A moderate budget of 8 latent tokens achieves the best overall balance across all three benchmarks, while larger budgets degrade performance, suggesting that compact latent trajectories provide more effective supervision than retaining redundant visual tokens.

Latent size	VideoMME	TempCompass	MVBench
4	58.9	71.1	55.0
8	**61.0**	**74.3**	**61.1**
16	52.2	67.3	55.5
256 (no average)	52.9	68.7	54.7

Table 6: Performance by two-stage training configuration. The full *STORM* pipeline consistently outperforms either stage alone, indicating that thought-video grounded latent alignment in Stage I and answer-focused consolidation in Stage II are complementary.

Setting	VideoMME	TempCompass	MVBench
Stage I only	58.8	72.0	56.0
Stage II only	56.6	72.6	56.8
Full *STORM* (Stage I + Stage II)	**61.0**	**74.3**	**61.1**

improve performance, suggesting that preserving more visual tokens is not necessarily beneficial for internalized reasoning. We therefore use $K = 8$ in all main experiments.

Effects of different training stages. Table 6 compares using only Stage I, using only Stage II, and the full two-stage *STORM* pipeline. Both single-stage variants underperform the full model: Stage I learns grounded latent dynamics from thought-video supervision but lacks large-scale answer alignment, while Stage II benefits from answer-token supervision but lacks latent-space grounding. The best performance therefore requires both thought-video-grounded latent supervision in Stage I and answer-focused consolidation in Stage II.

Latent token analysis. We sample 20 examples across three benchmark categories and visualize video-frame, keyframe, and latent-token embeddings after PCA and t-SNE projection. As shown in Figure 5, keyframe embeddings form sub-clusters within the broader video-frame distribution, and the sparse latent-token embeddings consistently fall inside or adjacent to the corresponding keyframe regions. This indicates that *STORM* learns informative latent representations aligned with true keyframe semantics, enabling more effective video reasoning.

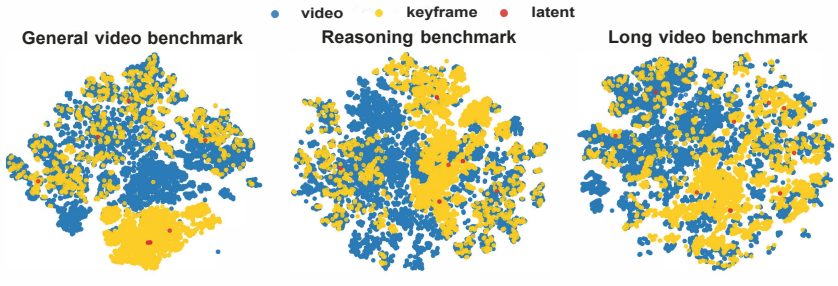

Figure 5: t-SNE plot of latent-token, keyframe, and video-frame embeddings.

5 Conclusion

In this paper, we present *STORM*, a framework for internalized spatial-temporal reasoning in video-language models. Instead of relying on explicit textual CoT, repeated frame reinsertion, or external visual tools, *STORM* captures key temporal-visual evidence through bounded latent trajectories. By using generated thought videos only during training, *STORM* shifts visual reasoning from inference-time tool orchestration to an internalized latent process, enabling strong benchmark performance while reducing test-time overhead.We hope this work provides a practical step toward efficient native visual reasoning in video-language models.

AI USE STATEMENT

Generative AI models were used as components of the data construction pipeline in this work. Specifically, GPT-4o was used to select question-relevant keyframes from source videos, and a teacher model was used to generate chain-of-thought plans for constructing the supervision data. Wan 2.1 was then used to generate synthetic thought videos conditioned on the selected keyframes and teacher-generated reasoning plans; these generated videos were used only as training-time supervision for latent-state alignment. The authors are responsible for the experimental design, analysis, and conclusions presented in the paper.

REPRODUCIBILITY STATEMENT

We provide the main implementation and experimental details required to reproduce our results throughout the paper. Section 3 describes the latent simulation mechanism, thought-video construction pipeline, and two-stage training objectives, while Section 4.2 specifies the model backbone, optimization settings, latent-token budget, training data construction, and hardware configuration. In particular, TORM is built on Qwen2.5-VL, uses a latent budget of eight tokens, and is trained for one epoch in each stage with a learning rate of 1×10^{-5} and batch size 1 on four NVIDIA GB200 GPUs. We additionally provide the keyframe-selection and thought-video-generation prompts in Appendix B and report evaluations across all benchmarks and ablation settings used in the main experiments.

REFERENCES

Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Temporal chain of thought: Long-video understanding by thinking in frames, 2025. URL https://arxiv.org/abs/2507.02001.

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.

Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G. Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models, 2024. URL https://arxiv.org/abs/2412.03548.

Meng Cao, Xingyu Li, Xue Liu, Ian Reid, and Xiaodan Liang. Spatialdreamer: Incentivizing spatial reasoning via active mental imagery, 2025. URL https://arxiv.org/abs/2512.07733.

Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Tuomas Rintamaki, Tyler Poon, Max Ehrlich, Tuomas Rintamaki, Tyler Poon, Tong Lu, Limin Wang, Bryan Catanzaro, Jan Kautz, Andrew Tao, Zhiding Yu, and Guilin Liu. Eagle 2.5: Boosting long-context post-training for frontier vision-language models, 2025. URL https://arxiv.org/abs/2504.15271.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions, 2024. URL https://arxiv.org/abs/2406.04325.

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.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms, 2024. URL https://arxiv.org/abs/2406.07476.

Jiwan Chung, Junhyeok Kim, Siyeol Kim, Jaeyoung Lee, Min Soo Kim, and Youngjae Yu. v1: Learning to point visual tokens for multimodal grounded reasoning, 2026. URL https://arxiv.org/abs/2505.18842.

Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step, 2024. URL https://arxiv.org/abs/2405.14838.

Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos, 2024. URL https://arxiv.org/abs/2408.14023.

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, 2025. URL https://arxiv.org/abs/2503.21776.

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, 2025a. URL https://arxiv.org/abs/2405.21075.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive, 2024. URL https://arxiv.org/abs/2404.12390.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding, 2025b. URL https://arxiv.org/abs/2501.05452.

Sara Ghazanfari, Francesco Croce, Nicolas Flammarion, Prashanth Krishnamurthy, Farshad Khorrami, and Siddharth Garg. Chain-of-frames: Advancing video understanding in multimodal llms via frame-aware reasoning, 2026. URL https://arxiv.org/abs/2506.00318.

Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection, 2024. URL https://arxiv.org/abs/2411.14794.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space, 2026. URL https://arxiv.org/abs/2412.06769.

Daeun Lee, Shoubin Yu, Yue Zhang, and Mohit Bansal. Visioncoach: Reinforcing grounded video reasoning via visual-perception prompting, 2026. URL https://arxiv.org/abs/2603.14659.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning, 2025a. URL https://arxiv.org/abs/2509.24251.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. URL https://arxiv.org/abs/2408.03326.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2024b. URL https://arxiv.org/abs/2311.17005.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models, 2023. URL https://arxiv.org/abs/2311.17043.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models, 2025b. URL https://arxiv.org/abs/2506.18883.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection, 2024. URL https://arxiv.org/abs/2311.10122.

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input, 2024a. URL https://arxiv.org/abs/2408.15542.

Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners, 2024b. URL https://arxiv.org/abs/2404.00308.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos?, 2024c. URL https://arxiv.org/abs/2403.00476.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben

Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. Sat: Dynamic spatial aptitude training for multimodal language models, 2025. URL https://arxiv.org/abs/2412.07755.

Arijit Ray, Ahmed Abdelkader, Chengzhi Mao, Bryan A. Plummer, Kate Saenko, Ranjay Krishna, Leonidas Guibas, and Wen-Sheng Chu. Mull-tokens: Modality-agnostic latent thinking, 2026. URL https://arxiv.org/abs/2512.10941.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration, 2025a. URL https://arxiv.org/abs/2411.16044.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation, 2025b. URL https://arxiv.org/abs/2502.21074.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire

Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix

de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer,

Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Shijian Wang, Jiarui Jin, Xingjian Wang, Linxin Song, Runhao Fu, Hecheng Wang, Zongyuan Ge, Yuan Lu, and Xuelian Cheng. Video-thinker: Sparking "thinking with videos" via reinforcement learning, 2025a. URL https://arxiv.org/abs/2510.23473.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025b. URL https://arxiv.org/abs/2508.18265.

Xidong Wang, Dingjie Song, Shunian Chen, Junyin Chen, Zhenyang Cai, Chen Zhang, Lichao Sun, and Benyou Wang. Longllava: Scaling multi-modal llms to 1000 images efficiently via a hybrid architecture, 2025c. URL https://arxiv.org/abs/2409.02889.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-r1: Post-training large vision language model for temporal video grounding, 2025d. URL https://arxiv.org/abs/2503.13377.

Qiucheng Wu, Handong Zhao, Michael Saxon, Trung Bui, William Yang Wang, Yang Zhang, and Shiyu Chang. Vsp: Assessing the dual challenges of perception and reasoning in spatial planning tasks for vlms, 2024. URL https://arxiv.org/abs/2407.01863.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava : Parameter-free llava extension from images to videos for video dense captioning, 2024. URL https://arxiv.org/abs/2404.16994.

Mingze Xu, Mingfei Gao, Shiyu Li, Jiasen Lu, Zhe Gan, Zhengfeng Lai, Meng Cao, Kai Kang, Yinfei Yang, and Afshin Dehghan. Slowfast-llava-1.5: A family of token-efficient video large language models for long-form video understanding, 2025. URL https://arxiv.org/abs/2503.18943.

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images, 2026. URL https://arxiv.org/abs/2505.11409.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning?, 2025a. URL https://arxiv.org/abs/2402.16837.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens, 2025b. URL https://arxiv.org/abs/2506.17218.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action, 2023. URL https://arxiv.org/abs/2303.11381.

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, 2026. URL https://arxiv.org/abs/2511.20785.

Shoubin Yu, Yue Zhang, Zun Wang, Jaehong Yoon, Huaxiu Yao, Mingyu Ding, and Mohit Bansal. When and how much to imagine: Adaptive test-time scaling with world models for visual spatial reasoning, 2026. URL https://arxiv.org/abs/2602.08236.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding, 2025a. URL https://arxiv.org/abs/2501.13106.

Guanghao Zhang, Tao Zhong, Yan Xia, Mushui Liu, Zhelun Yu, Haoyuan Li, Wanggui He, Fangxun Shu, Dong She, Yi Wang, and Hao Jiang. Cmmcot: Enhancing complex multi-image comprehension via multi-modal chain-of-thought and memory augmentation, 2025b. URL https://arxiv.org/abs/2503.05255.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision, 2024a. URL https://arxiv.org/abs/2406.16852.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. Adaptive chain-of-focus reasoning via dynamic visual search and zooming for efficient vlms, 2025c. URL https://arxiv.org/abs/2505.15436.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024b. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data, 2025d. URL https://arxiv.org/abs/2410.02713.

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.

Yiyang Zhou, Yangfan He, Yaofeng Su, Siwei Han, Joel Jang, Gedas Bertasius, Mohit Bansal, and Huaxiu Yao. Reagent-v: A reward-driven multi-agent framework for video understanding, 2025. URL https://arxiv.org/abs/2506.01300.

Hanlin Zhu, Shibo Hao, Zhiting Hu, Jiantao Jiao, Stuart Russell, and Yuandong Tian. Emergence of superposition: Unveiling the training dynamics of chain of continuous thought, 2026. URL https://arxiv.org/abs/2509.23365.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

A Additional Analysis of Latent Slot Representations

A.1 Same-video retrieval probe

To further examine whether the learned latent states encode video-specific temporal information, we conduct a same-video retrieval probe. For each query sample, we use its learned latent representation to retrieve other samples and measure whether the nearest retrieved examples originate from the same source video.

Figure 6 compares retrieval using latent representations, textual representations, and a random baseline. Latent representations achieve substantially stronger same-video retrieval performance, reaching 71% Hit@1, 79% Hit@5, and 75% MRR. In comparison, textual representations achieve 29% Hit@1, 71% Hit@5, and 50% MRR, while random retrieval gives 0% under all three metrics. This suggests that the learned latent states retain video-specific information more reliably than text-only representations, especially at the top retrieval rank. The large gap in Hit@1 indicates that latent states are more effective at identifying the most relevant same-video neighbor, rather than only recovering a same-video example somewhere in the top-ranked set.

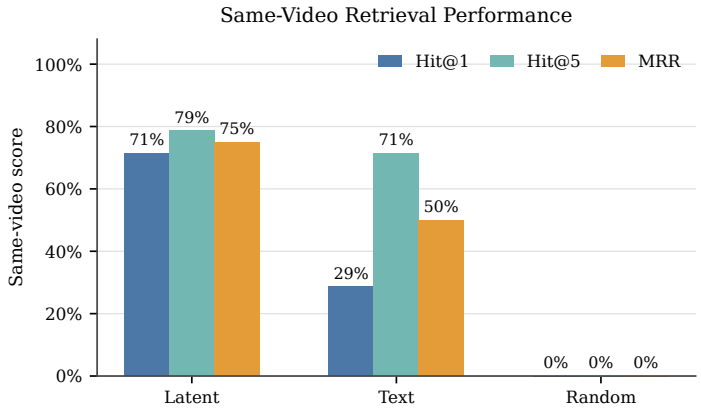

Figure 6: **Same-video retrieval performance.** We compare latent representations, textual representations, and random retrieval using Hit@1, Hit@5, and MRR. Latent representations achieve the strongest same-video retrieval performance, indicating that the latent states preserve video-specific information.

A.2 Per-query retrieval behavior

We additionally visualize the reciprocal rank of the first same-video match for each query in Figure 7. This per-query view shows that the improvement from latent representations is not caused by a single outlier. For several queries, latent representations retrieve a same-video match at rank 1, while text representations either fail or retrieve the match at a much lower rank. In particular, multiple examples achieve a reciprocal rank of 1.0 under the latent representation, corresponding to perfect top-1 same-video retrieval. Text representations occasionally retrieve same-video examples, but they more often assign them lower ranks. Random retrieval remains near zero across nearly all queries.

This result provides complementary evidence to the aggregate metrics in Figure 6. While the aggregate results show that latent representations improve overall retrieval quality, the per-query comparison shows that the gain is distributed across many samples. This supports the interpretation that *STORM* latent states encode temporally grounded video information rather than merely memorizing answer labels or exploiting a small number of easy examples.

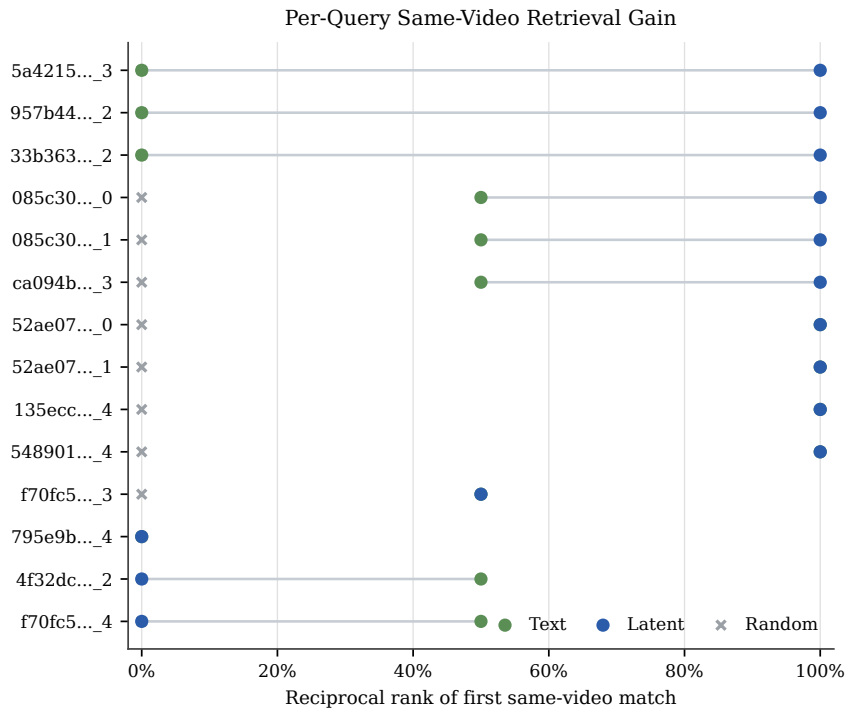

Figure 7: **Per-query same-video retrieval gain.** Each row corresponds to a query example, and the x-axis shows the reciprocal rank of the first retrieved same-video match. Latent representations more frequently retrieve same-video examples at high rank, especially at rank 1.

A.3 Latent slot ablation

We further analyze which latent slots contribute most to same-video retrieval. Figure 8 ranks different slot aggregation and ablation strategies according to same-video retrieval performance. Mean pooling over latent slots achieves the strongest performance among the tested variants, suggesting that video-specific information is distributed across multiple latent states rather than concentrated in a single slot. Dropping individual slots generally reduces performance only moderately, which indicates that the latent representation is relatively robust to the removal of any one slot.

Among single-slot variants, later slots such as Slot 7 and Slot 4 perform strongly, while earlier slots such as Slot 1 and Slot 2 are weaker. This trend suggests that later latent states may accumulate more complete temporal evidence after several rounds of latent rollout. However, no single slot consistently outperforms mean pooling, supporting the design choice of treating the latent rollout as a short trajectory rather than relying on a single hidden state. The reverse-order variant performs similarly to several slot-dropping variants, suggesting that the learned representation is not purely order-agnostic but also not catastrophically sensitive to slot ordering under this diagnostic.

Overall, the slot ablation supports two conclusions. First, the latent rollout contains meaningful video-level information that can be recovered through retrieval. Second, this information is distributed across the latent trajectory, with mean pooling providing a stable summary. These findings are consistent with the main design of *STORM*, where multiple latent slots jointly form a compact internal simulation before answer generation.

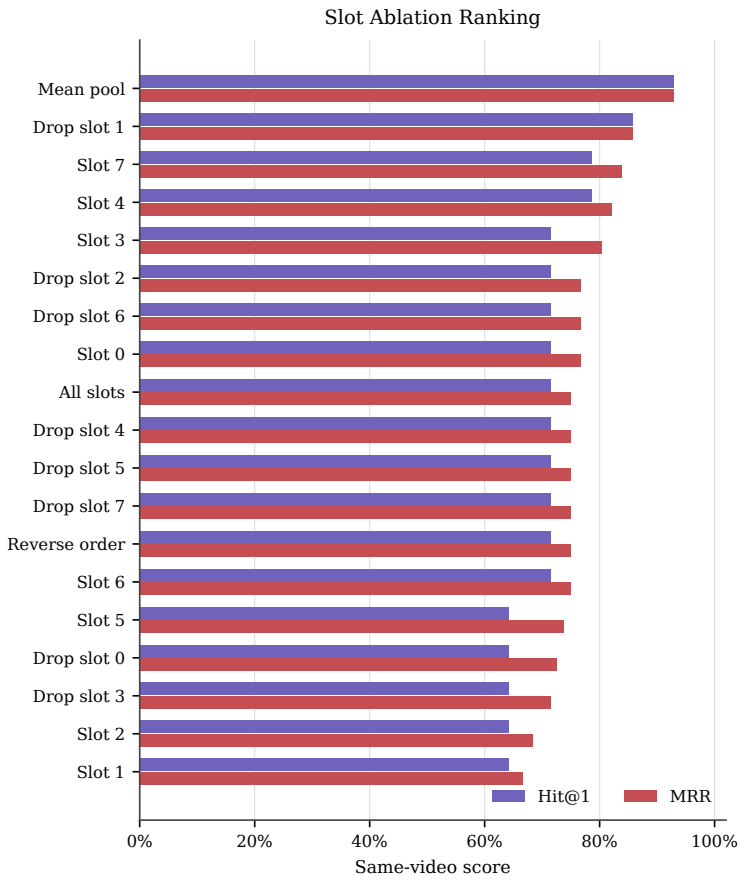

Figure 8: **Latent slot ablation ranking.** We compare same-video retrieval performance under different slot selection, dropping, reversal, and pooling strategies. Mean pooling performs best, indicating that useful video-specific information is distributed across the latent trajectory.

B Prompt Templates for Data Construction

This appendix provides the prompt templates used in our data construction pipeline. Appendix B.1 presents the prompt for selecting question-relevant keyframes from candidate frames, and Appendix B.2 presents the prompt for generating thought videos from the selected keyframe and teacher-generated reasoning plan.

B.1 Keyframe Selection Prompt

Figure 9 shows the prompt used to select question-relevant keyframes from candidate frames. Given a question, optional answer choices, the correct answer, and a list of candidate frames with timestamps, GPT-4o is instructed to return only the selected frame indices in JSON format.

Prompt Template for Keyframe Selection

```
You are given a video represented by candidate frames.
Task: pick EXACTLY {need_n} frames from the candidates below that help answer the question.
Return ONLY JSON with this schema: {"selected":[i0,i1,...]}
Rules:
- Indices must come from the provided candidate list; they must be unique integers.
- JSON only. No extra text.

Question: {question}

Choices:
- A. {choice_0}
- B. {choice_1}
- ...

Correct answer: {answer}

frame_index: {idx}, t={t:.2f}s
[image_url: candidate frame]
```

Figure 9: **Prompt template for keyframe selection.** The prompt asks GPT-4o to select exactly the required number of question-relevant frames from a candidate-frame list and return only the selected frame indices in JSON format.

B.2 Thought-Video Generation Prompt

Figure 10 shows the prompt used to generate thought videos. The video generator is conditioned on the selected keyframe, the question, and the teacher-generated reasoning plan. The generated thought video is used only as training-time supervision for latent-state alignment.

Prompt Template for Thought Video Generation

```
You are given a keyframe from a source video and a reasoning plan for answering a video
question. Generate a short thought video that visualizes the temporal evolution needed to
answer the question.

Requirements:
- Preserve the scene layout, object identity, appearance, and camera viewpoint from the
  keyframe as much as possible.
- Animate only the question-relevant motion, interaction, or state change described by the
  reasoning plan.
- Keep the temporal order physically plausible and visually coherent.
- Do not add irrelevant objects, subtitles, watermarks, text overlays, or extra scenes.
- The video should focus on the evidence needed for the final answer, rather than producing
  a generic continuation.

Question: {question}

Reasoning plan:
{teacher_cot}

Initial keyframe:
[image_url: selected keyframe]

Output:
A short coherent thought video conditioned on the keyframe and reasoning plan.
```

Figure 10: **Prompt template for thought-video generation.** The prompt conditions the video generator on the selected keyframe, the question, and the teacher-generated reasoning plan, encouraging a short coherent video that visualizes the temporal evidence needed to answer the question.

C LIMITATIONS AND FUTURE WORK

Although STORM provides an efficient framework for internalized spatial-temporal reasoning, several limitations remain. First, it relies on generated thought-video supervision during training, whose quality depends on keyframe extraction, teacher CoT prompts, and the video generation model; noisy keyframes or inaccurate generated motion may lead to imperfect latent targets. Second, our evaluation mainly focuses on multiple-choice video QA benchmarks and Qwen2.5-VL backbones, so broader validation on open-ended video reasoning, dense temporal grounding, embodied tasks, and other LVLM architectures remains future work. Finally, while STORM reduces inference-time cost, it introduces additional training-time cost for constructing thought-video supervision. Future work will explore stronger filtering, multi-keyframe conditioning, architecture transfer, and lighter supervision strategies such as keyframe-only targets or self-distilled latent targets.