VideoLatent: Video-Language Learning via Latent Self-Forcing

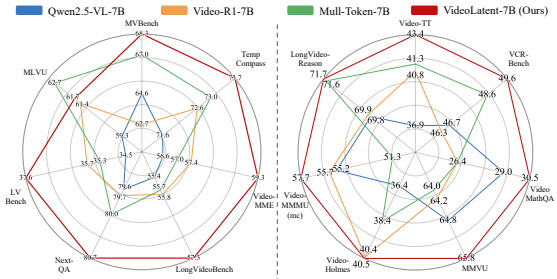
Anonymous ACL submission

Abstract

Recent advancements in chain-of-thought (CoT) reasoning have shown promise in enhancing video understanding and reasoning capabilities of multimodal large language models (MLLMs). However, existing CoT-based MLLMs require labor-intensive CoT annotations and incur substantial training and inference overhead. While visual latent reasoning has emerged as a more efficient alternative, existing methods primarily focus on image tasks and heavily rely on additional supervision signals for visual latent generation (*e.g.*, CoT traces, auxiliary images, or fine-grained annotations), limiting their scalability and transferability to video tasks. To bridge this gap, we introduce **VideoLatent**, a novel MLLM equipped with a latent injection module tailored for video understanding and reasoning. Specifically, VideoLatent learns to perform visual latent reasoning using a new latent self-forcing training paradigm, which comprises latent alignment and latent diversity objectives, and relies solely on standard video-question-answer triplets. Extensive experiments across 14 benchmarks demonstrate that our model consistently outperforms existing standard and latent MLLMs on general video understanding and complex video reasoning. Compared with Video-R1, our VideoLatent achieves superior computational efficiency, reducing training/inference overhead by $\sim$6$\times$/$\sim$68$\times$. Moreover, experiments demonstrate that our method has strong generalizability to different MLLM backbones and different model scales.

1 Introduction

Recently, large language models (LLMs) (Vaswani et al., 2017; Touvron et al., 2023; DeepSeek-AI, 2025; OpenAI, 2025; Team, 2025b) have substantially enhanced complex reasoning capability through chain-of-thought (CoT) reasoning (Wei et al., 2022). Building upon these advances, multimodal learning efforts have increasingly focused



(a) General Video Understanding (b) Complex Video Reasoning

Figure 1: Our VideoLatent-7B consistently outperforms existing standard and latent MLLMs across fourteen benchmarks under the same experimental settings (see Tab. 1 and Tab. 2 for more details), covering both general video understanding and complex video reasoning.

on equipping multimodal large language models (MLLMs) with CoT reasoning to enhance multimodal understanding and reasoning (OpenAI, 2024b; Huang et al., 2025; Li et al., 2025c; Xu et al., 2025a; Wang et al., 2026a; Zhan et al., 2026; Gao et al., 2025). Particularly in the video-language domain, numerous studies have been proposed to investigate the effectiveness of CoT reasoning in improving both general video understanding and complex video reasoning capabilities (Feng et al., 2025a; Wang et al., 2025a; Meng et al., 2025; Feng et al., 2025b; Fei et al., 2024; Ghazanfari et al., 2025; Liu et al., 2025b; Han et al., 2025).

Despite their improved performance and interpretability, existing CoT-based video reasoning approaches exhibit two critical limitations: (1) Their explicit reasoning traces may suffer from hallucination and overthinking (Liu et al., 2026a; Zhong et al., 2025), and may not consistently improve performance on both general video understanding and complex video reasoning benchmarks (as demonstrated in Tab. 1); and (2) Their training incurs substantial computational and memory overhead, and often requires labor-intensive CoT annotations (Feng et al., 2025a,b; Meng et al., 2025), limiting training scalability and inference efficiency.

To address these limitations, visual latent reasoning (Yang et al., 2025c; Li et al., 2025a; Ray et al.,

2025; Yu et al., 2026) has emerged as a promising paradigm to perform implicit reasoning beyond language (Hao et al., 2024; Xu et al., 2025b), enabling more efficient and flexible reasoning. Specifically, it alternates between language-space and latent-space reasoning, with the latter generating continuous latent thoughts rather than discrete textual ones. However, existing methods mainly focus on image understanding and reasoning (Yang et al., 2025c; Li et al., 2025a; Wang et al., 2025b), leaving visual latent reasoning for video tasks underexplored. Moreover, they largely rely on additional, image-centric supervision signals for visual latent generation (*e.g.*, CoT traces (Shen et al., 2025a), helper images (Yang et al., 2025c), pretrained vision foundation models (Qin et al., 2025), bounding boxes (Li et al., 2025a), and other fine-grained annotations (Wang et al., 2025b, 2026b)), limiting scalability and transferability to video reasoning.

In this paper, we propose **VideoLatent**, a novel MLLM framework focused on visual latent reasoning for video understanding and reasoning tasks, mitigating the limitations of both standard and latent MLLMs. As in Fig. 3, VideoLatent learns to perform visual latent reasoning through a new **latent injection module** and **latent self-forcing training paradigm**. Specifically, the latent injection module is designed to prevent the self-generated latent thoughts from drifting away from the video and question context. Latent self-forcing further enhances video-language learning via two complementary objectives (*i.e.*, latent alignment and latent diversity), while relying solely on video-question-answer triplets without requiring any additional supervision for visual latent generation.

To ensure a comprehensive evaluation, we conduct experiments on fourteen video-language benchmarks, covering both general video understanding and complex video reasoning. Our experiments yield several **key findings**: (1) VideoLatent consistently outperforms both standard and latent MLLMs across diverse benchmarks and experimental settings (*e.g.*, Tab. 1 and Tab. 2), demonstrating the **effectiveness** of our proposed method; (2) Our method exhibits strong **generalizability** to different MLLM backbones and different model sizes (*e.g.*, Tab. 3 and Tab. 4); (3) Our VideoLatent achieves superior computational **efficiency** compared with CoT-based MLLMs (*e.g.*, Fig. 2); and (4) Visualization and ablation studies further demonstrate the effectiveness of our design.

Our contributions are summarized as follows:

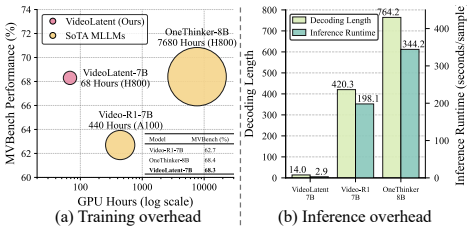


Figure 2: Our VideoLatent achieves stronger or comparable performance with existing CoT-based MLLMs, while achieving superior computational efficiency.

(1) We introduce VideoLatent, a novel MLLM equipped with a latent injection module to enhance video-language learning; (2) We propose a new latent self-forcing training paradigm that enables VideoLatent to perform visual latent reasoning, relying solely on standard video-question-answer triplets; and (3) Extensive experiments conducted on fourteen benchmarks, covering both general video understanding and complex video reasoning, demonstrate the effectiveness, efficiency, and generalizability of our proposed method.

2 Related Work

MLLLMs for Video Understanding and Reasoning. With the rapid evolution of large language models (LLMs) (Vaswani et al., 2017; Touvron et al., 2023; DeepSeek-AI, 2025; Brown et al., 2020; OpenAI, 2023, 2025; Team, 2023; Reid et al., 2024; Yang et al., 2024a; Team, 2025b), multimodal large language models (MLLMs) have witnessed remarkable progress in the vision-language domain (Liu et al., 2023, 2024a; Li et al., 2024a, 2025b; Bai et al., 2023; Li et al., 2026, 2023; Lin et al., 2024; Zhang et al., 2025d; Ye et al., 2025; Hu et al., 2023; Maaz et al., 2024; Deitke et al., 2025; Wang et al., 2025d; Liu et al., 2025c; Dai et al., 2023; He et al., 2025; Assran et al., 2025; Chen et al., 2025a). For example, existing MLLMs have demonstrated strong performance on general video understanding benchmarks (Zhao et al., 2025; Krojer et al., 2025; Xue et al., 2025; Zhang et al., 2024a; Patraucean et al., 2023; Cai et al., 2024; Cores et al., 2024; Shangguan et al., 2025; Hong et al., 2025; Wang et al., 2025c). To further enhance complex video reasoning ability (Hu et al., 2025; Yang et al., 2025a,b; Chen et al., 2025c,b; Cheng et al., 2025; Zhu et al., 2025; Yu et al., 2025), recent efforts have explored incorporating explicit chain-of-thought (CoT) reasoning into MLLMs (Wei et al., 2022; Zhang et al., 2023, 2024b; Zhong et al., 2024; Mitra et al., 2024; Zheng et al., 2023; Chen et al., 2025d; Tian et al., 2025; Wang et al., 2025e).

Specifically, Video-R1 (Feng et al., 2025a) pro-

poses T-GRPO (*i.e.*, a variant of GRPO (Shao et al., 2024)) to efficiently enable R1-style (DeepSeek-AI, 2025) reasoning over videos, while VideoRFT (Wang et al., 2025a) strengthens the alignment between textual reasoning and visual evidence via a semantic-consistency reward. VideoChat-R1.5 (Yan et al., 2025) further improves multimodal reasoning by iteratively refining its focus on fine-grained regions. OneThinker (Feng et al., 2025b) optimizes multimodal reasoning by employing a new EMA-GRPO algorithm, whereas Open-o3-Video (Meng et al., 2025) learns to integrate explicit spatio-temporal evidence into the video reasoning process. Concurrently, VideoAuto-R1 (Liu et al., 2026a) introduces a reason-when-necessary strategy to facilitate dynamic CoT reasoning, while still incurring significant training overhead (*i.e.*, 1120 H100 GPU hours).

Despite their promising performance, existing CoT-based video reasoning approaches primarily rely on explicit CoT reasoning and additional labor-intensive CoT supervision (Feng et al., 2025a), resulting in substantial computational overhead and limited training scalability. In contrast, we propose VideoLatent, which introduces a new latent injection module and latent self-forcing to enhance video-language learning without explicit CoT annotations, while maintaining strong performance and high computational efficiency.

Visual Latent Reasoning. Latent reasoning has demonstrated strong effectiveness in NLP, enabling LLMs to perform reasoning beyond the constraints of language (Goyal et al., 2024; Deng et al., 2024; Shen et al., 2025b; Cheng and Durme, 2024; Xu et al., 2025c; Zou et al., 2025; Yu et al., 2026). For example, (Hao et al., 2024; Xu et al., 2025b) learn to switch between language and latent reasoning modes. Instead of generating explicit textual thoughts, latent reasoning performs implicit reasoning in continuous latent spaces, enabling more efficient and flexible reasoning. Recently, latent reasoning has been extended to a wide range of domains, including image reasoning (Wu et al., 2026; Han et al., 2026; Liu et al., 2025a; Shen et al., 2025a; Tong et al., 2025), 3D reasoning (Chen et al., 2025e), image generation (Chen et al., 2026), robotics (Ma et al., 2025b; Liu et al., 2026b), and autonomous driving (Luo et al., 2026), with some works using foundation models (*e.g.*, Diffusion (Ma et al., 2025a; Shao et al., 2026; Zhang et al., 2025a; He et al., 2024) and DINO (Jeon et al., 2026)) to supervise visual latent generation.

In particular, visual latent reasoning has emerged as a promising direction for enhancing the multimodal understanding and reasoning capabilities of MLLMs (Bigverdi et al., 2025; Dong et al., 2025). For instance, Mirage (Yang et al., 2025c) learns to reconstruct the image embeddings of helper images as visual latent representations, while ILVR (Dong et al., 2025) instead learns to selectively extract context-relevant visual signals from helper images. Furthermore, CoVT (Qin et al., 2025) leverages multiple pretrained vision foundation models (Kirillov et al., 2023; Yang et al., 2024b; Su et al., 2021; Oquab et al., 2024) to supervise the visual latent generation. Concurrently, LVR (Li et al., 2025a) learns to generate fine-grained visual latent representations by reconstructing the embeddings of cropped image regions based on bounding box annotations, whereas Monet (Wang et al., 2025b) further incorporates diverse fine-grained supervision signals (*e.g.*, cropping, grounding, highlighting) into visual latent reasoning. Laser (Wang et al., 2026b) learns to align visual latent representations with cognitive scanpaths annotated by GPT-4o (OpenAI, 2024a), while Mull-Tokens (Ray et al., 2025) enables modality-agnostic latent reasoning using additional annotations, including both textual CoT traces and subgoal images.

Despite recent progress, visual latent reasoning for video tasks remains underexplored and faces significant scalability challenges. Specifically, existing methods primarily rely on additional, image-centric supervision signals for visual latent generation (*e.g.*, CoT traces (Shen et al., 2025a), helper images (Yang et al., 2025c), pretrained vision foundation models (Qin et al., 2025), fine-grained annotations (Li et al., 2025a; Wang et al., 2025b, 2026b)), limiting both training scalability and transferability to video tasks. In contrast, we propose a latent injection module and a latent self-forcing training paradigm tailored for visual latent reasoning, without requiring additional latent supervision.

3 Methodology

In this section, we first describe the formulation of standard MLLMs in Sec. 3.1. To enhance video-language learning, we introduce VideoLatent, a new latent MLLM equipped with a latent injection module in Sec. 3.2. We further propose a new latent self-forcing training paradigm (Sec. 3.3) to enable effective video latent reasoning.

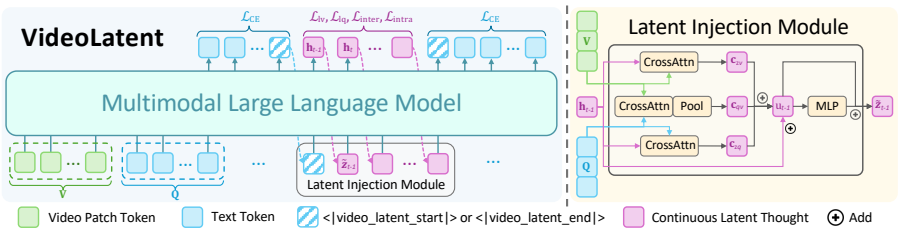


Figure 3: **Overview of VideoLatent**. Given an input video and a text question, our VideoLatent learns to perform visual latent reasoning (see Sec. 3.2) using our proposed latent self-forcing training paradigm (see Sec. 3.3). Specifically, we introduce a latent injection module to prevent self-generated latent thoughts from drifting away from the video and question context. Furthermore, our latent self-forcing covers both latent alignment and latent diversity objectives to enhance video-language learning. Detailed symbol definitions are provided in Sec. 3.

3.1 Preliminary: Standard MLLM

A standard MLLM typically consists of a vision encoder, an LLM $\mathcal{M}_\theta$, and a modality projector. Given an input video $\mathcal{V} = \{v_f\}_{f=1}^{F}$ with F frames and a text question $\mathcal{T}$, the model first extracts a sequence of visual patch tokens $\mathbf{V} = \{\mathbf{v}_{f,k}\}_{f=1,k=1}^{F,N_p}$, where $\mathbf{v}_{f,k} \in \mathbb{R}^d$ denotes the k-th projected patch token of the f-th video frame, d is the hidden dimension, and N_p is the number of patch tokens per frame. The text input is encoded as a sequence of text embeddings $\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^{N_q}$, where $\mathbf{q}_i \in \mathbb{R}^d$ denotes the i-th text token embedding. Subsequently, the model generates an output token sequence $\mathbf{Y} = \{y_t\}_{t=1}^{N_y}$ in an autoregressive manner:

$$P(\mathbf{Y} \mid \mathbf{V}, \mathbf{Q}; \theta) = \prod_{t=1}^{N_y} P(y_t \mid y_{<t}, \mathbf{V}, \mathbf{Q}; \theta), \quad (1)$$

where $y_{<t} = \{y_1, \ldots, y_{t-1}\}$ denotes previously generated tokens. For brevity, we denote the next-token distribution $P(y_t \mid y_{<t}, \mathbf{V}, \mathbf{Q}; \theta)$ as $\mathbf{p}_t$. At each decoding step t, the LLM backbone computes the last-layer hidden state $\mathbf{h}_t \in \mathbb{R}^d$, next-token distribution $\mathbf{p}_t$, and next predicted token y_t as follows (assuming greedy decoding for simplicity):

$$\begin{aligned} \mathbf{h}_t &= \mathcal{M}_\theta([\mathbf{V}, \mathbf{Q}, \mathbf{E}(y_{<t})]), \\ \mathbf{p}_t &= \mathrm{Softmax}(\mathbf{W}_o \mathbf{h}_t), \\ y_t &= \arg\max_{y \in \mathcal{Y}} \mathbf{p}_t(y), \end{aligned} \quad (2)$$

where $[\cdot]$ denotes sequence concatenation, $\mathbf{E}(\cdot)$ is the text embedding lookup function, $\mathbf{W}_o \in \mathbb{R}^{|\mathcal{Y}| \times d}$ denotes the MLLM's language head, and $|\mathcal{Y}|$ denotes the vocabulary size.

3.2 Our Model: VideoLatent

In this section, we detail the design of our VideoLatent, which comprises a latent injection module and learns to perform visual latent reasoning via our proposed latent self-forcing, as shown in Fig. 3.

VideoLatent Overview. Our VideoLatent extends standard MLLMs by augmenting the output space to include both discrete text tokens and continuous latent thoughts (*i.e.*, the last-layer hidden state $\mathbf{h}_t$). Specifically, VideoLatent learns to switch between a language mode and a latent mode, with the former serving as the default, as in standard MLLMs.

At a given decoding step, VideoLatent enters the latent mode upon predicting a special token `<|video_latent_start|>`, and performs visual latent reasoning in the subsequent decoding steps. During the latent mode, our VideoLatent directly generates the last-layer hidden state $\mathbf{h}_t$ as the predicted continuous latent thought, instead of generating a discrete text token y_t. VideoLatent reverts to the language mode either when the generated $\mathbf{h}_t$ is mapped to another special token `<|video_latent_end|>` via the language head (Eq. 2), or when a predefined number N_{max} of latent reasoning steps is reached. Assuming that the latent mode is triggered at step t_s and lasts for N_h steps, we denote the sequence of generated continuous latent thoughts as $\mathbf{H} = \{\mathbf{h}_t\}_{t=t_s+1}^{t_s+N_h}$. Consequently, the hybrid text-latent output sequence $\mathbf{Y}_h$ can be formulated as:

$$\mathbf{Y}_h = [y_{\leq t_s}, \mathbf{H}, y_{>t_s+N_h}], \quad (3)$$

where $y_{\leq t_s}$ and $y_{>t_s+N_h}$ denote the discrete text token sequences before and after the latent mode. In particular, $y_{\leq t_s}$ ends with the `<|video_latent_start|>` token, and $y_{>t_s+N_h}$ begins with the `<|video_latent_end|>` token.

Latent Injection Module. During the latent mode, to generate the continuous latent thought $\mathbf{h}_t$ at step $t \in [t_s + 1, t_s + N_h]$, we define the raw input representation $\mathbf{z}_{t-1} \in \mathbb{R}^d$ fed to our model as:

$$\mathbf{z}_{t-1} = \begin{cases} \mathbf{E}(y_{t_s}), & \text{if } t = t_s + 1, \\ \mathbf{h}_{t-1}, & \text{if } t_s + 1 < t \leq t_s + N_h. \end{cases} \quad (4)$$

To prevent the self-generated latent thoughts from drifting away from the input video and question context, we introduce a latent injection module

$\mathcal{F}_\phi$ tailored for enhancing video-language learning As shown in Fig. 3, the latent injection module dynamically injects the video features $\mathbf{V}$ and text question features $\mathbf{Q}$ into $\mathbf{z}_{t-1}$:

$$\tilde{\mathbf{z}}_{t-1} = \mathcal{F}_\phi(\mathbf{z}_{t-1}, \mathbf{V}, \mathbf{Q}), \tag{5}$$

where $\tilde{\mathbf{z}}_{t-1} \in \mathbb{R}^d$ denotes the ultimate context-aware representation feed into our VideoLatent.

Specifically, $\mathcal{F}_\phi$ first obtains the aggregated semantic information $\mathbf{u}_{t-1} \in \mathbb{R}^d$ from three complementary perspectives (*i.e.*, latent-to-video, latent-to-question, question-to-video) via cross attention:

$$\begin{aligned} \mathbf{c}_{zv} &= \mathrm{CrossAttn}(\mathbf{z}_{t-1}, \mathbf{V}) \in \mathbb{R}^d, \\ \mathbf{c}_{zq} &= \mathrm{CrossAttn}(\mathbf{z}_{t-1}, \mathbf{Q}) \in \mathbb{R}^d, \\ \mathbf{c}_{qv} &= \mathrm{Pool}(\mathrm{CrossAttn}(\mathbf{Q}, \mathbf{V})) \in \mathbb{R}^d, \\ \mathbf{u}_{t-1} &= \mathbf{z}_{t-1} + \mathbf{c}_{zv} + \mathbf{c}_{zq} + \mathbf{c}_{qv} \in \mathbb{R}^d, \end{aligned} \tag{6}$$

where $\mathrm{Pool}(\cdot)$ denotes mean pooling along the sequence dimension. $\mathcal{F}_\phi$ then produces $\tilde{\mathbf{z}}_{t-1}$ via an MLP with a residual connection:

$$\tilde{\mathbf{z}}_{t-1} = \mathbf{u}_{t-1} + \mathrm{MLP}(\mathbf{u}_{t-1}). \tag{7}$$

3.3 Our Training: Latent Self-Forcing

Prior works on visual latent reasoning mainly focus on optimizing the continuous latent thoughts $\mathbf{H} = \{\mathbf{h}_t\}_{t=t_s+1}^{t_s+N_h}$ using additional, image-centric supervision signals for visual latent generation (*e.g.*, CoT traces (Shen et al., 2025a), helper images (Yang et al., 2025c), pretrained vision foundation models (Qin et al., 2025), bounding boxes (Li et al., 2025a), and other fine-grained annotations (Wang et al., 2025b, 2026b)), which limits their scalability and transferability to video reasoning.

In contrast, we introduce a latent self-forcing mechanism specifically designed to enhance video-language learning, without requiring additional annotations (*i.e.*, training relies solely on video–question–answer triplets). Specifically, latent self-forcing comprises latent alignment and latent diversity objectives to supervise the generation of continuous latent thoughts.

Latent Alignment. To prevent the self-generated continuous latent thoughts from drifting away from the visual and linguistic context, we explicitly align $\mathbf{H}$ with $\mathbf{V}$ and $\mathbf{Q}$ via contrastive learning, where we follow (He et al., 2020) to maintain dynamic memory banks with a queue size of K to store negative samples (*e.g.*, videos, questions, generated latent thoughts) from preceding batches.

For *Latent-Video Alignment*, we first apply adaptive pooling to $\mathbf{V}$ to obtain $\bar{\mathbf{V}} \in \mathbb{R}^{N_h \times d}$. We then flatten and L_2-normalize both $\mathbf{H}$ and $\bar{\mathbf{V}}$ into $\tilde{\mathbf{h}} \in \mathbb{R}^{N_h d}$ and $\tilde{\mathbf{v}} \in \mathbb{R}^{N_h d}$. The latent-video alignment loss $\mathcal{L}_{\text{lv}}$ is formulated as:

$$\mathcal{L}_{\text{lv}} = -\log \frac{e^{\mathrm{sim}(\tilde{\mathbf{h}}, \tilde{\mathbf{v}})/\tau}}{e^{\mathrm{sim}(\tilde{\mathbf{h}}, \tilde{\mathbf{v}})/\tau} + \sum_{k=1}^{K} e^{\mathrm{sim}(\tilde{\mathbf{h}}, \tilde{\mathbf{v}}_k^-)/\tau}}, \tag{8}$$

where $\mathrm{sim}(\cdot)$ denotes the cosine similarity, τ is the temperature hyperparameter, and $\tilde{\mathbf{v}}_k^-$ represents the negative video samples from the memory bank.

For *Latent-Question Alignment*, we first employ mean pooling to $\mathbf{Q}$ to extract the global question representation $\bar{\mathbf{q}} \in \mathbb{R}^d$. To ensure the question-awareness of each latent thought, the latent-question alignment loss $\mathcal{L}_{\text{lq}}$ is defined as:

$$\mathcal{L}_{\text{lq}} = -\frac{1}{N_h} \sum_{\mathbf{h}_t \in \mathbf{H}} \log \frac{e^{\mathrm{sim}(\mathbf{h}_t, \bar{\mathbf{q}})/\tau}}{e^{\mathrm{sim}(\mathbf{h}_t, \bar{\mathbf{q}})/\tau} + \sum_{k=1}^{K} e^{\mathrm{sim}(\mathbf{h}_t, \mathbf{q}_k^-)/\tau}}, \tag{9}$$

where $\mathbf{q}_k^-$ represents the negative question samples.

Latent Diversity. To avoid latent representation collapse, we adopt contrastive learning for *Inter-Latent Diversity*. Similar to Eq. 8, the inter-latent diversity loss is formulated as:

$$\mathcal{L}_{\text{inter}} = -\log \frac{e^{\mathrm{sim}(\tilde{\mathbf{h}}, \tilde{\mathbf{h}})/\tau}}{e^{\mathrm{sim}(\tilde{\mathbf{h}}, \tilde{\mathbf{h}})/\tau} + \sum_{k=1}^{K} e^{\mathrm{sim}(\tilde{\mathbf{h}}, \tilde{\mathbf{h}}_k^-)/\tau}}, \tag{10}$$

where $\tilde{\mathbf{h}}_k^-$ denotes the negative samples.

For *Intra-Latent Diversity*, we penalize the similarity between different latent thoughts within $\mathbf{H}$, which is defined as:

$$\mathcal{L}_{\text{intra}} = \frac{1}{N_h(N_h-1)} \sum_{\mathbf{h}_i, \mathbf{h}_j \in \mathbf{H}, i \neq j} (\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_j))^2. \tag{11}$$

Overall Training Objectives. The overall training objectives combine the standard next token prediction loss $\mathcal{L}_{\text{CE}}$ for discrete text token prediction, along with our proposed latent-centric losses:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{cl}}(\mathcal{L}_{\text{lv}} + \mathcal{L}_{\text{lq}} + \mathcal{L}_{\text{inter}}) + \lambda_{\text{intra}}\mathcal{L}_{\text{intra}}, \tag{12}$$

where λ_{cl} and λ_{intra} are balancing coefficients.

4 Experiments

4.1 Experimental Settings

Implementation Details. In this work, our VideoLatent models are built upon Qwen2.5-VL (Bai et al., 2025) and Qwen3-VL (Team, 2025c). During model training, we strictly follow (Feng et al., 2025a; Wang et al., 2025a) to limit the maximum number of input video frames to 16 to ensure training efficiency, where each frame is processed

Model	Frames	General Video Understanding				Complex Video Reasoning		
		MVBench	TempCompass	Video-MME	LongVideoBench	Video-TT	VCR-Bench	VideoMathQA
Standard MLLMs								
Qwen2.5-VL-7B	16	64.6	71.6	56.6	55.4	36.9	46.7	29.0
Qwen2.5-VL-7B (CoT)	16	57.4♠	72.2♠	53.1♠	-	-	-	-
Video-R1-7B	16	62.7♠	72.6♠	57.4♠	55.8	40.8	46.3	26.4
Latent MLLMs								
Monet-7B	16	37.1	31.2	28.9	19.7	20.6	24.7	11.7
LVR-7B	16	64.6	72.0	59.1	56.4	37.0	45.3	26.9
Mull-Token-7B	16	67.0	73.0	57.0	55.7	41.3	48.6	26.4
VideoLatent-7B (Ours)	16	**68.3**	**73.7**	**59.3**	**57.3**	**43.4**	**49.6**	**30.5**
Δ *(vs. Qwen2.5-VL-7B)*		*+3.7*	*+2.1*	*+2.7*	*+1.9*	*+6.5*	*+2.9*	*+1.5*
Standard MLLMs								
Qwen2.5-VL-7B	32	67.4	72.3	59.4	58.8	37.9	50.4	27.9
Qwen2.5-VL-7B (CoT)	32	59.0♠	72.6♠	56.6♠	-	-	-	-
Video-R1-7B	32	63.9♠	73.2♠	59.3♠	56.8	41.8	46.9	24.3
VideoRFT-7B	32	62.1♠	73.7♠	59.8♠	-	-	-	-
Open-o3-Video-7B	32	66.0	-	61.3	58.8	39.3	52.5	25.7
Latent MLLMs								
Monet-7B	32	38.9	34.4	33.2	18.3	20.5	25.5	14.0
LVR-7B	32	65.1	72.0	60.5	**60.5**	38.3	49.0	25.0
Mull-Token-7B	32	68.5	73.6	59.5	57.2	40.8	50.6	24.8
VideoLatent-7B (Ours)	32	**69.4**	**74.1**	**61.4**	58.3	**44.4**	**53.1**	**30.0**
Δ *(vs. Qwen2.5-VL-7B)*		*+2.0*	*+1.8*	*+2.0*	*-0.5*	*+6.6*	*+2.7*	*+2.1*
Standard MLLMs								
Qwen2.5-VL-7B	64	67.1	72.3	62.7	59.8	38.7	51.4	**29.8**
Qwen2.5-VL-7B (CoT)	64	59.2♠	72.9♠	59.6♠	-	-	-	-
Video-R1-7B	64	64.8♠	73.2♠	61.4♠	57.6	41.8	50.8	25.2
Open-o3-Video-7B	64	65.8	-	61.4	59.7	39.7	49.2	25.0
Latent MLLMs								
Monet-7B	64	39.3	34.5	34.8	19.8	21.5	25.1	12.9
LVR-7B	64	65.5	72.0	62.0	**60.1**	39.3	48.0	28.8
Mull-Token-7B	64	68.3	73.6	62.1	58.3	41.1	50.6	25.0
VideoLatent-7B (Ours)	64	**69.2**	**74.2**	**63.8**	**60.1**	**43.9**	**52.5**	29.5
Δ *(vs. Qwen2.5-VL-7B)*		*+2.1*	*+1.9*	*+1.1*	*+0.3*	*+5.2*	*+1.1*	*-0.3*

Table 1: **Main experiments: comparisons with standard and latent MLLMs** on general video understanding and complex video reasoning under varying numbers of input frames. Our model consistently outperforms existing baselines and achieves the best or second-best performance on most benchmarks. ♠: Results are borrowed from their papers, while others are produced under the same experimental setting (*e.g.*, using a frame resolution of 256×32×32 pixels; see Sec. 4.1 for more details). **Bold** / Underlined: Best / Second-best result.

at a max resolution of 128×28×28 pixels and 128×32×32 pixels for Qwen2.5-VL-based and Qwen3-VL-based models, respectively. During evaluation, we further follow (Feng et al., 2025a; Wang et al., 2025a) to double the frame resolution (*e.g.*, 256×28×28 pixels and 256×32×32 pixels for VideoLatent-7B and VideoLatent-8B, respectively) to facilitate a fair comparison. More implementation details are provided in the Appendix.

Evaluation Benchmarks. To ensure a comprehensive evaluation, our experiments are conducted on 14 benchmarks, covering general video understanding and complex video reasoning. General video understanding comprises 3 short-video understanding benchmarks (*i.e.*, MVBench (Li et al., 2024b), TempCompass (Liu et al., 2024b), Next-QA (Xiao et al., 2021)) and 4 long-video understanding benchmarks (*i.e.*, Video-MME (Fu et al., 2025), LongVideoBench (Wu et al., 2024b), LVBench (Wang et al., 2025c), MLVU (Wang et al., 2025c)), while complex video reasoning includes Video-TT (Zhang et al., 2025c), VCR-Bench (Qi et al., 2025), VideoMathQA (Rasheed et al., 2025),

Model	Frames	General Video Understanding			Complex Video Reasoning			
		Next-QA	LV Bench	ML VU	MM VU	Video-Holmes	Video-MMMU	LongVideo-Reason
Standard MLLMs								
Qwen2.5-VL-7B	16	79.6	34.5	59.3	64.8	36.4	55.2	69.8
Qwen2.5-VL-7B (CoT)	16	-	-	-	59.2♠	-	-	-
Video-R1-7B	16	79.7	35.7	61.4	64.2♠	40.4	55.7	69.9
Latent MLLMs								
Mull-Token-7B	16	80.0	35.3	**62.7**	64.0	38.4	51.3	71.6
VideoLatent-7B (Ours)	16	**80.7**	**37.6**	61.7	**65.8**	**40.5**	**57.7**	**71.7**
Δ *(vs. Qwen2.5-VL-7B)*		*+1.1*	*+3.1*	*+2.4*	*+1.0*	*+4.1*	*+2.5*	*+1.9*

Table 2: **Additional experiments** on seven more benchmarks, covering general video understanding and complex video reasoning. Our model achieves the best or second-best performance on all benchmarks.

MMVU (Zhao et al., 2025), Video-Holmes (Cheng et al., 2025), Video-MMMU (Hu et al., 2025), LongVideo-Reason (Chen et al., 2025c).

Comparison Baselines. To facilitate a thorough comparison, we compare our models against both standard and latent MLLMs. Standard MLLMs comprise Qwen2.5-VL-3B/7B (Bai et al., 2025), Video-R1-7B (Feng et al., 2025a), TinyLLaVA-Video-R1 (Zhang et al., 2025b), VideoRFT-7B (Wang et al., 2025a), Open-o3-Video-7B (Meng et al., 2025), VideoChat-R1.5 (Yan et al., 2025), Qwen3-VL-8B (Team, 2025c), OneThinker-

Model	Frames	General Video Understanding				Complex Video Reasoning		
		MV Bench	Temp Compass	Video-MME	LongVideo Bench	Video-TT	VCR-Bench	Video MathQA
GPT-4o♠	-	-	-	71.9	66.7	45.2	46.9	20.2
Gemini-2.5-Pro♠	-	-	-	84.3	-	-	-	-
Seed1.5-VL♠	-	74.4	83.7	77.9	74.0	-	-	-
VideoChat-R1.5♠	2048	70.6	-	65.2	61.4	-	-	-
VideoAuto-R1♠	2048	72.0	-	71.7	67.4	-	-	-
Qwen3-VL-8B	16	67.7	75.6	**60.7**	**58.3**	39.2	49.4	29.0
OneThinker-8B	16	68.4	**76.6**	57.4	56.3	38.7	46.9	29.8
VideoLatent-8B	16	**70.6**	76.5	60.5	57.8	**44.4**	**51.0**	**31.4**
Δ *(vs. Qwen3-VL-8B)*		*+2.9*	*+0.9*	*-0.2*	*-0.5*	*+5.2*	*+1.6*	*+2.4*
Qwen3-VL-8B	32	68.9	75.5	**64.1**	60.1	42.1	53.9	30.0
OneThinker-8B	32	69.9	**76.3**	61.4	59.6	39.2	53.1	30.5
VideoLatent-8B	32	**70.6**	76.2	63.5	**60.4**	**45.1**	**58.4**	**31.7**
Δ *(vs. Qwen3-VL-8B)*		*+1.7*	*+0.7*	*-0.6*	*+0.3*	*+3.0*	*+4.5*	*+1.7*
Qwen3-VL-8B	64	69.0	75.7	**67.2**	62.5	43.2	55.3	30.2
OneThinker-8B	64	69.9	76.3	63.8	59.7	40.8	55.7	**32.6**
VideoLatent-8B	64	**71.6**	**76.4**	65.7	**63.0**	**46.4**	**59.8**	31.4
Δ *(vs. Qwen3-VL-8B)*		*+2.6*	*+0.7*	*-1.5*	*+0.5*	*+3.2*	*+4.5*	*+1.2*

Table 3: **Generalization to a different MLLM backbone**. Our model improves the performance of Qwen3-VL-8B on most benchmarks, demonstrating its generalizability across architectures.

8B (Feng et al., 2025b), and VideoAuto-R1 (Liu et al., 2026a). Latent MLLMs encompass Monet-7B (Wang et al., 2025b), LVR-7B (Li et al., 2025a), and Mull-Token-7B (Ray et al., 2025). We also include three closed-source models, including GPT-4o (OpenAI, 2024a), Gemini-2.5-Pro (Team, 2025a), and Seed1.5-VL (Seed, 2025).

4.2 Main Experimental Results

To ensure a fair comparison, all models in Tab. 1 and Tab. 2 are built upon Qwen2.5-VL-7B.

Compared with standard MLLMs. As shown in Tab. 1 and Tab. 2, our VideoLatent-7B consistently outperforms existing standard MLLMs across all 14 benchmarks under varying numbers of input frames (from 16 to 64). For example, compared with the MLLM backbone, VideoLatent-7B exhibits stronger or comparable performance across all benchmarks (*e.g.*, +3.7 on MVBench (short-video understanding; 16 frames), +2.7 on Video-MME (long-video understanding; 16 frames), and +6.5 on Video-TT (complex video reasoning; 16 frames)). In contrast, we observe that the CoT-based MLLMs fail to achieve consistent performance improvement over the MLLM backbone. For example, Video-R1-7B / Open-o3-Video-7B achieves -2.3 / -1.3 on MVBench (64 frames), -1.3 / -1.3 on Video-MME (64 frames), and -4.6 / -4.8 on VideoMathQA (64 frames).

Moreover, our VideoLatent-7B surpasses the CoT-based MLLMs by a notable margin on most benchmarks. For example, compared with Video-R1-7B / Open-o3-Video-7B, VideoLatent-7B achieves +5.5 / +3.4 on MVBench (32 frames), +2.4 / +2.4 on Video-MME (64 frames), and +5.7 / +4.3 on VideoMathQA (32 frames).

Model	Frames	General Video Understanding				Complex Video Reasoning		
		MV Bench	Temp Compass	Video-MME	LongVideo Bench	Video-TT	VCR-Bench	Video MathQA
Qwen2.5-VL-3B	16	62.9	68.0	53.9	52.4	38.7	44.7	24.8
TinyLLaVA-Video-R1♠	16	49.5	-	46.6	-	-	-	-
VideoRFT-3B♠	32	59.5	61.0	45.4	-	-	-	-
VideoLatent-3B (Ours)	16	**65.2**	**69.8**	**54.4**	**54.2**	**44.6**	**45.5**	**25.7**
Δ *(vs. Qwen2.5-VL-3B)*		+2.3	+1.8	+0.5	+1.8	+5.9	+0.8	+0.9

Table 4: **Generalization to a small-scale model**. Our model outperforms other models, further demonstrating its effectiveness across different model scales.

Compared with latent MLLMs. Tab. 1 and Tab. 2 further demonstrate that our VideoLatent-7B outperforms latent MLLMs on most benchmarks. For Monet-7B, we observe that it often struggles to follow instructions, leading to relatively weak performance on video-language tasks. Compared with LVR-7B / Mull-Token-7B, our VideoLatent-7B achieves +3.7 / +1.3 on MVBench (short-video understanding; 16 frames), +0.9 / +1.6 on LongVideoBench (long-video understanding; 16 frames), and +3.6 / +4.1 on VideoMathQA (complex video reasoning; 16 frames).

Overall. These extensive comparative experiments validate the effectiveness of our model design and the proposed latent self-forcing for enhancing video understanding and reasoning.

4.3 Generalization Analysis

Generalization to different MLLM backbones. To validate the generalizability of our approach to different MLLM backbones, we extend our method to Qwen3-VL-8B (using $N_{\max} = 2$ to reduce training overhead). Tab. 3 shows that VideoLatent-8B consistently yields superior or competitive performance against the standard MLLMs under the same experimental setting. Specifically, compared with the MLLM backbone / OneThinker-8B, our VideoLatent-8B achieves +2.9 / +2.2 on MVBench (short-video understanding; 16 frames), +0.5 / +3.3 on Video-MME (long-video understanding; 64 frames), +5.2 / +5.7 on Video-TT (complex video reasoning; 16 frames). Moreover, our VideoLatent-8B (64 frames) even surpasses VideoChat-R1.5 (using extremely large numbers of input frames) on general video understanding (*e.g.*, +1.0 on MVBench, +0.5 on Video-MME, and +1.6 on LongVideoBench), while also outperforming GPT-4o on some complex video reasoning tasks (*e.g.*, +1.2 on Video-TT and +12.9 on VCR-Bench).

Generalization to different model scales. To assess the generalizability of our approach under lim-

Training	Frames	General Video Understanding				Complex Video Reasoning		
		MV Bench	Temp Compass	Video-MME	LongVideo Bench	Video-TT	VCR-Bench	Video MathQA
SFT	16	65.8	72.4	57.5	53.0	**43.9**	46.3	28.8
Ours	16	**68.3**	**73.7**	**59.3**	**57.3**	43.4	**49.6**	**30.5**
Δ *(vs. SFT)*		*+2.5*	*+1.3*	*+1.8*	*+4.3*	*-0.5*	*+3.3*	*+1.7*

Table 5: **Comparison between different training methods**. Our latent self-forcing consistently outperforms standard SFT on most benchmarks.

Figure 4: Case study and heatmap visualization of the generated latent thoughts. See more details in Appendix.

ited computational resources, we further extend our method to Qwen-2.5VL-3B (using $N_{\text{max}} = 2$ to reduce training overhead). Tab. 4 shows that our VideoLatent-3B achieves the best performance across all benchmarks, outperforming both the MLLM backbone and prior works with more input frames (*i.e.*, VideoRFT-3B (Wang et al., 2025a)).

Overall. These results demonstrate the strong generalizability of our method across different MLLM backbones and model scales.

4.4 Further Analyses

SFT vs. Latent self-forcing. Tab. 5 shows that our latent self-forcing consistently outperforms standard SFT on most benchmarks, validating the effectiveness of the proposed method.

Computational efficiency. Fig. 2 demonstrates that our method exhibits significantly improved computational efficiency (training overheads are borrowed from (Feng et al., 2025a,b), inference overheads are calculated based on VideoMathQA (64 frames). Specifically, compared with Video-R1-7B, VideoLatent-7B significantly reduces training overhead by ∼6× and inference overhead by ∼68×, while maintaining strong performance (as shown in Tab. 1 and Tab. 2).

Qualitative analysis. Fig. 4 presents qualitative comparisons between CoT-based MLLMs and our model. The case study shows that, while CoT-based MLLMs exhibit overthinking behaviors and incorrect reasoning, our model effectively internalizes the video reasoning with significant inference efficiency, consistent with the results in Fig. 2.

Moreover, to better understand the implicit reasoning process, we provide a heatmap visualization of the self-generated continuous latent thoughts based on the similarity between video patch tokens and latent thoughts. Specifically, we observe that the latent thoughts dynamically capture key objects across different frames, thereby enhancing video understanding and reasoning capabilities.

N_{max}	0	2	4	6	8
LongVideoBench (16 frames)	55.4	56.3	56.7	**57.3**	56.4

Table 6: Ablation of latent reasoning steps N_{max}.

Model	Latent Self-Forcing				Latent Injection Module			Full
	w/o $\mathcal{L}_{\text{lv}}$	w/o $\mathcal{L}_{\text{lq}}$	w/o $\mathcal{L}_{\text{inter}}$	w/o $\mathcal{L}_{\text{intra}}$	w/o $\mathbf{c}_{zv}$	w/o $\mathbf{c}_{zq}$	w/o $\mathbf{c}_{qv}$	
LongVideoBench	55.6	56.1	56.6	56.6	55.6	55.2	55.3	**57.3**

Table 7: Ablation of training and module designs.

4.5 Ablation Studies

Latent reasoning steps. As shown in Tab. 6, we ablate the effect of the maximum number of latent reasoning steps (*i.e.*, N_{max} in Sec. 3.2). The results demonstrate that $N_{\text{max}} = 6$ achieves the best performance, while even $N_{\text{max}} = 2$ outperforms $N_{\text{max}} = 0$ (*i.e.*, the MLLM backbone). The performance first increases and then decreases as N_{max} increases, suggesting that an overly small N_{max} is insufficient for effective video-language learning, while an excessively large N_{max} may impair the backbone's pretrained text reasoning capability.

Latent self-forcing. Tab 7 presents ablations of different latent-centric objectives, where removing any loss consistently degrades performance, demonstrating the effectiveness of our latent self-forcing training paradigm. Among these objectives, $\mathcal{L}_{\text{lv}}$ and $\mathcal{L}_{\text{lq}}$ contribute the most (+1.7 and +1.2), highlighting the importance of aligning latent thoughts with both video and question contexts.

Latent injection module. Tab 7 further ablates the design of the latent injection module, where removing any injected component consistently leads to performance degradation, indicating the importance of injecting both video and question information for enhancing video-language learning.

5 Conclusion

In this paper, we present VideoLatent, a novel framework specially designed for visual latent reasoning on video understanding and reasoning. Specifically, VideoLatent involves a new latent injection module and latent self-forcing training paradigm, enabling effective latent reasoning without relying on additional supervision signals. Extensive experiments conducted on fourteen video-language benchmarks demonstrate that our VideoLatent consistently outperforms existing standard and latent MLLMs across general video understanding and complex video reasoning, while achieving superior efficiency and generalizability.

Limitations

We summarize the limitations as follows: (1) Despite our model achieving strong performance across various general video understanding and complex video reasoning benchmarks, the results may not generalize to all video-language benchmarks; (2) Due to computational constraints, the maximum number of input video frames in our current training is limited to 16. Future work could benefit from leveraging higher-quality training data with an increased number of frames to further enhance video-language learning; (3) Since our method relies on self-generated latent thoughts without additional supervision signals, the generated latent thoughts may still be irrelevant to the video and question context; and (4) Despite its superior inference efficiency, similar to prior visual latent reasoning works, our implicit latent reasoning may provide relatively lower interpretability compared with explicit CoT reasoning.

References

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew J. Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, and 10 others. 2025. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *CoRR*, abs/2506.09985.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL technical report. *CoRR*, abs/2502.13923.

Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G. Shapiro, and Ranjay Krishna. 2025. Perception tokens enhance visual reasoning in multimodal language models. In *CVPR*, pages 3836–3845. Computer Vision Foundation / IEEE.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *NeurIPS*.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. 2024. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *CoRR*, abs/2410.10818.

Delong Chen, Mustafa Shukor, Théo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. 2025a. VL-JEPA: joint embedding predictive architecture for vision-language. *CoRR*, abs/2512.10942.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. 2025b. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *ICLR*. OpenReview.net.

Harold Haodong Chen, Xinxiang Yin, Wen-Jie Shu, Hongfei Zhang, Zixin Zhang, Chenfei Liao, Litao Guo, Qifeng Chen, and Ying-Cong Chen. 2026. Show, don't tell: Morphing latent reasoning into image generation. *CoRR*, abs/2602.02227.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. 2025c. Scaling RL to long videos. *CoRR*, abs/2507.07966.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. 2025d. Scaling RL to long videos. *CoRR*, abs/2507.07966.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. 2025e. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. *CoRR*, abs/2510.18632.

Jeffrey Cheng and Benjamin Van Durme. 2024. Compressed chain of thought: Efficient reasoning through dense representations. *CoRR*, abs/2412.13171.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. 2025. Video-holmes: Can MLLM think like holmes for complex video reasoning? *CoRR*, abs/2505.21374.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. 2024. Tvbench: Redesigning video-language evaluation. *CoRR*, abs/2410.07752.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, Yen-Sung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 31 others. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *CVPR*, pages 91–104. Computer Vision Foundation / IEEE.

Yuntian Deng, Yejin Choi, and Stuart M. Shieber. 2024. From explicit cot to implicit cot: Learning to internalize cot step by step. *CoRR*, abs/2405.14838.

Shuai Dong, Siyuan Wang, Xingyu Liu, and Zhongyu Wei. 2025. Interleaved latent visual reasoning with selective perceptual modeling. *CoRR*, abs/2512.05665.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *ICML*, Proceedings of Machine Learning Research, pages 13109–13125. PMLR / OpenReview.net.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. 2025a. Video-r1: Reinforcing video reasoning in mllms. *CoRR*, abs/2503.21776.

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. 2025b. Onethinker: All-in-one reasoning model for image and video. *CoRR*, abs/2512.03043.

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, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, pages 24108–24118. Computer Vision Foundation / IEEE.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. 2025. Interleaved-modal chain-of-thought. In *CVPR*, pages 19520–19529. Computer Vision Foundation / IEEE.

Sara Ghazanfari, Francesco Croce, Nicolas Flammarion, Prashanth Krishnamurthy, Farshad Khorrami, and Siddharth Garg. 2025. Chain-of-frames: Advancing video understanding in multimodal llms via frame-aware reasoning. *CoRR*, abs/2506.00318.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2024. Think before you speak: Training language models with pause tokens. In *ICLR*. OpenReview.net.

Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. 2025. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *CVPR*, pages 26181–26191. Computer Vision Foundation / IEEE.

Yudong Han, Yong Wang, Zaiquan Yang, Zhen Qu, Liyuan Pan, and Xiangxiang Chu. 2026. Visual enhanced depth scaling for multimodal latent reasoning. *CoRR*, abs/2604.10500.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *CoRR*, abs/2412.06769.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735. Computer Vision Foundation / IEEE.

Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. 2024. Multi-modal latent space learning for chain-of-thought reasoning in language models. In *AAAI*, pages 18180–18187. AAAI Press.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. 2025. Videossr: Video self-supervised reinforcement learning. *CoRR*, abs/2511.06281.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. 2025. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *CVPR*, pages 8450–8460. Computer Vision Foundation / IEEE.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *CoRR*, abs/2501.13826.

Zi-Yuan Hu, Yanyang Li, Michael R. Lyu, and Liwei Wang. 2023. VL-PET: vision-and-language parameter-efficient tuning via granularity control. In *ICCV*, pages 2998–3008. IEEE.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *CoRR*, abs/2503.06749.

Byungwoo Jeon, Yoonwoo Jeong, Hyunseok Lee, Minsu Cho, and Jinwoo Shin. 2026. Vision-aligned latent reasoning for multi-modal large language model. *CoRR*, abs/2602.04476.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment anything. In *ICCV*, pages 3992–4003. IEEE.

Benno Krojer, Mojtaba Komeili, Candace Ross, Quentin Garrido, Koustuv Sinha, Nicolas Ballas, and Mido Assran. 2025. A shortcut-aware video-qa benchmark for physical understanding via minimal video pairs. *Trans. Mach. Learn. Res.*, 2025.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. 2025a. Latent visual reasoning. *CoRR*, 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. 2025b. LLaVA-OneVision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulic, and Furu Wei. 2025c. Imagine while reasoning in space: Multimodal visualization-of-thought. In *ICML*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024a. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *CoRR*, abs/2407.07895.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pages 22195–22206. IEEE.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2026. Mini-gemini: Mining the potential of multi-modality vision language models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(3):3530–3543.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. VILA: on pre-training for visual language models. In *CVPR*, pages 26679–26689. IEEE.

Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyue Wei, Sheng Liu, and Xin Eric Wang. 2025a. Reasoning within the mind: Dynamic multimodal interleaving in latent space. *CoRR*, abs/2512.12623.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *CVPR*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Shuming Liu, Mingchen Zhuge, Changsheng Zhao, Jun Chen, Lemeng Wu, Zechun Liu, Chenchen Zhu, Zhipeng Cai, Chong Zhou, Haozhe Liu, Ernie Chang, Saksham Suri, Hongyu Xu, Qi Qian, Wei Wen, Balakrishnan Varadarajan, Zhuang Liu, Hu Xu, Florian Bordes, and 4 others. 2026a. Videoauto-r1: Video auto reasoning via thinking once, answering twice. *CoRR*, abs/2601.05175.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. 2025b. Videomind: A chain-of-lora agent for long video reasoning. *CoRR*, abs/2503.13444.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024b. Tempcompass: Do video llms really understand videos? In *ACL (Findings)*, Findings of ACL, pages 8731–8772. Association for Computational Linguistics.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Haotian Tang, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Jinyi Hu, Sifei Liu, and 6 others. 2025c. NVILA: efficient frontier visual language models. In *CVPR*, pages 4122–4134. Computer Vision Foundation / IEEE.

Zhuoyang Liu, Jiaming Liu, Hao Chen, Jiale Yu, Ziyu Guo, Chengkai Hou, Chenyang Gu, Xiangju Mi, Renrui Zhang, Kun Wu, Zhengping Che, Jian Tang, Pheng-Ann Heng, and Shanghang Zhang. 2026b. Last_0: Latent spatio-temporal chain-of-thought for robotic vision-language-action model. *CoRR*, abs/2601.05248.

Yuechen Luo, Fang Li, Shaoqing Xu, Yang Ji, Zehan Zhang, Bing Wang, Yuannan Shen, Jianwei Cui, Long Chen, Guang Chen, Hangjun Ye, Zhi-Xin Yang, and Fuxi Wen. 2026. Last-vla: Thinking in latent spatio-temporal space for vision-language-action in autonomous driving. *CoRR*, abs/2603.01928.

Jizheng Ma, Xiaofei Zhou, Yanlong Song, and Han Yan. 2025a. Cocova: Chain of continuous vision-language thought for latent space reasoning. *CoRR*, abs/2511.02360.

Xiangkai Ma, Lekai Xing, Han Zhang, Wenzhong Li, and Sanglu Lu. 2025b. Unifying perception and action: A hybrid-modality pipeline with implicit visual chain-of-thought for robotic action generation. *CoRR*, abs/2511.19859.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL (1)*, pages 12585–12602. Association for Computational Linguistics.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, and Zhuochen Wang. 2025. Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. *CoRR*, abs/2510.20579.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *CVPR*, pages 14420–14431. IEEE.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2024a. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-07-29.

OpenAI. 2024b. Openai o1 system card. *CoRR*, abs/2412.16720.

OpenAI. 2025. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/. Accessed: 2025-08-07.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adrià Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, and 5 others. 2023. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS*.

Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. 2025. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. *CoRR*, abs/2504.07956.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and Xudong Wang. 2025. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *CoRR*, abs/2511.19418.

Hanoona Abdul Rasheed, Abdelrahman M. Shaker, Anqi Tang, Muhammad Maaz, Ming-Hsuan Yang, Salman H. Khan, and Fahad Shahbaz Khan. 2025. Videomathqa: Benchmarking mathematical reasoning via multimodal understanding in videos. *CoRR*, abs/2506.05349.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD*, pages 3505–3506. ACM.

Arijit Ray, Ahmed Abdelkader, Chengzhi Mao, Bryan A. Plummer, Kate Saenko, Ranjay Krishna, Leonidas J. Guibas, and Wen-Sheng Chu. 2025. Mull-tokens: Modality-agnostic latent thinking. *CoRR*, abs/2512.10941.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, and 34 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

ByteDance Seed. 2025. Seed1.5-vl technical report. *CoRR*, abs/2505.07062.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. 2025. TOMATO: assessing visual temporal reasoning capabilities in multimodal foundation models. In *ICLR*. OpenReview.net.

Yifei Shao, Kun Zhou, Ziming Xu, Mohammad Atif Quamar, Shibo Hao, Zhen Wang, Zhiting Hu, and Biwei Huang. 2026. Learning modal-mixed chain-of-thought reasoning with latent embeddings. *CoRR*, abs/2602.00574.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Xuan Shen, Yizhou Wang, Xiangxi Shi, Yanzhi Wang, Pu Zhao, and Jiuxiang Gu. 2025a. Efficient reasoning with hidden thinking. *CoRR*, abs/2501.19201.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025b. CODI: compressing chain-of-thought into continuous space via self-distillation. In *EMNLP*, pages 677–693. Association for Computational Linguistics.

Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu. 2021. Pixel difference networks for efficient edge detection. In *ICCV*, pages 5097–5107. IEEE.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Gemini Team. 2025a. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261.

Qwen Team. 2025b. Qwen3 technical report. *CoRR*, abs/2505.09388.

Qwen Team. 2025c. Qwen3-vl technical report. *CoRR*, abs/2511.21631.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. 2025. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *CoRR*, abs/2506.13654.

Jintao Tong, Jiaqi Gu, Yujing Lou, Lubin Fan, Yixiong Zou, Yue Wu, Jieping Ye, and Ruixuan Li. 2025. Sketch-in-latents: Eliciting unified reasoning in mllms. *CoRR*, abs/2512.16584.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. 2025a. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *CoRR*, abs/2505.12434.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. 2025b. Monet: Reasoning in latent visual space beyond images and language. *CoRR*, abs/2511.21395.

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. 2025c. Lvbench: An extreme long video understanding benchmark. In *ICCV*, pages 22958–22967. IEEE.

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, and 56 others. 2025d. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *CoRR*, abs/2508.18265.

Yifan Wang, Shiyu Li, Peiming Li, Xiaochen Yang, Yang Tang, and Zheng Wei. 2026a. Render-of-thought: Rendering textual chain-of-thought as images for visual latent reasoning. *CoRR*, abs/2601.14750.

Yubo Wang, Juntian Zhang, Yichen Wu, Yankai Lin, Nils Lukas, and Yuhan Liu. 2026b. Forest before trees: Latent superposition for efficient visual reasoning. *CoRR*, abs/2601.06803.

Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. 2025e. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In *EMNLP*, pages 28126–28140. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. 2024a. STAR: A benchmark for situated reasoning in real-world videos. *CoRR*, abs/2405.09711.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024b. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *NeurIPS*.

Linquan Wu, Tianxiang Jiang, Yifei Dong, Haoyu Yang, Fengji Zhang, Shichaang Meng, Ai Xuan, Linqi Song, and Jacky Keung. 2026. Lavit: Aligning latent visual thoughts for multi-modal reasoning. *CoRR*, abs/2601.10129.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786. Computer Vision Foundation / IEEE.

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulic. 2025a. Visual planning: Let's think only with images. *CoRR*, abs/2505.11409.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025b. Softcot: Soft chain-of-thought for efficient reasoning with llms. In *ACL (1)*, pages 23336–23351. Association for Computational Linguistics.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025c. Softcot++: Test-time scaling with soft chain-of-thought reasoning. *CoRR*, abs/2505.11484.

Zihui Xue, Mi Luo, and Kristen Grauman. 2025. Seeing the arrow of time in large multimodal models. *CoRR*, abs/2506.03340.

Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. 2025. Videochat-r1.5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *CoRR*, abs/2509.21100.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024a. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025a. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, pages 10632–10643. Computer Vision Foundation / IEEE.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024b. Depth anything V2. In *NeurIPS*.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. 2025b. Cambrian-s: Towards spatial supersensing in video. *CoRR*, abs/2511.04670.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. 2025c. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *CoRR*, abs/2506.17218.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2025. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *ICLR*. OpenReview.net.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020. CLEVRER: collision events for video representation and reasoning. In *ICLR*. OpenReview.net.

Jiashuo Yu, Yue Wu, Meng Chu, Zhifei Ren, Zizheng Huang, Pei Chu, Ruijie Zhang, Yinan He, Qirui Li, Songze Li, Zhenxiang Li, Zhongying Tu, Conghui He, Yu Qiao, Yali Wang, Yi Wang, and Limin Wang. 2025. Vrbench: A benchmark for multi-step reasoning in long narrative videos. In *ICCV*, pages 21655–21666. IEEE.

Xinlei Yu, Zhangquan Chen, Yongbo He, Tianyu Fu, Cheng Yang, Chengming Xu, Yue Ma, Xiaobin Hu, Zhe Cao, Jie Xu, Guibin Zhang, Jiale Tao, Jiayi Zhang, Siyuan Ma, Kaituo Feng, Haojie Huang, Youxing Li, Ronghao Chen, Huacan Wang, and 18 others. 2026. The latent space: Foundation, evolution, mechanism, ability, and outlook. *CoRR*, abs/2604.02029.

Yu-Liang Zhan, Xinyu Tang, Han Wan, Jian Li, Jirong Wen, and Hao Sun. 2026. L2v-cot: Cross-modal transfer of chain-of-thought reasoning via latent intervention. In *AAAI*, pages 12358–12366. AAAI Press.

Huanyu Zhang, Wenshan Wu, Chengzu Li, Ning Shang, Yan Xia, Yangyu Huang, Yifan Zhang, Li Dong, Zhang Zhang, Liang Wang, Tieniu Tan, and Furu Wei. 2025a. Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in mllms. *CoRR*, abs/2510.24514.

Jianrui Zhang, Mu Cai, and Yong Jae Lee. 2024a. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *CoRR*, abs/2410.02763.

Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. 2025b. Tinyllava-video-r1: Towards smaller lmms for video reasoning. *CoRR*, abs/2504.09641.

Yuanhan Zhang, Yunice Chew, Yuhao Dong, Aria Leo, Bo Hu, and Ziwei Liu. 2025c. Towards video thinking test: A holistic benchmark for advanced video reasoning and understanding. In *ICCV*, pages 20626–20636. IEEE.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2025d. Llava-video: Video instruction tuning with synthetic data. *Trans. Mach. Learn. Res.*, 2025.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *ICLR*. OpenReview.net.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024b. Multimodal chain-of-thought reasoning in language models. *Trans. Mach. Learn. Res.*, 2024.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, Chengye Wang, Ziyao Shangguan, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. 2025. MMVU: measuring expert-level multi-discipline video understanding. In *CVPR*, pages 8475–8489. Computer Vision Foundation / IEEE.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *NeurIPS*.

Yiwu Zhong, Zi-Yuan Hu, Yin Li, and Liwei Wang. 2025. Rethinking chain-of-thought reasoning for videos. *CoRR*, abs/2512.09616.

Yiwu Zhong, Zi-Yuan Hu, Michael R. Lyu, and Liwei Wang. 2024. Beyond embeddings: The promise of visual table in visual reasoning. In *EMNLP*, pages 6876–6911. Association for Computational Linguistics.

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. 2025. MLVU: benchmarking multi-task long video understanding. In *CVPR*, pages 13691–13701. Computer Vision Foundation / IEEE.

Kejian Zhu, Zhuoran Jin, Hongbang Yuan, Jiachun Li, Shangqing Tu, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2025. MMR-V: what's left unsaid? A benchmark for multimodal deep reasoning in videos. *CoRR*, abs/2506.04141.

Jiaru Zou, Xiyuan Yang, Ruizhong Qiu, Gaotang Li, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, James Zou, Mengdi Wang, and Ling Yang. 2025. Latent collaboration in multi-agent systems. *CoRR*, abs/2511.20639.

In the Appendix, we provide more implementation details in Sec. A, including MLLM backbones, training details, training time cost, evaluation details, and instruction prompt. Subsequently, we describe the statistics of both training data and evaluation data in Sec. B. Finally, we provide comprehensive visualization results in Sec. C.

A Implementation details

MLLM Backbones. In this work, our VideoLatent-7B, VideoLatent-8B models, and VideoLatent-3B are built upon Qwen2.5-VL-7B-Instruct (Bai et al., 2025), Qwen3-VL-8B-Instruct (Team, 2025c), and Qwen2.5-VL-3B-Instruct (Bai et al., 2025), respectively. Our main experiments are conducted on VideoLatent-7B (*e.g.*, Tab. 1, Tab. 2, Tab. 5, Tab. 7, Fig. 1, Fig. 2, and Fig. 4), while we validate the generalizability of our method using VideoLatent-8B models (*i.e.*, Tab. 3) and VideoLatent-3B (*i.e.*, Tab. 4).

Training details. As mentioned in the main paper, for training, we strictly follow (Feng et al., 2025a; Wang et al., 2025a) to limit the maximum number of input video frames to 16 to ensure training efficiency. Specifically, the maximum resolution of each frame is set to 128×28×28 pixels and 128×32×32 pixels for Qwen2.5-VL-based and Qwen3-VL-based models, respectively.

We train our models for 2 epochs with a per-device batch size of 1 and gradient accumulation steps of 16 on 4 NVIDIA H800 (80GB) GPUs, resulting in a total batch size of 64. During training, the modality projector, the LLM, and the proposed latent injection module are trainable, while the vision encoder remains frozen. To ensure training stability, we use the Adam optimizer with a learning rate of 1e-6, while employing a learning rate of 1e-5 on the newly introduced latent injection module. We further apply a linear decay scheduler with a warmup of 100 steps, a weight decay of 0.01, and a maximum gradient norm of 1.0.

For the latent self-forcing training paradigm, we set the maximum number of latent reasoning steps $N_{max} = 6$ for our VideoLatent-7B to achieve the best performance (as demonstrated in Tab. 6), while using $N_{max} = 2$ for VideoLatent-3B and VideoLatent-8B to reduce training overhead. The queue size of the memory banks and temperature hyperparameter used for contrastive learning (Eq. 8, Eq. 9, and Eq. 10) are set to $K = 64$ and $\tau = 0.2$, respectively. The balancing coefficients in Eq. 12

are set to $\lambda_{\text{cl}} = 0.1$ and $\lambda_{\text{intra}} = 0.01$.

Training time cost. For training, we use with DeepSpeed Zero2 (Rasley et al., 2020) to accelerate the training, where VideoLatent-7B ($N_{max} = 6$) takes 17 hours (*i.e.*, 68 H800 GPU hours), VideoLatent-8B ($N_{max} = 2$) takes 14 hours (*i.e.*, 64 H800 GPU hours), and VideoLatent-3B ($N_{max} = 2$) takes 8 hours (*i.e.*, 32 H800 GPU hours).

Evaluation details. During evaluation, we further follow (Feng et al., 2025a; Wang et al., 2025a) to double the frame resolution (*e.g.*, 256×28×28 pixels and 256×32×32 pixels for VideoLatent-7B and VideoLatent-8B, respectively) to facilitate a fair comparison. All experiments are conducted with 1 NVIDIA H800 GPU.

Instruction prompt. We use the following instruction prompt to evaluate our VideoLatent: `Please think with the given video and answer the multiple-choice question with the option's letter.\nThe final answer MUST BE put in \\boxed{}.` When reproducing the experimental results of the baselines, we use their official system and instruction prompts provided in their paper or official code. For prior visual latent reasoning works, we adopt the multi-image setting to evaluate the performance of Monet-7B (Wang et al., 2025b)and LVR-7B (Li et al., 2025a), while Mull-Token-7B directly takes video inputs.

Dataset	#Samples
LLaVA-Video-178K (Zhang et al., 2025d)	59,641
NeXT-QA (Xiao et al., 2021)	6,095
PerceptionTest (Patraucean et al., 2023)	3,857
CLEVRER (Yi et al., 2020)	4,188
STAR (Wu et al., 2024a)	6,942
Total	**80,723**

Table 8: Statistics of the training data.

Dataset	#Samples
General Video Understanding (Short Video)	
MVBench (Li et al., 2024b)	4,000
TempCompass (Liu et al., 2024b)	7,540
Next-QA (Xiao et al., 2021)	8564
General Video Understanding (Long Video	
Video-MME (Fu et al., 2025)	2,700
LongVideoBench (Wu et al., 2024b)	1,337
LVBench (Wang et al., 2025c)	1549
MLVU (Wang et al., 2025c)	2174
Complex Video Reasoning	
Video-TT (Zhang et al., 2025c)	1,000
VCR-Bench (mc) (Qi et al., 2025)	510
VideoMathQA (mc) (Rasheed et al., 2025)	420
MMVU (mc) (Zhao et al., 2025)	625
Video-Holmes (Cheng et al., 2025)	1837
Video-MMMU (mc) (Hu et al., 2025)	600
LongVideo-Reason (Chen et al., 2025c)	1,000
Total	**33,856**

Table 9: Statistics of the evaluation data, where some data splits are borrowed from Video-R1 (Feng et al., 2025a) and OneThinker (Feng et al., 2025b).

B Data Statistics

In this section, we provide the statistics of both our training data and evaluation data.

Training data. To ensure reproducibility, our training is conducted on publicly available data, containing 81k video-question-answer triplets without any additional annotations (*e.g.*, CoT traces, bounding boxes). The detailed statistics of the training data are listed in Tab. 8.

Evaluation data. In the main paper, we conduct extensive experiments on 14 video-language benchmarks, covering 7 general video understanding benchmarks, and 7 complex video reasoning benchmarks. The detailed statistics of the evaluation data are listed in Tab. 9.

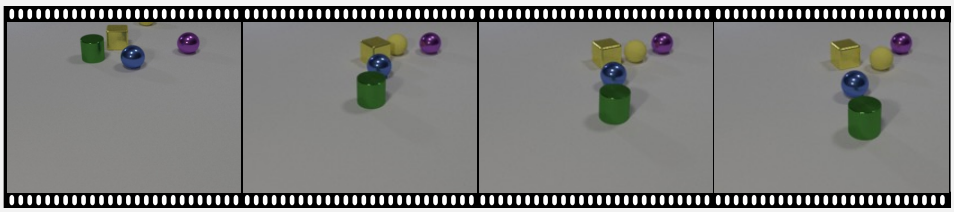
Our main experiments (*e.g.*, Tab. 1, Tab. 3, and Tab. 4) are conducted on seven widely used video-language benchmarks due to computational constraints. For general video understanding, we evaluate on five benchmarks, comprising two short-video benchmarks (*i.e.*, MVBench (Li et al., 2024b) and TempCompass (Liu et al., 2024b)) and two long-video benchmarks (*i.e.*, Video-MME (Fu et al., 2025) and LongVideoBench (Wu et al., 2024b)). For complex video reasoning, we evaluate on three benchmarks, including VCR-Bench (Qi et al., 2025), VideoMathQA (Rasheed et al., 2025), and Video-TT (Zhang et al., 2025c).

Additionally, we provide experimental results on seven more benchmarks (*i.e.*, Next-QA (Xiao et al., 2021), LVBench (Wang et al., 2025c), MLVU (Zhou et al., 2025), MMVU (Zhao et al., 2025), Video-MMMU (Hu et al., 2025), Video-Holmes (Cheng et al., 2025), and LongVideo-Reason (Chen et al., 2025c)) in Tab. 2 to further validate the effectiveness of our method.

C Visualizations

In this section, we provide comprehensive case studies and heatmap visualizations of the self-generated latent thoughts in Fig. 5 and Fig. 6.

Original video:



Question: What direction is the green cylinder moving in within the video?
A. The object is stationary. B. Up and to the right. C. Down and to the right. D. Up and to the left.

Video-R1-7B: <think>Let me think about this for a moment. The green cylinder appears to be moving in the video, so we can rule out option A immediately. Now, let's consider the other options.\n\nFirst, let's break it down by examining the direction of movement. The green cylinder seems to be moving from the left side of the image towards the center-right. This suggests that it is not moving up and to the right, which would be option B. \n\nNext, let's consider option C, which suggests that the green cylinder is moving down and to the right. While it does appear to be moving to the right, there is no clear indication that it is moving downwards. In fact, it seems to be maintaining a relatively constant height, which makes option C less likely.\n\nNow, let's consider option D, which suggests that the green cylinder is moving up and to the left. This option seems the most plausible because the cylinder is indeed moving towards the center-right, but it also appears to be slightly higher in the frame compared to its starting position. This aligns with the idea of moving up and to the left.\n\nHmm, after all this consideration, it seems that the most accurate description of the green cylinder's movement is indeed up and to the left. Oh, I see, the green cylinder is not just moving to the right but also slightly upwards.\n\nLet's verify one last time. The green cylinder is definitely not moving down, and it is not moving purely to the right without any upward component. Therefore, the most accurate option is D.\n\nFinal answer: D\n\nLet's double-check one last time to ensure there are no mistakes. The green cylinder is indeed moving up and to the left, and there is no evidence to suggest otherwise.</think>\n<answer>B</answer> (364 tokens)

Open-o3-Video-7B: The motion of the green cylinder can be observed by tracking its position relative to other objects. At the beginning, the <obj>green cylinder</obj> <box>[0,101,42,143]</box> at<t>0.5</t>s is seen entering the scene from the left. As it moves, the <obj>green cylinder</obj> <box>[76,101,118,143]</box> at<t>1.5</t>s is clearly visible rolling towards the center of the frame. By <t>2.6</t>s, the <obj>green cylinder</obj><box>[109,101,151,143]</box> has moved further to the right and is closer to the blue sphere. This consistent movement confirms that the green cylinder is traveling up and to the right.</think>\n<answer>D</answer> (199 tokens)

VideoLatent-7B (Ours): <|video_latent_pad|> *Continuous Latent Thought* ×6 <|video_latent_end|> \\boxed{C} (14 tokens)

Heatmap visualization of the generated continuous latent thoughts:

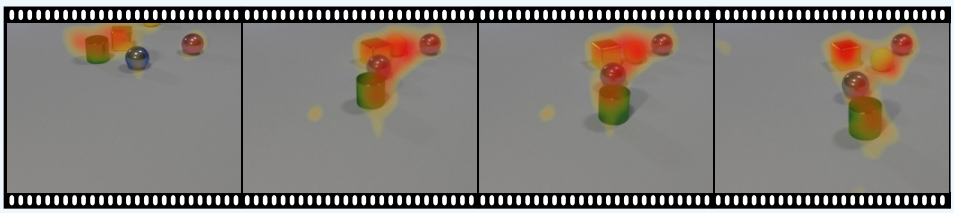
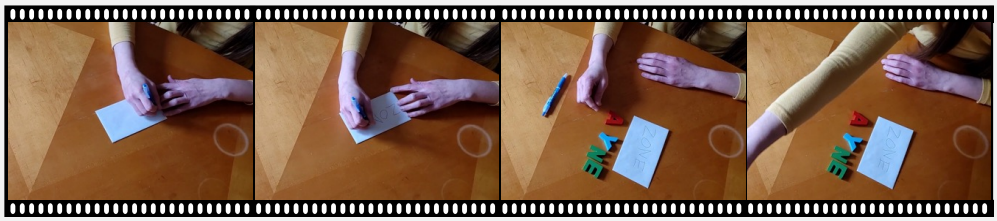


Figure 5: Comprehensive visualization result of a case study and heatmap visualization of the generated continuous latent thought, which is the detailed version of Fig. 4 presented in the main paper.

Original video:



Question: What letter did the person write first on the paper? A. z B. o C. n

Video-R1-7B: <think>Let me think about this for a moment. The person wrote something on the paper, but what was it exactly? Let's break it down step by step.\n\nFirst, I need to recall the sequence of actions. The person placed a pen on the table, then picked up a piece of paper, and started writing. Now, let's consider the options:\n\nA. z - This is a lowercase letter, and it's not immediately obvious if that's what was written.\nB. o - This is also a lowercase letter, and it's possible that the person wrote an \"o\" first.\nC. n - This is another lowercase letter, and it's possible that the person wrote an \"n\" first.\n\nNow, let's think about the context. The person was writing on a piece of paper, and they were using a pen. It's common for people to start writing with a lowercase letter, especially when they are just beginning to write something. \n\nHmm, I need to verify this. The person did indeed write something on the paper, and it appears to be a lowercase letter. Given the options, the most likely candidate is \"o\" because it's a common starting point for writing.\n\nOh, I see. The person likely wrote an \"o\" first on the paper. This makes sense because it's a common starting point for writing and it's a lowercase letter.\n\nLet's double-check. The person did write something on the paper, and it was a lowercase letter. The most logical choice among the options is \"o\".\n\nSo, the letter the person wrote first on the paper is \"o\".\n\nFinal answer: B\n\nLet's verify one last time. The person wrote a lowercase letter, and the most likely candidate is \"o\". Yes, that's correct.</think>\n<answer>B</answer> (373tokens)

Open-o3-Video-7B: <think>The video shows a person writing on a piece of paper. At the beginning, the person is holding a pen and a piece of paper. The person starts writing by placing their hand on the paper and begins to form letters. The first letter written is 'O', which is clearly visible as the person writes it. After completing the letter 'O', the person continues to write other letters. The sequence of writing is consistent with the options provided.</think>\n<answer>B</answer> (99 tokens)

VideoLatent-7B (Ours): <|video_latent_pad|> *Continuous Latent Thought* ×6 <|video_latent_end|> \\boxed{A} (14 tokens)

Heatmap visualization of the generated continuous latent thoughts:

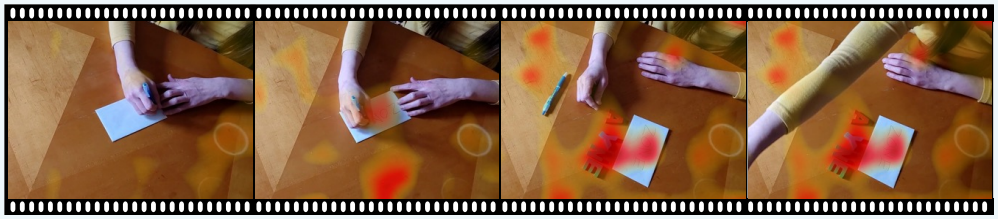


Figure 6: Comprehensive visualization result of a case study and heatmap visualization of the generated continuous latent thoughts.