From Seeing to Foreseeing: Unleashing LVLM Thinking in Dynamic Latent Space

Yuxuan Liang[1]*, **Xu Li**[1]*, **Xiaolei Chen**[1], **Rui Zhu**[1], **Zhe Liu**[1], **Haotian Chen**[1], **Yi Zheng**[1], , **Wenjuan Meng**[1], **Zhuolin He**[1], **Fan Shi**[1,2†], **Xiangyang Xue**[1†]
[1]College of Computer Science and Artificial Intelligence, Fudan University
[2]College of Computer and Data Science, Fuzhou University
{yxliang25, xu_li23, fshi22}@m.fudan.edu.cn, xyxue@fudan.edu.cn

Abstract

Although Large Vision-Language Models (LVLMs) have achieved remarkable progress, complex visual reasoning remains challenging. Existing approaches suffer from fundamental limitations: *Thinking about Image* mainly performs reasoning in text space, *Thinking with Image* relies on external tools to introduce additional visual evidence, and *Latent Visual Reasoning*, despite moving reasoning into latent space, still focuses largely on static visual modeling and lacks explicit characterization of temporal dynamics. We therefore propose **Dynamic Latent Visual Reasoning (DLVR)**, a training framework that enables LVLMs to reason about dynamics in continuous latent space from only a static image. We introduce Mahalanobis Novelty Token Selection and Novelty-Adaptive Temporal Quantization to construct dynamic latent supervision from open-source datasets, building DLVR-SFT-100K and DLVR-RL-4K. We further develop a two-stage SFT pipeline that first builds temporal grounding over explicit dynamic processes and then teaches the model to encode dynamic semantics and temporal structure into latent tokens. Finally, we propose Contrastive Dynamic Latent Policy Optimization, which encourages latent trajectories to align with real dynamics while moving away from counterfactual ones. DLVR empirically improves both dynamic-centric and general visual reasoning, achieving an average gain of 10.82% on BabyVision and a 10.50% improvement on HRBench4K, demonstrating the promise of dynamic latent reasoning for LVLMs. The code is available at https://github.com/Lyxxx1211/NeurIPS2026-DLVR.

1 Introduction

The rapid progress of Large Language Models (LLMs) [2, 16, 64, 21, 35, 65] has significantly advanced language understanding and reasoning, thereby accelerating the development of Large Vision-Language Models (LVLMs) [3, 45, 11, 18, 1, 37, 46]. By leveraging the strong reasoning capabilities of LLMs, LVLMs have achieved remarkable performance across a wide spectrum of multimodal tasks, such as image-text understanding, Visual Question Answering (VQA), and multimodal reasoning. Despite these impressive advances, existing LVLMs still face fundamental challenges in scenarios that require complex visual reasoning [43, 56, 22, 25, 59].

Early efforts on visual reasoning in LVLMs can be categorized into two paradigms. *Thinking about Image* grounds reasoning in a static visual context by translating visual features into discrete textual descriptions, after which reasoning is primarily performed in language space [85, 77, 29]. Although this paradigm enhances language-based reasoning, the conversion from visual signals to textual tokens inevitably compresses rich visual information and may discard fine-grained spatial

*Equal Contribution †Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

relations, geometric structures, and visual details that are difficult to verbalize [28, 61]. *Thinking with Image* expands the visual reasoning evidence by introducing external perceptual tools, such as image cropping, auxiliary lines, or bounding boxes, effectively turning images into an intermediate workspace for problem solving [33, 26, 87, 59]. While this improves access to localized visual evidence, it fragments the reasoning process and delegates part of the cognitive operation to discrete, rule-based visual manipulations rather than internalizing it within the model itself [61, 36, 70].

To model a more internalized form of visual reasoning, recent studies on *latent visual reasoning* shift the reasoning process from language space into latent space, enabling models to generate abstract visual thoughts [36, 70, 79, 53]. By reducing the reliance on verbose textual rationales and external visual tools, this paradigm provides a promising direction for alleviating both the language bottleneck and tool dependence in visual reasoning. However, existing methods still primarily focus on static scenes, whereas many complex visual reasoning problems require information that cannot be reliably inferred from a single static observation alone, such as temporal causal chains, motion directions, state transitions, and the order of interactions. These dynamic cues are not merely auxiliary details, but often constitute the key evidence required for correct reasoning [56, 52].

A defining property of human visual cognition is that it is not merely perceptual, but also anticipatory based on world knowledge. When observing a static image, humans do not simply recognize visible objects and scenes. Instead, they naturally extend the observation along the temporal axis through mental simulation and forward prediction [52, 10, 8, 19, 6], such as imagining whether and how an unstable block tower may collapse. Even in static visual question answering, humans often enrich perception by simulating plausible dynamics or performing implicit visual operations, such as zooming into relevant regions or highlighting target evidence [7, 23, 34]. In contrast, existing latent visual reasoning methods still center on static visual features or static auxiliary images, without explicitly modeling temporal dynamic evolution. As a result, they remain limited in forming coherent and dynamic implicit visual thoughts.

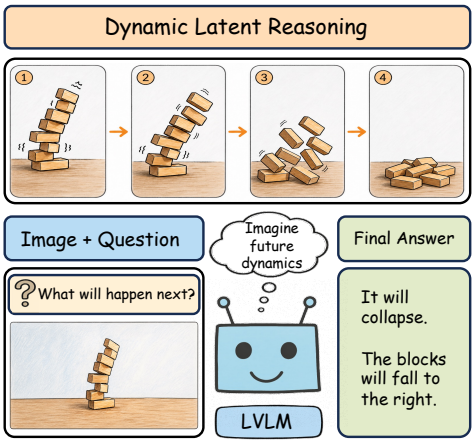


Figure 1: The core idea of our approach.

To overcome these limitations, we propose **Dynamic Latent Visual Reasoning (DLVR)**, a training framework that enables an LVLM to unfold coherent dynamic thinking in continuous latent space. As shown in Figure 1, the key intuition is that effective visual reasoning should not only perceive the static content, but also align its internal reasoning trajectory with plausible future evolution of the physical world. To build such trajectory-level supervision, we first introduce *Mahalanobis Novelty Token Selection* (MNTS) and *Novelty-Adaptive Temporal Quantization* (NATQ), which extract semantically distinctive dynamic cues from subsequent frames via the Gaussian distribution of previous-frame tokens and compress them into fixed-length latent targets. Based on this pipeline, we construct DLVR-SFT-100K and DLVR-RL-4K from open-source datasets covering physical dynamic question answering [68], real-world dynamic question answering [48], and general question answering scenarios [84]. We train the LVLM backbone through a two-stage Supervised Fine-Tuning (SFT), where the model first learns temporal grounding over explicit dynamic processes and then learns to autoregressively generate latent tokens that encode dynamic semantics and temporal structure. Finally, since existing GRPO-based Reinforcement Learning (RL) methods mainly optimize final answers rather than the latent reasoning process itself [70, 36, 57], we introduce *Contrastive Dynamic Latent Policy Optimization* (CDLPO), which directly supervises latent trajectories to stay close to real dynamics while moving away from counterfactual ones. In this way, the LVLM can reason about future dynamics in latent space from a static image. We summarize our contributions as follows:

- We propose **DLVR**, a dynamic latent visual reasoning framework that enables LVLMs to reason about plausible future dynamics in continuous latent space from static input images.
- We construct two datasets, DLVR-SFT-100K and DLVR-RL-4K, and introduce *Mahalanobis Novelty Token Selection* and *Novelty-Adaptive Temporal Quantization* to extract critical dynamic cues from temporal inputs and compress them into fixed-length supervision targets.

- We design a two-stage supervised training pipeline and *Contrastive Dynamic Latent Policy Optimization*, allowing the model to first acquire temporal grounding, then generate dynamic latent tokens autoregressively, and finally refine latent trajectories toward real dynamics and away from counterfactual ones.
- Extensive experiments demonstrate the effectiveness and generality of DLVR. DLVR achieves an average improvement of **10.61%** over the baseline on dynamic-centric benchmarks, including BabyVision [12] and VTBench [63]. It further improves general image understanding across eight benchmarks, including a **10.50%** gain on HRBench4K [72], showing that dynamic latent reasoning also benefits general image understanding.

2 Related Work

Large Vision-Language Models Large Vision-Language Models (LVLMs) connect visual encoders with Large Language Models (LLMs) to support joint understanding and reasoning over visual and textual inputs. Early LVLMs [18, 46] typically paired frozen CLIP-based vision transformers [54, 20] with LLMs through lightweight projection modules, achieving open-ended visual question answering and multimodal understanding. Recent LVLMs [47, 40, 15, 69, 78, 89, 80, 71, 41, 39, 58, 4, 42, 38] further improved visual perception through dynamic tile cropping, stronger visual tokenization, and advanced instruction tuning or reinforcement learning. Despite these advances, current LVLMs still struggle with complex visual reasoning that requires fine-grained perception, long-horizon inference, and deeper integration of world knowledge with visual evidence [56, 22, 25, 59].

Visual Reasoning in LVLMs. Current visual reasoning frameworks can be broadly grouped into three paradigms. The first paradigm, *Thinking about Image*, treats the image as static visual context and performs reasoning primarily through textual chain-of-thought [85, 77, 29, 30, 74, 50, 62, 55, 14]. This line of work enables multi-step reasoning in language space, but keeps visual information as passive context, which limits vision-centric and fine-grained reasoning over spatial, geometric, and dynamic evidence. The second paradigm, *Thinking with Image*, transforms the image into an intermediate workspace by introducing additional visual evidence through cropping, zooming, highlighting, auxiliary lines, or external visual tools [33, 26, 87, 59, 86, 75, 66, 31, 82, 17, 73, 60, 27]. Although these methods enrich visual evidence and improve interpretability, they are often constrained by predefined tool sets and require repeated visual encoding or discrete tool invocation. A more recent paradigm, *Latent Visual Reasoning*, moves intermediate reasoning into continuous latent space, using latent embeddings to carry visual semantics and reduce both the textual bottleneck and dependence on external tools [36, 70, 79, 53]. However, existing methods still mainly focus on static visual features [53], local regions of interest [36], or auxiliary and solution-image representations produced by external tools [79, 70], leaving the future dynamic evolution implied by a static image largely unexplored. Moreover, their reinforcement learning objectives typically optimize final-answer correctness rather than the latent reasoning process itself [36, 70, 79], thereby overlooking temporal dependencies and dynamic coherence among latent tokens. In contrast, DLVR enables LVLMs to internally imagine and temporally model dynamic evolution directly in latent space via two stage SFT and proposed CDLPO, providing a more coherent form of dynamic latent visual reasoning.

3 Method

3.1 Overview of DLVR

As shown in Figure 2a, we propose a **Dynamic Latent Visual Reasoning (DLVR)** framework that enables an LVLM to perform coherent dynamic thinking in continuous latent space from a static image. Given an image $\boldsymbol{I}$ and a question $\boldsymbol{q}$, DLVR generates N *dynamic latent tokens* $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N)$ in an autoregressive manner, interleaved with natural-language tokens to produce the final response.

Formally, let $\boldsymbol{h}_r$ and $\boldsymbol{e}_r$ denote the last-layer hidden state and the input embedding at decoding step r. When the special token `<Dynamic>` is generated, DLVR switches to dynamic latent reasoning. Instead of projecting $\boldsymbol{h}_r$ to the vocabulary space, we reuse it as the next input embedding, i.e., $\boldsymbol{z}_n = \boldsymbol{h}_{s+n}$ and $\boldsymbol{e}_{s+n+1} = \boldsymbol{z}_n$ for $n = 1, \ldots, N$, where s is the number of text tokens before `<Dynamic>`. This produces the latent trajectory $\boldsymbol{Z}$ capturing the internal dynamic reasoning. Following prior work [36, 79, 70, 53], we adopt a fixed latent length N for stability. After N steps, the model emits

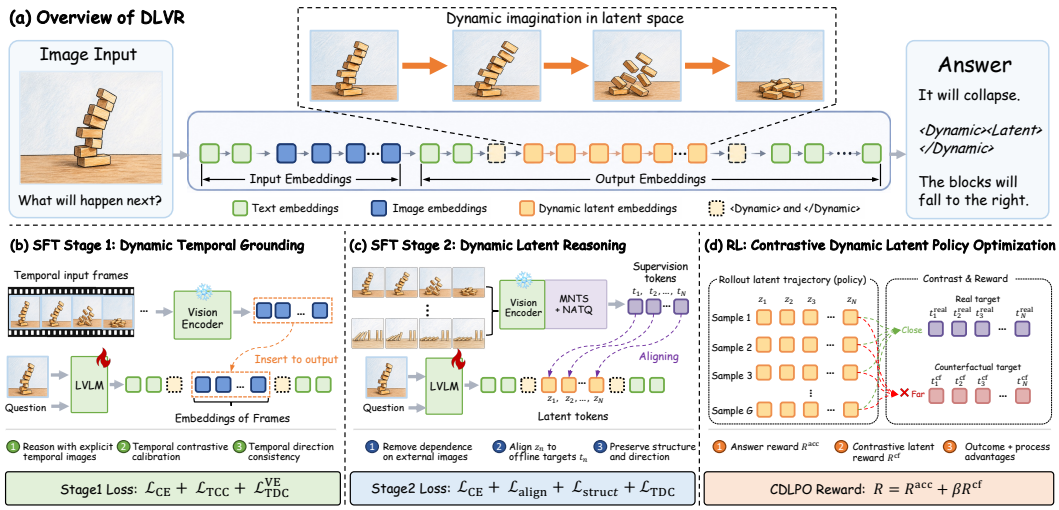


Figure 2: **Overview of the proposed DLVR.** DLVR interleaves text tokens with dynamic latent tokens, allowing the model to internally unfold plausible future scene evolution in continuous latent space before generating the final answer. The training pipeline consists of three stages: SFT Stage 1 teaches dynamic temporal grounding from explicit image sequences, SFT Stage 2 trains model to autoregressively generate dynamic latent tokens by aligning them with offline dynamic targets, and the RL stage further optimizes latent trajectories with contrastive dynamic latent policy optimization.

</Dynamic> and resumes language decoding. The complete output sequence can be illustrated as:

$$\boldsymbol{y} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_s, \texttt{<Dynamic>}, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N, \texttt{</Dynamic>}, \boldsymbol{w}_{s+1}, \ldots, \boldsymbol{w}_{T_y}), \tag{1}$$

where $\boldsymbol{w}_i$ denotes text tokens and T_y denotes the number of text tokens. This design allows symbolic reasoning and dynamic latent imagination to be integrated within a single autoregressive framework.

3.2 Training Data Construction

As illustrated in Figure 3, we construct DLVR-SFT-100K for dynamic latent reasoning that covers both dynamic scene evolution and static visual understanding, without relying on external visual tools and existing annotated visual-CoT datasets, which may overlap in distribution with evaluation benchmarks. Our goal is to enable the model to acquire dynamic world knowledge from the original visual input and internalize it into its generation process. For dynamic VQA, we collect 40K physical interaction cases from VBVR [68], such as ball bouncing and domino collapse, and 58K open-scene dynamic cases from Twiff [48]. We use DeepSeek-V3.2 [44] to clean and reformat the data, extract key frame indices, and distill textual descriptions to focus on the underlying dynamic evolution. For static VQA, we use the 1.8K PerceptionTest subset of LLaVA-Video-Train [84] to construct first-frame-based VQA pairs, while still associating them with subsequent frames as temporally extended visual context. In all cases, the inserted dynamic information is enclosed by <Dynamic></Dynamic>. This data provides two types of supervision: directly inserted frames teach the model coherent dynamic evolution, while their compressed representations supervise the latent tokens. Specifically, we introduce an offline supervision target construction algorithm that extracts discriminative dynamic cues from temporal frames and compresses the full dynamic process into fixed-length latent targets. For DLVR-RL-4K, we construct paired real and counterfactual samples from DualityVidQA [32]. Each pair shares a similar scene structure but differs in semantic or causal dynamics, providing structured negative evidence for penalizing counterfactual latent reasoning. We apply the same offline target construction pipeline to both real and counterfactual videos, producing fixed-length latent targets for contrastive latent rewards. More details on data construction are provided in Appendix D.

3.3 Offline Supervision Target Construction Algorithm

In DLVR, the first video frame serves as the visual input for question answering, while subsequent frames provide dynamic content that the model should imagine and predict in latent space. Since

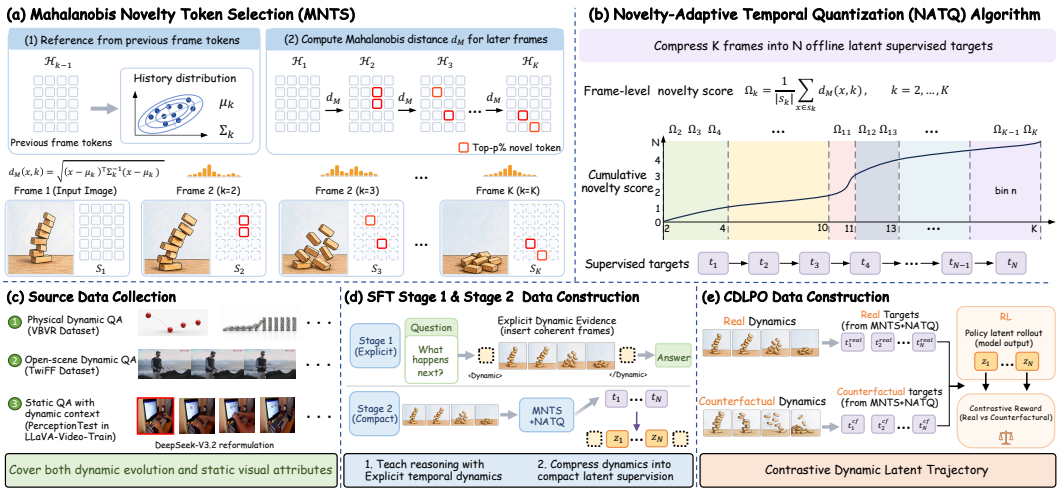


Figure 3: Illustration of the training data construction algorithm and pipeline.

visual tokens from later frames are often highly redundant, we design a two-stage algorithm to extract compact and informative supervision signals. We first apply *Mahalanobis Novelty Token Selection* (MNTS) to select informative tokens from dense temporal frames, and then use *Novelty-Adaptive Temporal Quantization* (NATQ) to compress them into a fixed number of supervision targets matching the latent-token budget. This enables the model to learn subsequent scene dynamics more effectively with limited latent tokens. The two-stage algorithm is detailed below.

Mahalanobis Novelty Token Selection. As shown in Figure 3a, we feed the K frames into the vision encoder and pass its final-layer hidden states through the model's built-in projector to obtain the visual representations $\{\mathcal{H}_k\}_{k=1}^{K}$ in the LLM embedding space, where $\mathcal{H}_k = \{\boldsymbol{h}_{k,1}^{\mathrm{VE}}, \cdots, \boldsymbol{h}_{k,S}^{\mathrm{VE}}\}$ and S indicates the number of visual tokens extracted from frame k. For $k \geq 2$, the history Gaussian distribution of frame k is defined as

$$\boldsymbol{\mu}_k = \frac{1}{S}\sum_{i=1}^{S} \boldsymbol{h}_{k-1,i}^{\mathrm{VE}}, \quad \boldsymbol{\Sigma}_k = \frac{1}{S-1}\sum_{i=1}^{S} (\boldsymbol{h}_{k-1,i}^{\mathrm{VE}} - \boldsymbol{\mu}_k)(\boldsymbol{h}_{k-1,i}^{\mathrm{VE}} - \boldsymbol{\mu}_k)^{\top}. \tag{2}$$

For each token $\boldsymbol{x} \in \mathcal{H}_k$, we compute its Mahalanobis distance to the history Gaussian distribution:

$$d_M(\boldsymbol{x}, k) = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu}_k)^{\top} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)}. \tag{3}$$

We select the top-$p\%$ tokens with the largest Mahalanobis distances, with $p = 10$, and form the novelty set $\mathcal{S}_k = \mathrm{Top}_{p\%}\{d_M(\boldsymbol{x}, k) \mid \boldsymbol{x} \in \mathcal{H}_k\}$ for frame k to capture the most distinctive dynamic cues. MNTS also admits an information-theoretic interpretation as maximizing the information preserved by the selected tokens, as detailed in Appendix B.1.

Novelty-Adaptive Temporal Quantization. As illustrated in Figure 3b, after selecting novelty tokens, we compress the full tokens of K frames into N latent supervision targets. As dynamic changes are often unevenly distributed over time, uniform temporal token budget partitioning is usually not optimal and we assign larger budget for intervals with stronger novelty. To this end, we compute a frame-level novelty score Ω_k as the average Mahalanobis distance of the selected tokens $\mathcal{S}_k$ to the historical Gaussian distribution, and construct a frame-level representation $\boldsymbol{v}_k$ via a novelty-weighted average of these tokens to capture the global information of frame k:

$$\Omega_k = \frac{1}{|\mathcal{S}_k|}\sum_{\boldsymbol{x}\in\mathcal{S}_k} d_M(\boldsymbol{x}, k), \quad \boldsymbol{v}_k = \frac{\sum_{\boldsymbol{x}\in\mathcal{S}_k} d_M(\boldsymbol{x}, k)\cdot \boldsymbol{x}}{\sum_{\boldsymbol{x}\in\mathcal{S}_k} d_M(\boldsymbol{x}, k)}, \quad k = 2, \cdots, K. \tag{4}$$

Next, we accumulate novelty over time as $\bar{\Omega}_k = \sum_{j=2}^{k} \Omega_j$ and $\Omega_{\mathrm{total}} = \bar{\Omega}_K$. Each frame is assigned to one of N bins according to its cumulative novelty ratio $\mathrm{bin}(k)$, and we compute the latent target $\boldsymbol{t}_n$ corresponding to bin n as the novelty-weighted average of frame-level representations:

$$\boldsymbol{t}_n = \frac{\sum_{k:\mathrm{bin}(k)=n-1} \Omega_k \boldsymbol{v}_k}{\sum_{k:\mathrm{bin}(k)=n-1} \Omega_k}, \quad \text{where} \quad \mathrm{bin}(k) = \min\left(\left\lfloor N \cdot \frac{\bar{\Omega}_k}{\Omega_{\mathrm{total}}} \right\rfloor, N-1\right). \tag{5}$$

This offline process produces N target vectors $\{\boldsymbol{t}_n\}_{n=1}^{N}$, which supervise the latent sequence generated inside the <Dynamic></Dynamic> segment during SFT Stage 2 and RL Stage.

3.4 SFT Stage 1: Dynamic Temporal Grounding

As shown in Figure 2, SFT Stage 1 aims to teach the model how to 1) use dynamic visual evidence in answer generation, 2) preserve temporal proximity among frames, and 3) capture the forward direction of scene evolution. For each frame k, let $\bar{\boldsymbol{h}}_k$ denote the average-pooled representation of its vision tokens, and let $\bar{\boldsymbol{e}}_k$ be the corresponding average-pooled input-layer representation, which serves as a stable reference from the frozen visual encoder. In this stage, we use the standard next-token prediction loss $\mathcal{L}_{\mathrm{CE}}$ for maintaining language reasoning and answer generation. To encode temporal proximity, we introduce *Temporal Contrastive Calibration* (TCC) loss, which encourages tokens of an anchor frame to be closer to those of temporally nearby frames than to distant ones:

$$\mathcal{L}_{\mathrm{TCC}} = \frac{1}{|\mathcal{T}|} \sum_{(i,j,k)\in\mathcal{T}} \max\Big(\cos\big(\bar{\boldsymbol{h}}_i, \bar{\boldsymbol{h}}_k\big) - \cos\big(\bar{\boldsymbol{h}}_i, \bar{\boldsymbol{h}}_j\big) + m, 0\Big), \tag{6}$$

where $\mathcal{T} = \{(i,j,k) : |i-j| < |i-k|\}$ and $m = 1/(K-1)$. Since temporal proximity alone is symmetric and cannot distinguish the directionality of evolution, we further introduce *Temporal Direction Consistency* (TDC) loss with respect to the visual encoder:

$$\mathcal{L}_{\mathrm{TDC}}^{\mathrm{VE}} = \frac{1}{K-1} \sum_{k=1}^{K-1} \Big[1 - \cos\big(\bar{\boldsymbol{h}}_{k+1} - \bar{\boldsymbol{h}}_k, \bar{\boldsymbol{e}}_{k+1} - \bar{\boldsymbol{e}}_k\big) \Big]. \tag{7}$$

The TDC loss aligns the transition direction in the model representation with that of the frozen visual encoder, providing an explicit temporal direction anchor. The overall objective of SFT Stage 1 is $\mathcal{L}_{\mathrm{S1}} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{TCC}} + \mathcal{L}_{\mathrm{TDC}}^{\mathrm{VE}}$, which equips the model with temporal grounding ability for subsequent dynamic latent reasoning. The detailed derivation is provided in Appendix B.2.

3.5 SFT Stage 2: Dynamic Latent Reasoning

The model is expected to reason with input temporal image sequences after Stage 1. As shown in Figure 2 (c), SFT Stage 2 removes this dependence by training the model to autoregressively generate compact dynamic latent tokens inside the <Dynamic></Dynamic> segment. These tokens are expected to encode dynamic semantics, temporal structure, and evolution direction, and then support reasoning. Using MNTS and NATQ, we construct offline target vectors $\{\boldsymbol{t}_n\}_{n=1}^{N}$ from the full dynamic sequence. In this stage, we still use $\mathcal{L}_{\mathrm{CE}}$ to preserve standard language reasoning and answer generation. To directly distill dynamic information into latent tokens, we use an alignment loss

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{N} \sum_{n=1}^{N} \Big[1 - \cos\big(\boldsymbol{z}_n, \boldsymbol{t}_n\big) \Big], \tag{8}$$

which anchors each generated latent token to its corresponding target. Since token-wise alignment alone may miss the global structure of the target tokens, we impose a structural preservation loss

$$\mathcal{L}_{\mathrm{struct}} = \frac{1}{\binom{N}{2}} \sum_{1\le i<j\le N} \Big[\cos\big(\boldsymbol{z}_i, \boldsymbol{z}_j\big) - \cos\big(\boldsymbol{t}_i, \boldsymbol{t}_j\big) \Big]^2. \tag{9}$$

This loss imposes a relational constraint by preserving pairwise similarities among latent tokens. Finally, to encode temporal order, we introduce a temporal direction consistency loss

$$\mathcal{L}_{\mathrm{TDC}} = \frac{1}{N-1} \sum_{n=1}^{N-1} \Big[1 - \cos\big(\boldsymbol{z}_{n+1} - \boldsymbol{z}_n, \boldsymbol{t}_{n+1} - \boldsymbol{t}_n\big) \Big], \tag{10}$$

which imposes a temporal constraint by aligning adjacent transition directions between $\boldsymbol{z}$ and $\boldsymbol{t}$. The overall objective is $\mathcal{L}_{\mathrm{S2}} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{align}} + \mathcal{L}_{\mathrm{struct}} + \mathcal{L}_{\mathrm{TDC}}$, which encourages the latent sequence to match the target dynamics at the token, structural, and directional levels. As a result, the model learns to perform dynamic reasoning through compact latent tokens without relying on auxiliary images.

3.6 RL Stage: Contrastive Dynamic Latent Policy Optimization

After the SFT stages, DLVR can generate dynamic latent trajectories, but these trajectories are not explicitly optimized for dynamic plausibility. Existing methods mainly reward final-answer correctness, leaving the latent reasoning process weakly supervised. To address this, we propose *Contrastive Dynamic Latent Policy Optimization* (CDLPO), which incorporates latent trajectories into the reward design. As illustrated in Figure 2 (d), CDLPO encourages generated latent tokens to align with real dynamics while contrasting them against counterfactual dynamics. For each training instance, we sample G trajectories $\{\boldsymbol{o}_i\}_{i=1}^{G}$ to estimate relative preferences and advantages. Since the latent segment consists of continuous tokens, we parameterize each latent token with a Gaussian distribution and sample from it during rollout [70]. This enables exploration of diverse latent trajectories in the continuous representation space. For rollout i, let a_i denote the predicted answer and $\{\boldsymbol{z}_n^{(i)}\}_{n=1}^{N}$ denote the generated latent sequence. We define the answer reward as $R_i^{\mathrm{acc}} = \mathbb{I}[a_i = a_{\mathrm{real}}]$, and further introduce a contrastive latent reward to supervise the latent reasoning process:

$$R_i^{\mathrm{cf}} = \frac{1}{N}\sum_{n=1}^{N}\left(\cos\left(\boldsymbol{z}_n^{(i)}, \boldsymbol{t}_n^{\mathrm{real}}\right) - \cos\left(\boldsymbol{z}_n^{(i)}, \boldsymbol{t}_n^{\mathrm{cf}}\right)\right). \tag{11}$$

Here, $\boldsymbol{t}^{\mathrm{real}}$ and $\boldsymbol{t}^{\mathrm{cf}}$ denote precomputed supervision targets from real and counterfactual videos. The contrastive reward R_i^{cf} encourages the generated latent trajectory to approach real dynamics and move away from counterfactual ones. The final reward is $R_i = R_i^{\mathrm{acc}} + \beta R_i^{\mathrm{cf}}$. Unlike standard GRPO, we compute both advantages by group-wise normalization over the G rollouts. The outcome advantage A_i^{out} is obtained by normalizing the total reward R_i, while the process advantage A_i^{proc} is obtained by normalizing the contrastive dynamic reward R_i^{cf}. The token-wise advantage $A_{i,t} = A_i^{\mathrm{out}} + \lambda A_i^{\mathrm{proc}}$ when token t is a latent token, and $A_{i,t} = A_i^{\mathrm{out}}$ otherwise. This design ensures that text tokens primarily respond to answer correctness, while latent tokens receive additional supervision on dynamic plausibility. DLVR is optimized through the objective

$$\mathcal{J}(\theta) = \frac{1}{\sum_i |\boldsymbol{o}_i|}\sum_{i=1}^{G}\sum_{t=1}^{|\boldsymbol{o}_i|}\min\Big(r_{i,t}(\theta)A_{i,t},\ \mathrm{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)A_{i,t}\Big), \tag{12}$$

where $r_{i,t}(\theta) = \frac{\pi_\theta(\boldsymbol{o}_{i,t}|\boldsymbol{I},\boldsymbol{q},\boldsymbol{o}_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(\boldsymbol{o}_{i,t}|\boldsymbol{I},\boldsymbol{q},\boldsymbol{o}_{i,<t})}$, π_θ denotes the policy distribution to optimize, and $\pi_{\theta_{\mathrm{old}}}$ is the old rollout policy. For latent tokens, the likelihood in $r_{i,t}(\theta)$ is computed under the Gaussian distribution; for text tokens, it is given by the softmax probabilities. We further include a KL regularization term $D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})$ to constrain policy deviation with the frozen reference policy π_{ref}. The final objective of RL Stage is $\mathcal{L}(\theta) = -\mathcal{J}(\theta) + \xi\, D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})$.

4 Experiment

Baselines. We include representative methods from both *Thinking With Images* and *Latent Visual Reasoning*. Specifically, we compare against VISER [33] and ICoT [26] in the former category, and against CoVT [53], LVR [36], and Monet [70] in the latter category, where CoVT enables reasoning with latent visual tokens distilled from multiple vision experts, LVR focuses latent reasoning on the core regions of the input image, Monet aligns latent token with auxiliary image representations, and further uses observation tokens as supervision signals derived from auxiliary images. Following these works, we apply DLVR on Qwen2.5-VL-7B [5]. More details can be found in Appendix C.2.

Benchmarks. We evaluate DLVR on two task categories. *Dynamic-Centric Image Understanding Tasks* include VTBench [63], Visual Cognitive Tasks [56], and BabyVision [12]. *General Image Understanding Tasks* include MMBench [49], MMStar [13], MMVet [81], V* [76], HRBench4K/8K [72], MME-RealWorld [83], and SpatialEval [67].

Hyperparameters. All models are trained and evaluated on 8 $\times$ A100 GPUs. We use *DLVR-SFT-100K* for SFT, with Stages 1 and 2 trained for 4 and 2 epochs, respectively. For CDLPO, we train DLVR on *DLVR-RL-4K* for 1 epoch. The latent size is 16 by default for both training and evaluation. We use a learning rate of 1×10^{-5} and a weight decay of 0.01. We additionally introduce the special tokens `<Dynamic>` and `</Dynamic>` into the vocabulary. Details can be found in Appendix C.3.

Table 1: **Performance comparison on Dynamic-Centric Image Understanding Tasks.** Blue denotes improvement, gray denotes degradation, the best-performing results are highlighted in bold.

Model	BabyVision					Visual Cognitive Tasks			VTBench	
	Fine-Grained	Vis. Tracking	Spatial Perc.	Vis. Pattern	Overall	Block Tower	Michotte	Astronaut	Eyeballing	Vis. Puzzles
Qwen2.5-VL-7B [5]	6.13 (0.00)	9.64 (0.00)	15.38 (0.00)	17.65 (0.00)	10.57 (0.00)	46.37 (0.00)	16.46 (0.00)	20.38 (0.00)	11.42 (0.00)	34.87 (0.00)
Think With Image										
+VISER [33](NeurIPS'25)	7.36 (1.23)	6.02 (3.62)	16.48 (1.10)	23.53 (5.88)	11.34 (0.77)	47.32 (0.95)	16.33 (0.13)	19.87 (0.51)	11.90 (0.48)	36.08 (1.21)
+ICoT [26](CVPR'25)	10.43 (4.30)	8.43 (1.21)	10.99 (4.39)	21.56 (3.91)	11.60 (1.03)	48.91 (2.54)	17.04 (0.58)	20.11 (0.27)	12.38 (0.96)	34.27 (0.60)
Latent Reasoning Method										
+CoVT [53](arXiv'25)	11.04 (4.91)	10.84 (1.20)	14.29 (1.09)	15.69 (1.96)	12.37 (1.80)	49.11 (2.74)	15.80 (0.66)	22.89 (2.51)	13.33 (1.91)	36.69 (1.82)
+LVR [36](ICLR'26)	9.82 (3.69)	12.05 (2.41)	18.68 (3.30)	17.66 (0.01)	13.40 (2.83)	51.07 (4.70)	18.59 (2.13)	22.31 (1.93)	13.80 (2.38)	36.89 (2.02)
+Monet [70](CVPR'26)	10.42 (4.29)	13.25 (3.61)	15.38 (0.00)	21.57 (3.92)	13.91 (3.34)	50.94 (4.57)	17.73 (1.27)	23.60 (3.22)	13.91 (2.49)	37.50 (2.63)
+DLVR (Ours)	**16.56** (10.43)	**18.07** (8.43)	**26.37** (10.99)	**33.33** (15.68)	**21.39** (10.82)	**57.79** (11.42)	**22.15** (5.69)	**30.03** (9.65)	**23.34** (11.92)	**42.39** (7.52)

Table 2: **Performance comparison on General Image Understanding Tasks.**

Model	Comprehensive				RealWorld			Spatial
	MMBench	MMStar	MMVet	V*	HRBench4K	HRBench8K	MME-RealWorld	SpatialEval
Qwen2.5-VL-7B [5]	81.16 (0.00)	61.18 (0.00)	68.95 (0.00)	76.38 (0.00)	64.13 (0.00)	60.25 (0.00)	42.97 (0.00)	59.20 (0.00)
Think With Image								
+VISER [33](NeurIPS'25)	82.06 (0.90)	64.29 (3.11)	70.28 (1.33)	78.15 (1.77)	65.13 (1.00)	60.75 (0.50)	42.40 (0.57)	61.23 (2.03)
+ICoT [26](CVPR'25)	82.18 (1.02)	63.87 (2.69)	70.65 (1.70)	79.79 (3.41)	66.62 (2.49)	61.00 (0.75)	43.36 (0.39)	59.54 (0.34)
Latent Reasoning Method								
+CoVT [53](arXiv'25)	81.82 (0.66)	63.75 (2.57)	72.49 (3.54)	78.60 (2.22)	67.88 (3.75)	65.13 (4.88)	46.24 (3.27)	62.42 (3.22)
+LVR [36](ICLR'26)	81.53 (0.37)	64.21 (3.03)	72.85 (3.90)	80.94 (4.56)	68.00 (3.87)	63.75 (3.50)	49.66 (6.69)	62.50 (3.30)
+Monet [70](CVPR'26)	82.90 (1.74)	62.56 (1.38)	73.67 (4.72)	83.11 (6.73)	67.38 (3.25)	64.38 (4.13)	51.70 (8.73)	61.94 (2.74)
+DLVR (Ours)	**86.01** (4.85)	**65.53** (4.35)	**75.83** (6.88)	**85.34** (8.96)	**74.63** (10.50)	**68.75** (8.50)	**52.53** (9.56)	**67.53** (8.33)

4.1 Main Results

Dynamic-Centric Image Understanding Tasks. As shown in Table 1, DLVR delivers consistent and substantial improvements over the Qwen2.5-VL-7B baseline and all compared reasoning methods. Notably, on BabyVision, DLVR raises the overall score from 10.57% to 21.39%, achieving a remarkable absolute gain of 10.82%. DLVR also achieves clear gains on Visual Cognitive Tasks, improving Block Tower from 46.37% to 57.79%. On VTBench, DLVR further improves Eyeballing by 11.92% and Visual Puzzles by 7.52%. These results demonstrate that modeling dynamic evolution in latent visual space provides a strong and generalizable advantage for image understanding tasks that require physical, causal, and future-oriented reasoning.

General Image Understanding Tasks. Table 2 shows that DLVR not only excels on dynamic-centric tasks, but also further improves general image understanding. For Comprehensive benchmarks, the most notable gain appears on V*, where DLVR improves the baseline from 76.38% to 85.34%, yielding an 8.96% increase. On RealWorld tasks, DLVR achieves a substantial improvement on HRBench4K, raising the score from 64.13% to 74.63%. The advantage also extends to spatial reasoning, with SpatialEval increasing from 59.20% to 67.53%. These results suggest that dynamic latent reasoning provides a broadly useful visual reasoning signal, benefiting not only dynamic perception but also general, real-world, and spatial image understanding. Detailed experimental results are provided in Appendix E.1. Qualitative visualization results are provided in Appendix F.

4.2 Ablation and Analysis

Ablation of the Main Components. Table 3 validates the effectiveness of each key component in DLVR. Both the vanilla SFT and vanilla SFT+GRPO baselines are trained on the same training data as DLVR. Vanilla SFT brings only moderate gains, while adding GRPO yields limited further improvement. In contrast, the full DLVR achieves substantial boosts across all benchmarks (e.g., BabyVision from 10.57% to 21.39%). Removing CDLPO consistently weakens performance, while removing DLVR-SFT Stage2 causes a larger drop but still outperform baseline. Overall, these results demonstrate that both the staged SFT design and CDLPO are essential to DLVR's strong performance.

Table 3: Ablation on key components of DLVR.

Model	BabyVision	VTBench	V*	HRBench8K	SpatialEval
Qwen2.5-VL-7B	10.57	19.03	76.38	60.25	59.20
+ vanilla SFT	12.88	22.12	76.43	61.50	61.17
+ vanilla SFT + GRPO	12.62	22.63	76.96	62.00	61.90
DLVR	**21.39**	**29.43**	**85.34**	**68.75**	**67.53**
w/o CDLPO	18.04	28.13	82.53	66.87	66.01
w/o DLVR-SFT Stage2	14.18	23.93	78.53	62.88	62.39

Table 4: Ablation on loss functions of DLVR.

Model	BabyVision	VTBench	V*	HRBench8K	SpatialEval
DLVR	**21.39**	**29.43**	**85.34**	**68.75**	**67.53**
SFT Stage 1					
w/o $\mathcal{L}_{\text{TCC}}$	18.81	28.01	82.72	66.75	64.27
w/o $\mathcal{L}_{\text{TDC-VE}}$	19.07	27.81	82.19	66.88	64.40
SFT Stage 2					
w/o $\mathcal{L}_{\text{struct}}$	18.29	27.49	81.15	66.13	63.52
w/o $\mathcal{L}_{\text{TDC}}$	18.04	25.94	80.10	65.63	63.81
w/o $\mathcal{L}_{\text{align}}$	17.53	24.64	80.63	65.38	63.72
RL Stage					
w/o Contrastive Latent Reward	19.33	28.13	83.62	67.75	65.80

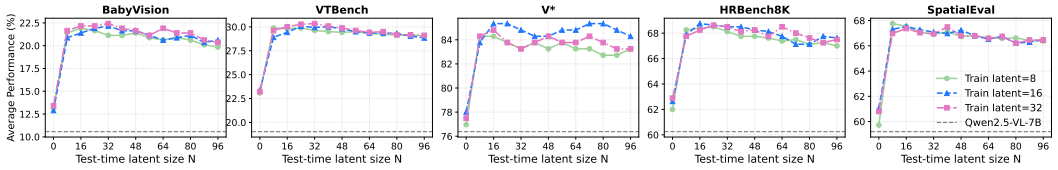


Figure 4: Ablation on the number of latent tokens in training and testing time.

Ablation of the Loss Function. Table 4 studies the contribution of each training loss in DLVR. Each time we remove a single loss while keeping the other losses unchanged. In SFT Stage 1, removing $\mathcal{L}_{\text{TCC}}$ or $\mathcal{L}_{\text{TDC-VE}}$ both lead to clear performance drops on BabyVision and V*. Removing $\mathcal{L}_{\text{align}}$ in SFT Stage 2 causes the largest decline, reducing HRBench8K score from 68.75% to 65.38%. In the RL stage, removing the contrastive latent reward also weakens performance across all benchmarks, confirming that contrastive optimization further aligns latent imagination with reliable dynamic evolution. Overall, these results verify that DLVR's gains come from the coordinated design of temporal supervision, latent structure learning, and contrastive latent reward optimization.

Number of Latent Tokens. Figure 4 shows that latent size $N = 16$ for both training and testing provides a balance between performance and training cost. The sharp gain after enabling latent tokens indicates that they carry the key visual imagination process behind improvements. Further increasing the test latent size can benefit some dynamic-centric tasks, suggesting that richer latent imagination helps capture more complex temporal and physical evolution. Following prior latent reasoning works [36, 70, 79], we use a simple and effective fixed-length setting, while leaving adaptive latent size as a promising direction for future research. Detailed results and analysis are provided in Appendix E.2.

Different Model Scales. We further apply our framework to Qwen2.5-VL-3B and compare it with the 7B counterpart. As shown in Table 5, DLVR consistently improves both backbones across all benchmarks. These results show that dynamic latent visual reasoning is not tied to a specific model size, but provides a transferable benefit across model scales.

Table 5: Performance comparison across different model scales.

Model	BabyVision	VTBench	V*	HRBench8K	SpatialEval
Qwen2.5-VL-7B	10.57	19.03	76.38	60.25	59.20
DLVR-7B	**21.39**	**29.43**	**85.34**	**68.75**	**67.53**
Qwen2.5-VL-3B	6.70	13.19	70.16	55.86	50.11
DLVR-3B	**11.86**	**21.02**	**77.49**	**62.63**	**58.39**

Different Model Backbones. We further apply our framework to three additional backbones, including Qwen3-VL-8B, InternVL3.5-8B, and LLaVA-1.5-7B. These models differ in their architectures, visual representations, and baseline capabilities, allowing us to assess the generality of DLVR across diverse backbone configurations. As shown in Table 6, DLVR consistently improves all three additional backbones across all evaluated benchmarks. For example, BabyVision scores increase by 8.76, 10.30, and 5.16 points on the three backbones, demonstrating substantial gains. These results show that dynamic latent visual reasoning is not specific to Qwen2.5-VL, but provides consistent benefits across different LVLM families.

Table 6: Performance comparison across different LVLM backbones.

Model	BabyVision	VTBench	V*	HRBench8K	SpatialEval
Qwen2.5-VL-7B	10.57	19.03	76.38	60.25	59.20
+DLVR	**21.39**	**29.43**	**85.34**	**68.75**	**67.53**
Qwen3-VL-8B	11.34	22.51	78.53	61.87	64.05
+DLVR	**20.10**	**31.05**	**86.39**	**71.13**	**69.19**
InternVL3.5-8B	11.86	29.35	70.25	65.52	58.00
+DLVR	**22.16**	**36.29**	**79.06**	**74.86**	**65.54**
LLaVA-1.5-7B	5.15	8.54	39.79	29.80	32.58
+DLVR	**10.31**	**17.98**	**47.12**	**34.36**	**38.81**

Table 7: Generalization performance on benchmarks beyond the training distribution.

Model	MMIQ				OpenWorldQA	CausalSpatial
	Logical	Object	Spatial	Temporal		
Qwen2.5-VL-7B [5]	25.70 (0.00)	23.68 (0.00)	21.19 (0.00)	22.89 (0.00)	32.20 (0.00)	43.48 (0.00)
+vanilla SFT+GRPO	25.95 (0.25)	26.32 (2.64)	23.73 (2.54)	24.10 (1.21)	34.20 (2.00)	46.54 (3.06)
+ICoT [26](CVPR'25)	26.21 (0.51)	28.95 (5.27)	18.64 (−2.55)	23.13 (0.24)	32.40 (0.20)	44.76 (1.28)
+Monet [70](CVPR'26)	28.50 (2.80)	28.95 (5.27)	24.58 (3.39)	25.54 (2.65)	35.40 (3.20)	47.04 (3.56)
+DLVR (Ours)	**29.26** (3.56)	**36.84** (13.16)	**27.74** (6.55)	**30.36** (7.47)	**40.80** (8.60)	**52.07** (8.59)

Generalization Beyond the Training Distribution. We further assess whether DLVR generalizes beyond the dynamics observed in its training videos by evaluating it on MMIQ [9], OpenWorldQA [88], and CausalSpatial [51]. These benchmarks are independent of the training sources and cover diverse abstract and physical reasoning tasks. As shown in Table 7, DLVR achieves the best performance among the compared methods across all four MMIQ categories and both other benchmarks. Its consistent advantage over vanilla SFT+GRPO trained on the same DLVR data without latent supervision indicates that the training data alone does not fully explain the improvements. Together with the gains across the eight general benchmarks, these results support the transferability of dynamic latent visual reasoning beyond the specific dynamics observed during training.

Ablation of MNTS and NATQ. We further evaluate the design choices of MNTS and NATQ by replacing them with random token sampling and uniform temporal compression, respectively. As shown in Table 8, both variants outperform the backbone but consistently underperform the DLVR across all benchmarks. The degradation under uniform compression supports the benefit of allocating latent targets according to temporal novelty. Random sampling leads to a larger performance drop, highlighting the importance of explicitly selecting tokens that capture novel dynamics. These results support the effectiveness of both modules.

Table 8: Comparison with simpler alternatives to MNTS and NATQ.

Model	BabyVision	VTBench	V*	HRBench8K	SpatialEval
Qwen2.5-VL-7B	10.57	19.03	76.38	60.25	59.20
+DLVR	**21.39**	**29.43**	**85.34**	**68.75**	**67.53**
NATQ → uniform	18.56	27.04	81.67	66.13	65.82
MNTS → random	16.24	24.13	80.10	65.00	63.49

4.3 Can DLVR Really Foresee Plausible Dynamics?

In the Table 9, DLVR significantly outperforms the static baseline and narrows the gap with the full input upper bound, demonstrating that its dynamic latent space can capture essential temporal cues and support reasoning over plausible scene evolution even without direct temporal observations, enabling it to bridge the gap between static visual understanding and the dynamic reasoning. Limited gain with $N = 0$ further confirms that dynamic latent tokens mainly drive the improvement.

Table 9: Comparison on Video-MME [24].

Model	Know.	Film	Sports	Art	Life	Multi.	Overall
Full Video Input							
Qwen2.5-VL-7B	63.33	65.00	60.22	63.06	57.78	53.33	61.37
Static Input (First Frame Only)							
Qwen2.5-VL-7B	33.09	31.67	23.33	41.67	31.11	30.00	31.85
DLVR(N=0)	36.42	33.06	25.56	41.39	31.90	31.11	33.59
DLVR	**46.91**	**46.11**	**42.00**	**46.67**	**42.38**	**44.44**	**44.81**

5 Conclusion

Inspired by human, we address a key limitation of LVLMs in complex visual reasoning, namely their lack of internal dynamic imagination from static visual input. We propose **Dynamic Latent Visual Reasoning (DLVR)**, a training framework that enables models to autoregressively infer future scene evolution in continuous latent space from a static image. Through dynamic and static training data, a two-stage SFT pipeline, and CDLPO with real-versus-counterfactual dynamic constraints, DLVR advances latent visual reasoning from static representation modeling to temporal and future-oriented reasoning. Extensive experiments show consistent gains on both dynamic-centric and general visual reasoning benchmarks, unleashing LVLMs thinking in dynamic latent space and demonstrating the potential of dynamic latent reasoning for building more capable visual reasoning systems.

Acknowledgments

This work is supported by the Lingang Laboratory (Grant No. LGL-1987-10) and the General Program of the National Natural Science Foundation of China (Grant No. 62576110).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

[4] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[5] 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.

[6] Halely Balaban and Tomer D Ullman. The capacity limits of moving objects in the imagination. *Nature Communications*, 16(1):5899, 2025.

[7] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the national academy of sciences*, 110(45):18327–18332, 2013.

[8] Johannes Bill, Samuel J Gershman, and Jan Drugowitsch. Visual motion perception as online hierarchical inference. *Nature communications*, 13(1):7403, 2022.

[9] Huanqia Cai, Yijun Yang, and Winston Hu. Mm-iq: Benchmarking human-like abstraction and reasoning in multimodal models. *arXiv preprint arXiv:2502.00698*, 2025.

[10] Kartik Chandra, Tony Chen, Tzu-Mao Li, Jonathan Ragan-Kelley, and Josh Tenenbaum. Inferring the future by imagining the past. *Advances in Neural Information Processing Systems*, 36:21196–21216, 2023.

[11] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[12] Liang Chen, Weichu Xie, Yiyan Liang, Hongfeng He, Hans Zhao, Zhibo Yang, Zhiqi Huang, Haoning Wu, Haoyu Lu, Yiping Bao, et al. Babyvision: Visual reasoning beyond language. *arXiv preprint arXiv:2601.06521*, 2026.

[13] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

[14] Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. MINT-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[15] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024.

[16] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[17] Jiwan Chung, Junhyeok Kim, Siyeol Kim, Jaeyoung Lee, Min Soo Kim, and Youngjae Yu. Don't look only once: Towards multimodal interactive reasoning with selective visual revisitation. *arXiv e-prints*, pages arXiv–2505, 2025.

[18] Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, pages 49250–49267, 2023.

[19] Ingmar EJ de Vries and Moritz F Wurm. Predictive neural representations of naturalistic dynamic input. *Nature Communications*, 14(1):3858, 2023.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[21] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[22] Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Xinze Guan, and Xin Eric Wang. GRIT: Teaching MLLMs to think with images. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[23] Jason Fischer, John G Mikhael, Joshua B Tenenbaum, and Nancy Kanwisher. Functional neuroanatomy of intuitive physical inference. *Proceedings of the national academy of sciences*, 113(34):E5072–E5081, 2016.

[24] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, 2025.

[25] Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Richard Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. In *Forty-second International Conference on Machine Learning*, 2025.

[26] Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19520–19529, 2025.

[27] Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Kuan Li, Yida Zhao, Huifeng Yin, Yong Jiang, Pengjun Xie, Fei Huang, Huaxiu Yao, Yi R. Fung, and Jingren Zhou. Webwatcher: Breaking new frontiers of vision-language deep research agent. In *The Fourteenth International Conference on Learning Representations*, 2026.

[28] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14953–14962, 2023.

[29] Zhitao He, Sandeep Polisetty, Zhiyuan Fan, Yuchen Huang, Shujin Wu, and Yi R Fung. Mmboundary: Advancing mllm knowledge boundary awareness through reasoning step confidence calibration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16427–16444, 2025.

[30] Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Fei zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

[31] Zeyi Huang, Yuyang Ji, Anirudh Sundara Rajan, Zefan Cai, Wen Xiao, Haohan Wang, Junjie Hu, and Yong Jae Lee. Visualtoolagent (vista): A reinforcement learning framework for visual tool selection. *arXiv preprint arXiv:2505.20289*, 2025.

[32] Zhe Huang, Hao Wen, Aiming Hao, Bingze Song, Meiqi Wu, Jiahong Wu, Xiangxiang Chu, Sheng Lu, and Haoqian Wang. Taming hallucinations: Boosting mllms' video understanding via counterfactual video generation. *arXiv preprint arXiv:2512.24271*, 2025.

[33] Amirmohammad Izadi, Mohammadali Banayeeanzade, Fatemeh Askari, Ali Rahimiakbar, Mohammad Mahdi Vahedi, Hosein Hasani, and Mahdieh Soleymani Baghshah. Visual structures help visual reasoning: Addressing the binding problem in LVLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[34] Zoe Kourtzi and Nancy Kanwisher. Activation in human mt/mst by static images with implied motion. *Journal of Cognitive Neuroscience*, 12(1):48–55, 2000. doi: 10.1162/08989290051137594.

[35] Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 71683–71702. Curran Associates, Inc., 2023.

[36] Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[38] Xu Li, Yuxuan Liang, Xiaolei Chen, Yi Zheng, Haotian Chen, Zhe Liu, Rui Zhu, Bin Li, and Xiangyang Xue. Hero: Rethinking visual token early dropping in high-resolution large vision-language models. *IEEE Transactions on Multimedia*, pages 1–15, 2026. doi: 10.1109/TMM.2026.3724758.

[39] Xu Li, Yi Zheng, Haotian Chen, Xiaolei Chen, Yuxuan Liang, Chenghang Lai, Bin Li, and Xiangyang Xue. Instruction-guided fusion of multi-layer visual features in large vision-language models. *Pattern Recognition*, 170:111932, 2026. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2025.111932.

[40] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *CVPR*, pages 26763–26773, 2024.

[41] Yuxuan Liang, Xu Li, Xiaolei Chen, Haotian Chen, Yi Zheng, Chenghang Lai, Bin Li, and Xiangyang Xue. Global semantic-guided sub-image feature weight allocation in high-resolution large vision-language models. *arXiv preprint arXiv:2501.14276*, 2025.

[42] Yuxuan Liang, Xu Li, Xiaolei Chen, Haotian Chen, Yi Zhen, Zhe Liu, Rui Zhu, Bin Li, and Xiangyang Xue. Pyramid token pruning for high-resolution large vision-language models via region, token, and instruction-guided importance. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2026. doi: 10.1109/TCSVT.2026.3663658.

[43] Yuxuan Liang, Fan Shi, Rui Zhu, Xu Li, Xiaolei Chen, Zhe Liu, Bin Li, and Xiangyang Xue. Envision, attend, then respond: Counterfactual hallucination mitigation in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18261–18272, June 2026.

[44] Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

[45] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, pages 34892–34916, 2023.

[46] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[47] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[48] Junhua Liu, Zhangcheng Wang, Zhike Han, Ningli Wang, Guotao Liang, and Kun Kuang. Twiff (think with future frames): A large-scale dataset for dynamic visual reasoning. *arXiv preprint arXiv:2602.10675*, 2026.

[49] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[50] Ruilin Luo, Zhuofan Zheng, Lei Wang, Yifan Wang, Xinzhe Ni, Zicheng Lin, Songtao Jiang, Yiyao Yu, Chufan Shi, Ruihang Chu, Jin zeng, and Yujiu Yang. Unlocking multimodal mathematical reasoning via process reward model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[51] Wenxin Ma, Chenlong Wang, Ruisheng Yuan, Hao Chen, Nanru Dai, S Kevin Zhou, Yijun Yang, Alan Yuille, and Jieneng Chen. Causalspatial: A benchmark for object-centric causal spatial reasoning. *arXiv preprint arXiv:2601.13304*, 2026.

[52] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, pages 508–524. Springer, 2020.

[53] Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[55] Gabriel Herbert Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[56] Luca M Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 7(1):96–106, 2025.

[57] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[58] Fan Shi, Yuxuan Liang, Xiaolei Chen, Haiyang Yu, Xu Li, Yi Zheng, Rui Zhu, Xiangyang Xue, and Bin Li. Decomposition of concept-level rules in visual scenes. In *The Fourteenth International Conference on Learning Representations*, 2026.

[59] Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[60] Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025.

[61] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025.

[62] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-RFT: Reinforcement fine-tuning for visual reasoning of vision language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[63] Jingqi Tong, Yurong Mou, Hangcheng Li, Mingzhe Li, Yongzhuo Yang, Ming Zhang, Qiguang Chen, Tianyi Liang, Xiaomeng Hu, Yining Zheng, et al. Thinking with video: Video generation as a promising multimodal reasoning paradigm. In *CVPR*, 2026.

[64] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[65] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[66] Jiacong Wang, Zijian Kang, Haochen Wang, LiangXiao, Ya Wang, Jiawen Li, Bohong Wu, Ran Jiao, Haiyong Jiang, ChaoFeng, and Jun Xiao. VGR: Visual grounded reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[67] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024.

[68] Maijunxian Wang, Ruisi Wang, Juyi Lin, Ran Ji, Thaddäus Wiedemer, Qingying Gao, Dezhi Luo, Yaoyao Qian, Lianyu Huang, Zelong Hong, et al. A very big video reasoning suite. *arXiv preprint arXiv:2602.20159*, 2026.

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

[70] Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond images and language. In *CVPR*, 2026.

[71] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.

[72] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7907–7915, 2025.

[73] Yikun Wang, Siyin Wang, Qinyuan Cheng, Zhaoye Fei, Liang Ding, Qipeng Guo, Dacheng Tao, and Xipeng Qiu. Visuothink: Empowering lvlm reasoning with multimodal tree search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21707–21719, 2025.

[74] Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru WANG, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[75] Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, ChengXiang Zhai, and Klara Nahrstedt. VTool-r1: VLMs learn to think with images via reinforcement learning on multimodal tool use. In *The Fourteenth International Conference on Learning Representations*, 2026.

[76] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024.

[77] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024.

[78] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

[79] Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. In *CVPR*, 2026.

[80] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13040–13051, 2024.

[81] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning*, 2024.

[82] Guanghao Zhang, Tao Zhong, Yan Xia, Zhelun Yu, Haoyuan Li, Wanggui He, Fangxun Shu, Mushui Liu, Dong She, Yi Wang, and Hao Jiang. Cmmcot: Enhancing complex multi-image comprehension via multi-modal chain-of-thought and memory augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[83] YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. MME-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *The Thirteenth International Conference on Learning Representations*, 2025.

[84] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[85] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[86] Jiyang Zheng, Jialiang Shen, Yu Yao, Min Wang, Yang Yang, Dadong Wang, and Tongliang Liu. Chain-of-focus prompting: Leveraging sequential visual cues to prompt large autoregressive vision models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[87] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and XingYu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[88] Yucheng Zhou, Wei Tao, Yiwen Guo, and Jianbing Shen. World models meet language models: On the complementarity of concrete and abstract reasoning. *arXiv preprint arXiv:2606.03603*, 2026.

[89] Deyao Zhu, jun chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 18378–18394, 2024.

A Appendix Overview

This appendix provides additional theoretical analysis, implementation details, and experimental results of DLVR. Appendix B presents the theoretical analysis of the DLVR training objectives. Appendix C describes the implementation details, covering the comparison methods, training and evaluation hyperparameters. Appendix D provides the construction of our dynamic and static training data. Appendix E reports additional experimental results, including detailed main results, the effect of latent size, and analysis of dynamic latent trajectories. Appendix F presents qualitative case studies of dynamic latent visual reasoning. Appendix G discusses limitations and future work.

B Derivations and Proofs

B.1 Information-Theoretic Perspective of MNTS

We denote the set of all selected tokens as $\mathcal{S}$ and the remaining tokens as $\mathcal{R} = \bigcup_{k=2}^{K} \mathcal{R}_k$ where $\mathcal{R}_k = \mathcal{H}_k \setminus \mathcal{S}_k$. If the selected tokens cover the information of all video tokens, $\mathcal{S}$ should make the unselected $\mathcal{R}$ highly predictable. The conditional entropy $H(\mathcal{R} \mid \mathcal{S})$ measures the uncertainty of the unselected tokens $\mathcal{R}$ after observing the selected tokens $\mathcal{S}$. If the conditional entropy is small, then $\mathcal{S}$ nearly determines $\mathcal{R}$, meaning $\mathcal{S}$ retains enough information to reconstruct the full token set. We minimize $H(\mathcal{R} \mid \mathcal{S})$ so that $\mathcal{S}$ retains sufficient information about the full frame tokens:

$$\min_{\mathcal{S}} H(\mathcal{R} \mid \mathcal{S}) = \min_{\mathcal{S}} \sum_{k=2}^{K} H(\mathcal{R}_k \mid \mathcal{R}_{<k}, \mathcal{S}). \tag{13}$$

$H(\mathcal{R}_k \mid \mathcal{R}_{<k}, \mathcal{S})$ measures the remaining uncertainty of $\mathcal{R}_k$ after observing the past unselected tokens $\mathcal{R}_{<k}$ and the entire selected set $\mathcal{S}$. We introduce an assumption that the tokens at time step k are only related to $\mathcal{H}_{\leq k}$. Following the principle that conditioning on fewer variables leads to larger entropy, we replace the condition with the prefix $\mathcal{S}_{\leq k}$ to obtain an upper bound for the original objective:

$$H(\mathcal{R} \mid \mathcal{S}) \leq \sum_{k=2}^{K} H(\mathcal{R}_k \mid \mathcal{R}_{<k}, \mathcal{S}_{\leq k}). \tag{14}$$

Equation 14 involves the intractable true conditional distribution $p(\mathcal{R}_k \mid \mathcal{R}_{<k}, \mathcal{S}_{\leq k})$, which is approximated via a variational distribution $q(\mathcal{R}_k \mid \mathcal{R}_{<k}, \mathcal{S}_{\leq k})$. Considering the KL divergence

$$\begin{aligned} &\mathrm{KL}\big(p(\mathcal{R}_k \mid \mathcal{R}_{<k}, \mathcal{S}_{\leq k}) \parallel q(\mathcal{R}_k \mid \mathcal{R}_{<k}, \mathcal{S}_{\leq k})\big) \\ &\quad = \mathbb{E}_p \left[\log \frac{p(\mathcal{R}_k \mid \mathcal{R}_{<k}, \mathcal{S}_{\leq k})}{q(\mathcal{R}_k \mid \mathcal{R}_{<k}, \mathcal{S}_{\leq k})} \right] \geq 0, \end{aligned} \tag{15}$$

we have

$$\mathbb{E}_p \left[-\log p(\mathcal{R}_k \mid \mathcal{R}_{<k}, \mathcal{S}_{\leq k}) \right] \leq \mathbb{E}_p \left[-\log q(\mathcal{R}_k \mid \mathcal{R}_{<k}, \mathcal{S}_{\leq k}) \right]. \tag{16}$$

Then we obtain the variational upper bound

$$\begin{aligned} H(\mathcal{R} \mid \mathcal{S}) &\leq \sum_{k=2}^{K} \mathbb{E}_p \big[-\log q(\mathcal{R}_k \mid \mathcal{R}_{<k}, \mathcal{S}_{\leq k}) \big] \\ &= \sum_{k=2}^{K} \mathbb{E}_p \left[-\log \frac{q(\mathcal{R}_k, \mathcal{S}_k \mid \mathcal{R}_{<k}, \mathcal{S}_{<k})}{q(\mathcal{S}_k \mid \mathcal{R}_{<k}, \mathcal{S}_{<k})} \right] \\ &= \sum_{k=2}^{K} \mathbb{E}_p \big[\log q(\mathcal{S}_k \mid \mathcal{H}_{<k}) - \log q(\mathcal{H}_k \mid \mathcal{H}_{<k}) \big]. \end{aligned} \tag{17}$$

In the variational upper bound, $q(\mathcal{H}_k \mid \mathcal{H}_{<k})$ is unrelated to the selection of tokens and regarded as a constant. Therefore, the objective in Equation 13 becomes

$$\min_{\mathcal{S}} \sum_{k=2}^{K} \mathbb{E}_p \big[\log q(\mathcal{S}_k \mid \mathcal{H}_{<k}) \big]. \tag{18}$$

Algorithm 1 Mahalanobis Novelty Token Selection

Require: Frame token sets $\{\mathcal{H}_k\}_{k=1}^{K}$
Ensure: Selected tokens $\mathcal{S}$
$\mathcal{S} \leftarrow \{\}$ ▷ Initialize the selected token set
for $k = 2$ to K **do**
  Compute $\boldsymbol{\mu}_k = \frac{1}{S}\sum_{i=1}^{S} \boldsymbol{h}_{k-1,i}^{\mathrm{VE}}$ ▷ Estimate the history mean
  Compute $\boldsymbol{\Sigma}_k = \frac{1}{S-1}\sum_{i=1}^{S}(\boldsymbol{h}_{k-1,i}^{\mathrm{VE}} - \boldsymbol{\mu}_k)(\boldsymbol{h}_{k-1,i}^{\mathrm{VE}} - \boldsymbol{\mu}_k)^\top$ ▷ Estimate the history covariance
  for each $\boldsymbol{x} \in \mathcal{H}_k$ **do**
    $d_M(\boldsymbol{x}, k) = (\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)$ ▷ Compute novelty score
  end for
  $\mathcal{S}_k \leftarrow \text{Top-}10\%\{\, \boldsymbol{x} \in \mathcal{H}_k \mid d_M(\boldsymbol{x}, k) \,\}$ ▷ Select the most novel tokens
  $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{S}_k$ ▷ Update selected token set
end for
return $\mathcal{S}$

To build the variational distribution, we model the history Gaussian distribution as

$$q(\mathcal{S}_k \mid \mathcal{H}_{<k}) = \prod_{\boldsymbol{x}\in\mathcal{S}_k} q(\boldsymbol{x} \mid \mathcal{H}_{<k}) = \prod_{\boldsymbol{x}\in\mathcal{S}_k} \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{19}$$

where

$$\boldsymbol{\mu}_k = \frac{1}{S}\sum_{i=1}^{S} \boldsymbol{h}_{k-1,i}^{\mathrm{VE}}, \quad \boldsymbol{\Sigma}_k = \frac{1}{S-1}\sum_{i=1}^{S}(\boldsymbol{h}_{k-1,i}^{\mathrm{VE}} - \boldsymbol{\mu}_k)(\boldsymbol{h}_{k-1,i}^{\mathrm{VE}} - \boldsymbol{\mu}_k)^\top. \tag{20}$$

Since $\boldsymbol{\Sigma}_k$ can be rank-deficient in high-dimensional hidden space, we use a shrinkage covariance $(\boldsymbol{\Sigma}_k + \mathrm{diag}(\boldsymbol{\Sigma}_k))/2$. According to the objective in Equation 18, we find the tokens that minimizes $\log q(\mathcal{S}_k \mid \mathcal{H}_{<k})$ at each time step as $\mathcal{S}_k$. For each candidate token $\boldsymbol{x} \in \mathcal{H}_k$, we estimate the history Gaussian distribution from the token set $\mathcal{H}_{<k}$:

$$q(\boldsymbol{x} \mid \mathcal{H}_{<k}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}_k|^{-1/2} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right). \tag{21}$$

The log-likelihood is

$$\log q(\boldsymbol{x} \mid \mathcal{H}_{<k}) = -\frac{1}{2}d_M(\boldsymbol{x}, k) \underbrace{-\frac{d}{2} \cdot \log(2\pi) - \frac{1}{2} \cdot \log|\boldsymbol{\Sigma}_k|}_{\text{Constant w.r.t. } \boldsymbol{x}}, \tag{22}$$

where $d_M(\boldsymbol{x}, k)$ is the Mahalanobis distance $\sqrt{(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)}$. Thus selecting tokens with maximum $d_M(\boldsymbol{x}, k)$ is exactly selecting tokens with maximum information of $\mathcal{H}_k$ under the previous-frame density. Finally, we select the top-$p\%$ tokens from $\mathcal{H}_k$ with the largest Mahalanobis scores to construct $\mathcal{S}_k$.

B.2 Feasibility of Temporal Ranking Constraint

We show that the temporal ranking constraint is feasible under the chosen margin $m = 1/(K-1)$. That is, the margin does not make the ranking objective intrinsically contradictory.

Proposition 1. *Assume that $d \geq K$. There exists a set of average-pooled representations $\{\bar{\boldsymbol{h}}_k\}_{k=1}^{K}$ such that $\mathcal{L}_{\mathrm{TCC}} = 0$ when the margin is set to $m = 1/(K-1)$ and $\bar{\boldsymbol{h}}_k \in \mathbb{R}^d$.*

Proof. We construct such a set of representations through the matrix $G \in \mathbb{R}^{K\times K}$ that

$$G_{ij} = 1 - \frac{|i-j|}{K-1}, \quad 1 \leq i, j \leq K. \tag{23}$$

This matrix assigns a larger inner product to temporally closer indices and a smaller inner product to temporally farther indices. We first verify that G is a valid Gram matrix. Clearly, G is symmetric and satisfies $G_{ii} = 1$ for all i. It remains to show that G is positive semidefinite. The function

$$\kappa(i, j) = 1 - \frac{|i-j|}{K-1} \tag{24}$$

is the discrete triangular kernel on the index set $\{1, \dots, K\}$. This kernel is positive semidefinite. One way to see this is that a triangular kernel can be obtained as the autocorrelation of a rectangular window, and autocorrelation kernels are positive semidefinite. Hence, $G \succeq 0$. Since G is positive semidefinite and has unit diagonal entries, it is a valid correlation matrix. Therefore, there exist representations $\{\boldsymbol{u}_k\}_{k=1}^K \subset \mathbb{R}^r$, where $r = \operatorname{rank}(G) \leq K$, such that $\boldsymbol{u}_i^\top \boldsymbol{u}_j = G_{ij}$. Moreover, because $G_{ii} = 1$, each vector has unit norm $\|\boldsymbol{u}_i\|_2^2 = G_{ii} = 1$. When $d \geq K$, these vectors can be embedded into $\mathbb{R}^d$ by zero-padding. We denote the resulting unit representations by $\{\bar{\boldsymbol{h}}_k\}_{k=1}^K \subset \mathbb{R}^d$. For these vectors, the cosine similarity is exactly their inner product:

$$\cos\left(\bar{\boldsymbol{h}}_i, \bar{\boldsymbol{h}}_j\right) = \bar{\boldsymbol{h}}_i^\top \bar{\boldsymbol{h}}_j = G_{ij} = 1 - \frac{|i-j|}{K-1}. \tag{25}$$

Now consider any triplet (i, j, k) used in the temporal ranking constraint, where j is temporally closer to i than k, i.e., $|i-j| < |i-k|$. Then

$$\cos\left(\bar{\boldsymbol{h}}_i, \bar{\boldsymbol{h}}_j\right) - \cos\left(\bar{\boldsymbol{h}}_i, \bar{\boldsymbol{h}}_k\right) = \left(1 - \frac{|i-j|}{K-1}\right) - \left(1 - \frac{|i-k|}{K-1}\right) = \frac{|i-k| - |i-j|}{K-1}. \tag{26}$$

Since $|i-j|$ and $|i-k|$ are integers and $|i-j| < |i-k|$, we have $|i-k| - |i-j| \geq 1$. Therefore,

$$\cos\left(\bar{\boldsymbol{h}}_i, \bar{\boldsymbol{h}}_j\right) - \cos\left(\bar{\boldsymbol{h}}_i, \bar{\boldsymbol{h}}_k\right) \geq \frac{1}{K-1} = m. \tag{27}$$

Thus every triplet satisfies the ranking constraint with margin m. Consequently, every hinge term in $\mathcal{L}_{\mathrm{TCC}}$ is zero, and hence $\mathcal{L}_{\mathrm{TCC}} = 0$. This proves the feasibility of the temporal ranking constraint. □

C Detailed Experimental Settings

Appendix C.1 introduces the benchmarks used in our evaluation, Appendix C.2 describes the compared methods, and Appendix C.3 provides detailed hyperparameter settings.

C.1 Benchmarks

We evaluate DLVR on two categories of benchmarks: *Dynamic-Centric Image Understanding Tasks*, including VTBench [63], Visual Cognitive Tasks [56], and BabyVision [12], and *General Image Understanding Tasks*, including MMBench [49], MMStar [13], MMVet [81], V* [76], HRBench4K/8K [72], MME-RealWorld [83], and SpatialEval [67]. Collectively, these benchmarks include diverse tasks such as BabyVision, VTBench, and document OCR in MME-RealWorld to validate the generality of DLVR. With the General Image Understanding Tasks (e.g., MMBench, HRBench, SpatialEval) containing numerous samples, thereby testing DLVR's ability to generalize dynamic latent reasoning beyond the training distribution.

BabyVision [12]. BabyVision is a visual reasoning benchmark designed to test whether LVLMs can handle vision-centric problems that are easy for young children but still challenging for current models. It provides evaluation track for multimodal language models, and covers four core categories: Fine-Grained Discrimination, Visual Tracking, Spatial Perception, and Visual pattern recognition. Its tasks include finding subtle differences, solving mazes, understanding 3D views, cube unfolding, paper folding, block counting, and completing visual patterns, making it especially suitable for evaluating low-level perception, spatial reasoning, and non-linguistic visual intelligence. The benchmark reports overall, category-wise, and subtype-wise accuracy, and its leaderboard shows a large gap between current LVLMs and human performance, highlighting that strong language reasoning does not yet imply mature visual reasoning ability.

Visual Cognitive Tasks [56]. Visual Cognitive Tasks is a cognitively motivated benchmark for evaluating whether multimodal large language models can perform higher level visual reasoning beyond surface recognition. In this work, we focus on three representative tasks from the benchmark: Block Tower, Michotte, and Astronaut. Block Tower evaluates intuitive physics by asking models to judge the stability of real block structures, requiring an understanding of support, balance, and physical plausibility. Michotte examines causal perception through launching style interactions, where models must infer whether one object's motion causes another object's movement. Astronaut targets intuitive psychology by testing whether models can infer an agent's preference or utility from

visual observations of goal directed behavior. We futher include the evaluation on the Help or Hinder in the detailed experimental results provided in the Appendix E.These tasks assess complementary aspects of visual cognition, covering physical dynamics, cause effect understanding, and intention inference. We follow the original evaluation protocol and report the corresponding task metrics.

VTBench [63]. VTBench is a vision-centric benchmark designed to evaluate complex reasoning from static visual inputs. Eyeballing focuses on spatial estimation and geometric intuition, requiring models to make approximate visual judgments about positions, lengths, alignments, or relative quantities without relying on explicit measurements. Visual Puzzles evaluates inductive visual reasoning, where models must identify hidden rules, infer visual transformations, and select the correct continuation or missing element from structured patterns. These two tasks provide complementary tests of spatial perception and abstract pattern reasoning, making them well suited for assessing whether LVLMs can move beyond object recognition toward deeper visual understanding. We follow the original evaluation protocol and report the corresponding task accuracy.

MMBench [49]. MMBench is a systematically designed benchmark for evaluating the overall multimodal understanding ability of large vision language models. It contains over 3,000 multiple choice questions covering 20 fine grained ability dimensions, including object localization, OCR, attribute recognition, action understanding, image description, and relation reasoning. To improve evaluation reliability, MMBench adopts CircularEval, which tests models under different option orders and reduces the bias caused by unstable choice prediction. It also provides both English and Chinese versions, enabling consistent bilingual evaluation. Overall, MMBench offers a broad and objective testbed for measuring whether an MLLM can serve as an all around visual understanding model rather than excelling only on a narrow set of tasks.

MMStar [13]. MMStar is a carefully curated benchmark for evaluating the genuine multimodal capabilities of LVLMs. It contains human selected, vision indispensable samples, where each question is designed to require visual evidence rather than being solvable from language priors alone. The benchmark covers 6 core capabilities and 18 fine grained axes, with each core capability balanced to contain samples. By emphasizing visual dependency, reduced data leakage, and challenging multimodal reasoning, MMStar provides a more reliable testbed for measuring whether performance gains truly come from visual understanding rather than dataset bias or textual shortcuts.

MM-Vet [81]. MM-Vet is an open-ended benchmark for evaluating whether large multimodal models can solve complex visual tasks by integrating multiple vision-language capabilities. Instead of testing isolated skills, MM-Vet emphasizes capability composition across six dimensions: recognition, OCR, knowledge, language generation, spatial awareness, and math. Its questions often require models to jointly perceive visual content, read text, reason over spatial relations, apply external knowledge, and generate fluent answers, making it a strong test of generalist multimodal intelligence. To handle diverse answer formats, MM-Vet adopts an LLM-based evaluation protocol that grades free-form model responses with a unified scoring metric. Overall, it provides a challenging and comprehensive benchmark for measuring integrated multimodal reasoning beyond simple visual question answering.

V* [76]. V* is a fine-grained visual search benchmark, designed to evaluate whether LVLMs can locate and reason over task-relevant details in high-resolution and visually crowded images. Unlike conventional VQA benchmarks that often focus on salient objects or global scene understanding, V* requires models to actively search for small, subtle, or easily overlooked visual evidence before answering. It therefore tests a model's ability to combine precise visual grounding, selective attention, and localized reasoning under challenging real-world visual conditions. Overall, V* highlights visual search as a crucial capability for robust multimodal reasoning, especially when the answer depends on finding the right evidence rather than simply understanding the image as a whole.

HRBench [72]. HRBench is a high-resolution image understanding benchmark designed to evaluate whether multimodal large language models can perceive and reason over fine-grained details in ultra-high-resolution visual inputs. It contains two settings, HRBench4K and HRBench8K, corresponding to 4K and 8K image resolutions, and is built to expose the limitations of models that appear strong on standard low-resolution benchmarks but lose critical visual information after downsampling. The benchmark emphasizes detail-sensitive recognition, localized perception, and high-resolution visual reasoning, making it a rigorous testbed for assessing whether LVLMs can effectively exploit rich visual signals in real-world high-resolution images. Overall, HRBench highlights the importance of preserving fine visual details for robust multimodal understanding.

MME-RealWorld [83]. MME-RealWorld is a large-scale benchmark designed to evaluate multimodal large language models under challenging high-resolution real-world scenarios. It contains high-quality images with an average resolution of around 2000×1500, together with manually annotated question-answer pairs spanning 43 subtasks across five practical domains, including OCR in the wild, remote sensing, diagrams and tables, autonomous driving, and monitoring. Unlike conventional benchmarks that often rely on low-resolution images or simplified visual questions, MME-RealWorld emphasizes fine-grained perception, real-world complexity, and application-oriented reasoning. It therefore provides a rigorous testbed for assessing whether LVLMs can reliably understand detailed visual evidence in realistic high-resolution environments.

SpatialEval [67]. SpatialEval is a benchmark designed to evaluate the spatial intelligence of Large Vision-Language Models. It covers four complementary dimensions of spatial reasoning, including spatial relationships, positional understanding, object counting, and navigation. The benchmark consists of four task settings, Spatial-Map, Maze-Nav, Spatial-Grid, and Spatial-Real, which jointly test whether models can understand object layouts, reason about relative positions, navigate through structured spaces, and count spatially arranged objects. By combining synthetic and real-world scenarios, SpatialEval provides a focused and controlled testbed for measuring spatial reasoning beyond general visual recognition, revealing whether models can build reliable spatial representations from visual inputs.

C.2 Other Methods

We compare DLVR with several existing thinking with image and latent visual reasoning methods for improving reasoning in LVLMs. All methods are implemented on Qwen2.5-VL-7B, or evaluated using their officially released checkpoints trained on Qwen2.5-VL-7B. For all methods, we adopt the best-performing configurations reported in their papers and reproduce the results by implementing their pipelines based on the released code and documented procedures, ensuring a fair comparison.

C.2.1 Thinking With Image

VISER [33] is a visual structuring method designed to mitigate the binding problem in LVLMs, where models fail to correctly associate visual attributes with their corresponding objects. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, VISER applies a lightweight visual scaffold operator $\mathcal{S}_n$ that inserts n horizontal lines and optional row indices into the image:

$$I^{\text{vis}} = \mathcal{S}_n(I), \tag{28}$$

where the horizontal lines divide the image into $n+1$ ordered regions $\{R_i\}_{i=1}^{n+1}$. The structured image is paired with a sequential scanning instruction, yielding the final model input:

$$q^{\text{vis}} = [q;\ p_{\text{scan}}], \quad p_{\text{scan}} = \text{“Scan the image sequentially based on the horizontal lines.”} \tag{29}$$

The model prediction is then obtained by

$$\hat{y} = f_\theta(I^{\text{vis}}, q^{\text{vis}}). \tag{30}$$

Conceptually, VISER encourages the LVLM to replace global parallel perception with row-wise visual parsing:

$$h_i = \Phi_\theta(h_{i-1}, R_i, q), \quad i = 1, \ldots, n+1, \qquad \hat{y} = \Psi_\theta(h_{n+1}), \tag{31}$$

where each region is processed in an ordered manner before producing the final answer. By adding explicit spatial anchors to the visual input, VISER improves object-level binding and reduces feature interference in tasks such as counting, visual search, scene description, and spatial relationship reasoning.

ICoT [26] is a multimodal reasoning method that extends conventional text-only CoT by inserting fine-grained visual evidence into intermediate reasoning steps. Given an input image I and question q, standard multimodal CoT generates only textual rationales:

$$r_1, r_2, \ldots, \hat{y} = f_\theta(I, q, p_{\text{cot}}), \tag{32}$$

where r_i denotes the i-th textual rationale and $\hat{y}$ is the final answer. In contrast, ICoT formulates the reasoning process as an interleaved sequence of visual and textual rationales:

$$(v_1, r_1), (v_2, r_2), \ldots, (v_T, r_T), \hat{y} = f_\theta(I, q, p_{\text{icot}}), \tag{33}$$

where each v_t is a fine-grained visual rationale selected from the original image and paired with the following textual rationale r_t. Since existing LVLMs usually cannot generate such visual rationales directly, ICoT introduces Attention-driven Selection (ADS) to retrieve relevant visual tokens from the input image. Let the visual encoder produce visual tokens

$$V = E(I) = \{v_i\}_{i=1}^{N}, \quad v_i \in \mathbb{R}^d. \tag{34}$$

When the model generates a predefined signal token s_t, ADS extracts the attention map A_t between this signal token and all visual tokens, and selects the top-K attended visual tokens:

$$\mathcal{I}_t = \mathrm{TopK}(A_t, K), \qquad V_t^{\mathrm{sel}} = \{v_i \mid i \in \mathcal{I}_t\}. \tag{35}$$

The selected visual tokens are then restored according to their original spatial order and inserted into the current decoding context:

$$x_{t+1} = \mathrm{Concat}\left(x_{\le t}, V_t^{\mathrm{sel}}\right), \qquad r_t = f_\theta(x_{t+1}). \tag{36}$$

Through this process, ICoT enables the model to reason with localized image regions and textual explanations in an alternating manner, making the generated rationale more visually grounded and interpretable. Compared with text-only multimodal CoT, ICoT better preserves fine-grained image evidence, reduces ambiguous linguistic references, and improves reasoning on tasks that require detailed visual association, without model fine-tuning or additional external tools.

C.2.2 Latent Visual Reasoning

CoVT [53] enables LVLMs to reason with latent visual tokens distilled from multiple vision experts. Instead of restricting intermediate reasoning to discrete language tokens, CoVT introduces a compact set of continuous visual tokens that encode complementary perceptual cues, including segmentation, depth, edge structure, and DINO-style patch features. Given visual features V and text features T, a standard LVLM predicts a textual response as

$$P(Y \mid V, T; \theta) = \prod_{i=1}^{n} P(y_i \mid y_{<i}, V, T). \tag{37}$$

CoVT extends this formulation by allowing each token y_i to be either a text token or a continuous visual thought token:

$$Y = \{y_1, \ldots, y_n\}, \qquad y_i \in \mathcal{V}_{\mathrm{text}} \cup \mathcal{Z}_{\mathrm{vis}}, \tag{38}$$

where $\mathcal{Z}_{\mathrm{vis}}$ denotes the latent visual token space. These visual tokens are supervised by multiple lightweight vision experts. Formally, for expert type $m \in \{\mathrm{seg}, \mathrm{depth}, \mathrm{edge}, \mathrm{dino}\}$, CoVT predicts a group of latent visual tokens

$$Z^m = \{z_1^m, \ldots, z_{K_m}^m\}, \qquad z_k^m \in \mathbb{R}^d, \tag{39}$$

which are projected and decoded to reconstruct the corresponding expert target:

$$\hat{O}^m = D_m\big(\Pi_m(Z^m), F_m(I)\big). \tag{40}$$

Here, D_m is the task-specific decoder, Π_m is the projection layer, and $F_m(I)$ denotes dense features or targets provided by the corresponding vision expert. The training objective jointly optimizes language generation and visual-token alignment:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{CE}} + \gamma\left(\lambda_{\mathrm{seg}}\mathcal{L}_{\mathrm{seg}} + \lambda_{\mathrm{depth}}\mathcal{L}_{\mathrm{depth}} + \lambda_{\mathrm{edge}}\mathcal{L}_{\mathrm{edge}} + \lambda_{\mathrm{dino}}\mathcal{L}_{\mathrm{dino}}\right). \tag{41}$$

During inference, CoVT performs reasoning directly in the continuous visual token space:

$$\hat{y} = f_\theta\left(I, q, \underbrace{Z^{\mathrm{seg}}, Z^{\mathrm{depth}}, Z^{\mathrm{edge}}, Z^{\mathrm{dino}}}_{\text{latent visual thoughts}}\right). \tag{42}$$

The dense outputs can still be optionally decoded for interpretability. By distilling multiple expert signals into a small number of latent visual tokens, CoVT provides LVLMs with compact visual thoughts that preserve fine-grained perception, spatial layout, and geometric awareness, thereby improving vision-centric reasoning while maintaining efficient autoregressive inference.

CoVT, as a highly valuable work, uses 774K training datas. Despite its strong performance, CoVT still relies on supervision from **multiple external vision experts**, which may increase the complexity of data construction and tie the learned latent tokens to the capability and coverage of these predefined experts. Moreover, its visual thoughts **mainly encode static perceptual cues** such as depth, edges, segmentation, and patch-level features, while temporal evolution, motion direction, and causal state transitions are not explicitly modeled. Therefore, CoVT provides a strong latent visual reasoning baseline, but remains less suited for scenarios where the key evidence not only lies in static perceptual structure.

LVR [36] is a multimodal reasoning paradigm that focuses latent reasoning on the core regions of the input image. Unlike text-only CoT, which verbalizes visual evidence into discrete language tokens, LVR performs autoregressive reasoning directly in the visual embedding space. Given an image-question pair (X_v, X_t), the visual encoder first extracts visual features and maps them into the language-aligned semantic space:

$$V = \mathrm{Enc}_v(X_v), \qquad V_T = \mathrm{Proj}(V), \tag{43}$$

where $V_T = \{v_1, \ldots, v_N\}$ denotes the projected visual tokens. During supervised fine-tuning, LVR uses the query-relevant region of interest to select the core visual tokens:

$$\mathcal{I}_{\mathrm{roi}} = \{i \mid p_i \in \mathrm{ROI}(X_v, X_t)\}, \qquad V_{\mathrm{core}} = \{v_i \mid i \in \mathcal{I}_{\mathrm{roi}}\}. \tag{44}$$

When the model generates the special token $\langle \mathrm{lvr_start} \rangle$, it enters the latent visual reasoning mode. Instead of producing text tokens, the model propagates its last hidden states as continuous latent thoughts:

$$h_t = f_\theta(h_{<t}, V_T, X_t), \qquad t = 1, \ldots, T_v. \tag{45}$$

These latent states are trained to reconstruct the selected core visual tokens:

$$\mathcal{L}_{\mathrm{LVR}} = \frac{1}{T_v} \sum_{t=1}^{T_v} \|h_t - v_t^{\mathrm{core}}\|_2^2 . \tag{46}$$

After the latent reasoning segment ends with $\langle \mathrm{lvr_end} \rangle$, the model resumes standard text generation conditioned on the reconstructed latent visual thoughts:

$$\mathcal{L}_{\mathrm{NTP}} = -\frac{1}{T_y} \sum_{t=1}^{T_y} \log p_\theta \left(y_t \mid y_{<t}, h_{1:T_v}, V_T, X_t \right). \tag{47}$$

The final supervised objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{NTP}} + \lambda_{\mathrm{LVR}} \mathcal{L}_{\mathrm{LVR}}. \tag{48}$$

After supervised fine-tuning, LVR further adopts a latent-space variant of GRPO, denoted as $\mathrm{GRPO}_{\mathrm{latent}}$, to refine the interaction between latent visual reasoning and textual answer generation. Since latent reasoning states are continuous and do not define an explicit token distribution, LVR computes the policy-gradient objective only over the generated text tokens, while replaying the latent hidden states produced during rollout. Formally, for an output sequence y_i with recorded latent states

$$\tilde{h}_i^{\mathrm{latent}} = \{h_{i,1}^{\mathrm{latent}}, \ldots, h_{i,L}^{\mathrm{latent}}\}, \tag{49}$$

the token-level importance ratio is computed as

$$r_{i,t}(\theta) = \frac{\pi_\theta \left(y_{i,t} \mid q, I, \tilde{h}_i^{\mathrm{latent}}, y_{i,<t} \right)}{\pi_{\theta_{\mathrm{old}}} \left(y_{i,t} \mid q, I, \tilde{h}_i^{\mathrm{latent}}, y_{i,<t} \right)}. \tag{50}$$

The optimization objective follows the clipped GRPO form:

$$\mathcal{J}_{\mathrm{GRPO}_{\mathrm{latent}}}(\theta) = \mathbb{E} \left[\frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \min \left(r_{i,t}(\theta) \hat{A}_{i,t}, \mathrm{clip} \left(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon \right) \hat{A}_{i,t} \right) - \beta D_{\mathrm{KL}} \left(\pi_\theta \| \pi_{\mathrm{ref}} \right) \right]. \tag{51}$$

The reward is computed from the final textual response rather than directly from the latent states. Specifically, LVR uses an accuracy reward to evaluate whether the final answer is correct, and a format reward to encourage the model to generate both $\langle \text{lvr_start} \rangle$ and $\langle \text{lvr_end} \rangle$:

$$R(y_i) = R_{\text{acc}}(y_i) + R_{\text{fmt}}(y_i). \tag{52}$$

The group-normalized advantage is then given by

$$\tilde{R}_i = \frac{R(y_i) - \text{mean}\left(R(y_1), \ldots, R(y_G)\right)}{\text{std}\left(R(y_1), \ldots, R(y_G)\right)}, \qquad \hat{A}_{i,t} = \tilde{R}_i. \tag{53}$$

In this way, LVR teaches the model to reconstruct visual semantics from the most answer-critical regions before producing the final response. By concentrating latent reasoning on query-relevant image regions, LVR strengthens fine-grained visual understanding without relying on external cropping tools or explicit image editing. It therefore provides a compact and end-to-end mechanism for bridging visual perception and language generation in perception-intensive reasoning tasks.

LVR is an effective and elegant step toward end-to-end latent visual reasoning, showing that continuous hidden states can help models focus on answer-critical visual evidence without explicit image editing. However, its latent reasoning is still **primarily grounded in static regions of interest (ROIs)** of the input image, where the supervision comes from reconstructing query-relevant visual tokens rather than modeling how the scene may evolve over time. Its training data depends on numbers of text-annotated visual-CoT data (438K). In addition, its RL stage **mainly optimizes the final textual response**, while the latent states themselves are not directly rewarded for temporal coherence or dynamic plausibility. As a result, LVR provides a strong baseline for perception-intensive reasoning, but may less tailored to scenarios where the decisive evidence lies in motion direction, causal evolution, or future state transitions.

Monet [70] enables LVLMs to reason directly in continuous visual space by generating latent embeddings as intermediate visual thoughts. Unlike methods that directly align latent embeddings with dense auxiliary-image tokens, Monet further uses *observation text tokens* as supervision signals derived from auxiliary images. Given an input image I, question q, and an auxiliary image I^{aux} appearing in an interleaved reasoning trajectory, Monet first identifies text tokens that describe key visual observations from I^{aux}:

$$\mathcal{O} = \{o_1, \ldots, o_M\}, \tag{54}$$

where each o_i corresponds to crucial visual evidence required for answering the question. A teacher model observes the auxiliary image and produces observation-token representations:

$$H^*_{\text{obs}} = \{h^{*(i,l)}_{\text{obs}}\}, \tag{55}$$

where i indexes observation tokens and l indexes model layers. The student model replaces the auxiliary image with generated latent embeddings:

$$Z_{\text{latent}} = \{z_1, \ldots, z_K\}, \tag{56}$$

and is trained to make its observation-token representations match those of the teacher:

$$\mathcal{L}_{\text{align-obs}} = \frac{1}{N} \sum_i \sum_l \left(1 - \cos\left(h^{*(i,l)}_{\text{obs}}, \hat{h}^{(i,l)}_{\text{obs}}\right)\right). \tag{57}$$

To ensure that this alignment truly improves latent visual thoughts rather than exploiting shortcut paths through text representations, Monet applies latent-only backpropagation, allowing gradients from $\mathcal{L}_{\text{align-obs}}$ to update the model only through the generated latent embeddings. In addition, Monet introduces a controlled attention flow:

$$I^{\text{aux}} \to Z_{\text{latent}} \to \mathcal{O}, \tag{58}$$

where auxiliary-image embeddings are visible to latent embeddings but invisible to subsequent text tokens. This design forces latent embeddings to selectively absorb useful visual cues from the auxiliary image before transferring them to observation-token prediction.

After obtaining high-quality target latent embeddings Z^*_{latent}, Monet trains the model to generate latent embeddings without access to auxiliary images:

$$\mathcal{L}_{\text{align-latent}} = \frac{1}{N} \sum_i \sum_l \left(1 - \cos\left(z^{*(i,l)}_{\text{latent}}, \hat{z}^{(i,l)}_{\text{latent}}\right)\right). \tag{59}$$

The supervised training objective combines text generation and latent alignment:

$$\mathcal{L}_{\mathrm{SFT}} = \mathcal{L}_{\mathrm{NTP}} + \beta \mathcal{L}_{\text{align-latent}}, \tag{60}$$

where

$$\mathcal{L}_{\mathrm{NTP}} = -\frac{1}{N} \sum_{i} \sum_{y_t \in T^{(i)}} \log p_\theta(y_t \mid y_{<t}). \tag{61}$$

Thus, Monet distills the effect of auxiliary images into compact latent embeddings, using observation tokens as fine-grained supervision for what visual information should be encoded.

Monet further proposes Visual-latent Policy Optimization (VLPO). Standard GRPO can only compute policy ratios over discrete text tokens, leaving continuous latent embeddings weakly optimized. To address this, VLPO estimates the probability of a rollout latent embedding under the current policy by modeling the latent action as a Gaussian sample. Let $h_{i,t}^{\mathrm{old}}$ be the latent embedding generated by the old policy during rollout, and let $h_{i,t}^{\theta}$ be the latent embedding generated by the current policy under the same context. Monet defines:

$$\pi_\theta\left(h_{i,t}^{\mathrm{old}} \mid q, I, o_{i,<t}\right) = \exp\left(-\frac{1}{2\sigma^2}\left\|h_{i,t}^{\mathrm{old}} - h_{i,t}^{\theta}\right\|_2^2 - C\right), \tag{62}$$

where σ is a predefined scalar and C is a constant. The latent-step importance ratio is then:

$$r_{i,t}^{\mathrm{latent}}(\theta) = \exp\left(-\frac{1}{2\sigma^2}\left\|h_{i,t}^{\mathrm{old}} - h_{i,t}^{\theta}\right\|_2^2\right). \tag{63}$$

Replacing the text-token ratio in GRPO with this latent ratio yields the VLPO objective:

$$\mathcal{J}_{\mathrm{VLPO}}(\theta) = \mathbb{E}\left[\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \min\left(r_{i,t}(\theta)\hat{A}_{i,t}, \mathrm{clip}\left(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right) - \beta D_{\mathrm{KL}}\left(\pi_\theta \| \pi_{\mathrm{ref}}\right)\right]. \tag{64}$$

For text steps, $r_{i,t}(\theta)$ is the standard token probability ratio; for latent steps, it is replaced by $r_{i,t}^{\mathrm{latent}}(\theta)$. Monet uses an accuracy reward and a format reward:

$$R(o_i) = R_{\mathrm{acc}}(o_i) + R_{\mathrm{fmt}}(o_i), \tag{65}$$

where the accuracy reward evaluates the final answer and the format reward encourages the required answer format. When the advantage $\hat{A}_{i,t}$ is positive, VLPO pulls the current latent embedding $h_{i,t}^{\theta}$ toward the rollout latent embedding $h_{i,t}^{\mathrm{old}}$ that led to a successful answer.

Monet is a strong and carefully designed latent reasoning framework, showing that auxiliary visual evidence can be distilled into compact latent embeddings and further optimized through latent-space policy learning supervised by text observation tokens. However, its training pipeline still depends on densely text-annotated visual-CoT data (125K), whose distribution may partially overlap with downstream evaluation benchmarks and whose construction can be costly. Its latent supervision also remains largely tied to **static auxiliary visual evidence**, while temporal evolution, causal dynamics, and trajectory-level coherence are not explicitly modeled. The RL reward in Monet **still only considers the correctness of the final answer** while overlooking the fact that latent tokens are expected to form a continuous and logically coherent reasoning trajectory. Therefore, Monet provides a strong baseline for latent visual reasoning, but is still less directly suited for learning dynamic imagination from temporal visual evolution.

C.3 Detailed Hyperparameters

All models are trained and evaluated on 8 × A100 GPUs. We use *DLVR-SFT-100K* for supervised fine-tuning, where Stage 1 and Stage 2 are trained for 4 and 2 epochs, respectively. For both SFT stages, the batch size is set to 128, the gradient accumulation steps are set to 16, and the maximum number of pixels per image is set to $2000 \times 28 \times 28$. In SFT Stage 2, we uniformly sample subsequent video frames and set $K = 2N$. For reinforcement learning, we use *DLVR-RL-4K* and train the model for 1 epoch. We set $\beta = 0.1$, $\lambda = 0.5$ and $\xi = 0.01$. The latent size is set to 16 during both training and evaluation as the default configuration. We use a learning rate of 1×10^{-5} and a weight decay of 0.01. We also add two special tokens, `<Dynamic>` and `</Dynamic>`, to the vocabulary, which mark the beginning and end of dynamic latent reasoning. During evaluation, we follow the official settings of all compared methods and use their publicly released checkpoints whenever available to ensure a fair comparison, where all methods are also trained based on Qwen2.5-VL-7B. We set `do_sample` = `False` in test time, corresponding to greedy decoding, to obtain deterministic outputs.

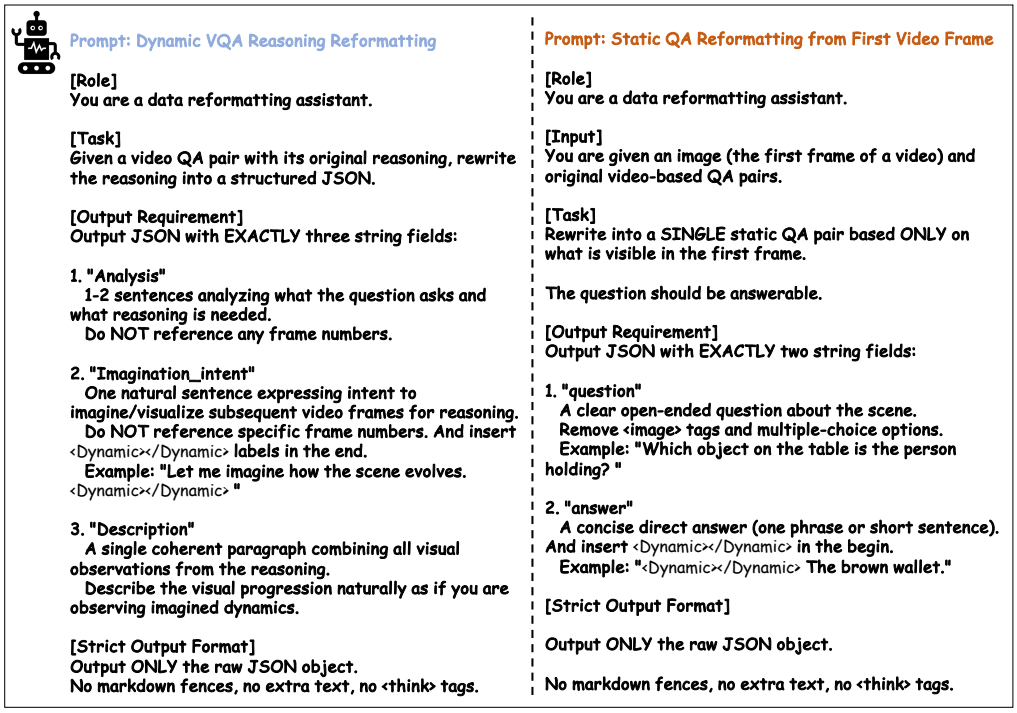


Figure 5: Prompts for data construction.

D Training Datasets Construction

We construct training data for dynamic latent reasoning that covers both dynamic scene evolution and static visual understanding, while avoiding external visual tools such as cropping, auxiliary drawing, or image editing. Importantly, our data construction does not rely on existing CoT-annotated visual reasoning datasets, which may overlap in distribution with evaluation benchmarks. Instead, it builds dynamic supervision directly from open-source video data and first-frame-based static QA, reducing potential benchmark-specific bias. Moreover, this construction pipeline can be extended to incorporate diverse dynamic video datasets, enabling the model to learn broader dynamic world knowledge and improving its generalization to broader scenarios.

D.1 Prompts For Data Construction

For physical dynamic question answering, we construct 40K training samples based on representative scenarios from the VBVR dataset [68], including ball bouncing and domino collapse. For each sample, we and insert key temporal frames that capture the essential physical evolution into the response as continuous dynamic evidence.

As shown in Figure 5 left, for open-scene dynamic question answering, we use DeepSeek-V3.2 [44] to reformulate 58K future-prediction questions from Twiff [48]. During this process, we preserve the visual evidence most relevant to future scene evolution by inserting corresponding key frames into the response, while simplifying the original textual reasoning to emphasize the underlying dynamic process. This construction allows the model to learn dynamic reasoning from both controlled physical interactions and diverse real-world scenes. As illustrated in Figure 5 right, for static question answering, we build 1.8K training samples from the existing QA annotations in the PerceptionTest subset of LLaVA-Video-Train [84]. Each question is reformulated to be grounded in the first frame, while the following frames are incorporated into the response as temporally continuous world knowledge. This design connects static visual understanding with the underlying scene evolution, enabling the model to learn dynamic context even from static QA samples. We use the token pair `<Dynamic>` and `</Dynamic>` as placeholders for the inserted dynamic information.

Table 10: **Performance comparison on BabyVision.** Blue denotes improvement, gray denotes degradation, the best-performing results are highlighted in bold.

Model	Fine-Grained	Vis. Tracking	Spatial Perc.	Vis. Pattern	Overall
Qwen2.5-VL-7B [5]	6.13 (0.00)	9.64 (0.00)	15.38 (0.00)	17.65 (0.00)	10.57 (0.00)
Think With Image					
+VISER [33](NeurIPS'25)	7.36 (1.23)	6.02 (3.62)	16.48 (1.10)	23.53 (5.88)	11.34 (0.77)
+ICoT [26](CVPR'25)	10.43 (4.30)	8.43 (1.21)	10.99 (4.39)	21.56 (3.91)	11.60 (1.03)
Latent Reasoning Method					
+CoVT [53](arXiv'25)	11.04 (4.91)	10.84 (1.20)	14.29 (1.09)	15.69 (1.96)	12.37 (1.80)
+LVR [36](ICLR'26)	9.82 (3.69)	12.05 (2.41)	18.68 (3.30)	17.66 (0.01)	13.40 (2.83)
+Monet [70](CVPR'26)	10.42 (4.29)	13.25 (3.61)	15.38 (0.00)	21.57 (3.92)	13.91 (3.34)
+DLVR (Ours)	**16.56** (10.43)	**18.07** (8.43)	**26.37** (10.99)	**33.33** (15.68)	**21.39** (10.82)

D.2 Cases of Training Dataset

Figure 6 illustrates representative examples from our Dynamic-centric training dataset, covering both physical dynamics and open-scene real-world dynamics. Each sample starts from a static input image and inserts subsequent key frames as dynamic evidence, enabling the answer to explicitly imagine how the scene evolves before producing the final response. In physical scenarios, the model is guided to infer motion trajectories and collision outcomes from visual cues, while in open scenes it learns to anticipate human actions and object interactions over time. These examples show how our dataset contains dynamic reasoning supervision, encouraging the model to move from simply recognizing the current scene toward predicting temporally coherent future evolution.

Figure 6 also shows that dynamic imagination can also benefit static visual question answering. Although the question only asks about static content, such as the word formed by blocks on the table, the inserted dynamic process helps the model mentally trace how the visual evidence may be used, for example by imagining the person typing the observed letters into the computer. This mirrors human cognition, where people often reinforce static perception through implicit action simulation. By associating static QA with temporally extended visual context, our dataset encourages the model to form more stable visual impressions and produce more reliable answers.

E Additional Results

E.1 Detailed Main Results

For BabyVision, Table 10 shows that DLVR consistently achieves the best performance across all four subcategories. Compared with the Qwen2.5-VL-7B baseline, DLVR improves the overall score from 10.57% to 21.39%, yielding a substantial gain of 10.82%. The improvement is especially pronounced on Vis. Pattern and Spatial Perc., where DLVR brings gains of 15.68% and 10.99%, respectively. These results indicate that dynamic latent reasoning is particularly effective for vision-centric tasks that require fine-grained perception, spatial understanding, and implicit prediction of visual evolution.

For Visual Cognitive Tasks and VTBench, Table 11 further demonstrates the strong generality of DLVR on dynamic and cognitively demanding visual reasoning. DLVR improves the overall score of Visual Cognitive Tasks from 25.11% to 33.36%, outperforming both think-with-image methods and prior latent reasoning methods across all sub-tasks. On VTBench, DLVR raises the overall score from 19.03% to 29.43%, with a particularly large gain on Eyeballing from 11.42% to 23.34%. These results show that DLVR not only strengthens physical and causal reasoning, but also improves spatial estimation and abstract visual puzzle solving through temporally coherent latent imagination which is out of training distribution.

E.2 Detailed Results of Latent Size

Tables 12, 13, and 14 provide detailed results for different test-time latent sizes under fixed training-time latent budgets. A clear pattern is that enabling latent tokens brings the dominant performance

Table 11: **Performance comparison on Visual Cognitive Tasks and VTBench.** Blue denotes improvement, gray denotes degradation, the best-performing results are highlighted in bold.

Model	Visual Cognitive Tasks					VTBench		
	Block Tower	Michotte	Astronaut	Help or Hinder	Overall	Eyeballing	Vis. Puzzles	Overall
Qwen2.5-VL-7B [5]	46.37 (0.00)	16.46 (0.00)	20.38 (0.00)	17.21 (0.00)	25.11 (0.00)	11.42 (0.00)	34.87 (0.00)	19.03 (0.00)
Think With Image								
+VISER [33](NeurIPS'25)	47.32 (0.95)	16.33 (0.13)	19.87 (0.51)	19.93 (2.72)	25.86 (0.75)	11.90 (0.48)	36.08 (1.21)	19.66 (0.71)
+ICoT [26](CVPR'25)	48.91 (2.54)	17.04 (0.58)	20.11 (0.27)	18.62 (1.41)	26.17 (1.06)	12.38 (0.96)	34.27 (0.60)	19.40 (0.45)
Latent Reasoning Method								
+CoVT [53](arXiv'25)	49.11 (2.74)	15.80 (0.66)	22.89 (2.51)	17.93 (0.72)	26.43 (1.32)	13.33 (1.91)	36.69 (1.82)	20.83 (1.88)
+LVR [36](ICLR'26)	51.07 (4.70)	18.59 (2.13)	22.31 (1.93)	18.44 (1.23)	27.60 (2.49)	13.80 (2.38)	36.89 (2.02)	21.22 (2.27)
+Monet [70](CVPR'26)	50.94 (4.57)	17.73 (1.27)	23.60 (3.22)	19.67 (2.46)	27.99 (2.88)	13.91 (2.49)	37.50 (2.63)	21.47 (2.52)
+DLVR (Ours)	**57.79** (11.42)	**22.15** (5.69)	**30.03** (9.65)	**23.48** (6.27)	**33.36** (8.25)	**23.34** (11.92)	**42.39** (7.52)	**29.43** (10.48)

Table 12: Effect of test latent size when training latent size is fixed to 8.

Model	BabyVision	VTBench	V*	HRBench8K	SpatialEval
Qwen2.5-VL-7B	10.57	19.03	76.38	60.25	59.20
DLVR-0	13.14	23.09	76.96	62.00	59.72
DLVR-8	21.39	**29.88**	**84.29**	68.25	**67.76**
DLVR-16	**21.91**	29.82	**84.29**	68.38	67.55
DLVR-24	21.65	**29.88**	83.77	**68.50**	67.08
DLVR-32	21.13	29.62	83.25	68.12	66.93
DLVR-40	21.13	29.50	83.77	67.75	67.11
DLVR-48	21.39	29.43	83.25	67.75	66.74
DLVR-56	20.88	29.50	83.77	67.62	66.80
DLVR-64	20.62	29.30	83.25	67.38	66.62
DLVR-72	20.88	29.24	83.25	67.50	66.59
DLVR-80	20.62	29.11	82.72	67.12	66.62
DLVR-88	20.10	29.17	82.72	67.25	66.45
DLVR-96	19.85	29.04	83.25	67.00	66.37

gain. When the test latent size increases from 0 to a positive value, DLVR shows a large and consistent improvement across all benchmarks, for example on BabyVision and VTBench, confirming that the model benefits mainly from latent dynamic reasoning, rather than from parameter tuning or text-only description without latent evidence. This strongly supports the effectiveness of our design, where latent tokens serve as the intermediate space for imagining and organizing dynamic visual evidence. DLVR can truly unleash LVLM thinking in dynamic latent space, shifting from seeing to foreseeing.

The results also show that the benefit of latent reasoning saturates after a moderate number of tokens. Increasing the test latent size from 8 to 16 or 32 can further improve some dynamic-centric tasks such as BabyVision and VTBench, suggesting that richer latent imagination helps capture complex temporal and physical evolution. However, longer latent sequences do not lead to monotonic gains and may introduce unnecessary computation or noisy reasoning. Therefore, we adopt 16 latent tokens for both training and testing as a balanced default setting, which achieves strong overall performance while keeping inference efficient. These findings validate that the main performance improvement comes from the latent reasoning tokens and motivate future work on adaptive latent length selection.

E.3 Analysis of Dynamic Latent Trajectories

We investigate whether the generated latent trajectories encode temporal structure and information specific to the input. We conduct perturbation and alignment analyses on the full Video-MME benchmark [24], which is independent of the DLVR training sources. The perturbation analysis examines the role of token ordering and input correspondence, while the alignment analysis measures the correspondence between generated latent positions and the temporal progression of reference trajectories derived from real videos.

Table 13: Effect of test latent size when training latent size is fixed to 16.

Model	BabyVision	VTBench	V*	HRBench8K	SpatialEval
Qwen2.5-VL-7B	10.57	19.03	76.38	60.25	59.20
DLVR-0	12.89	23.35	78.01	62.63	61.05
DLVR-8	20.87	28.91	83.77	67.88	67.31
DLVR-16	21.39	29.43	**85.34**	**68.75**	**67.53**
DLVR-24	21.90	**30.07**	**85.34**	68.63	67.27
DLVR-32	**22.16**	29.94	84.81	68.50	67.09
DLVR-40	21.64	30.01	84.29	68.50	66.98
DLVR-48	21.64	29.81	84.29	68.25	67.24
DLVR-56	21.13	29.50	84.81	68.13	66.80
DLVR-64	20.61	29.36	84.81	67.75	66.53
DLVR-72	20.87	29.36	**85.34**	67.13	66.75
DLVR-80	21.13	29.30	**85.34**	67.13	66.21
DLVR-88	20.36	29.04	84.81	67.75	66.30
DLVR-96	20.61	28.84	84.29	67.62	66.49

Table 14: Effect of test latent size when training latent size is fixed to 32.

Model	BabyVision	VTBench	V*	HRBench8K	SpatialEval
Qwen2.5-VL-7B	10.57	19.03	76.38	60.25	59.20
DLVR-0	13.40	23.22	77.49	62.88	60.80
DLVR-8	21.65	29.62	84.29	67.75	66.98
DLVR-16	22.16	30.01	**84.81**	68.25	67.37
DLVR-24	22.16	30.27	83.77	**68.62**	67.04
DLVR-32	**22.42**	**30.34**	83.25	68.50	66.95
DLVR-40	21.91	30.08	83.77	68.12	**67.49**
DLVR-48	21.65	29.88	84.29	68.25	66.80
DLVR-56	21.13	29.62	83.77	67.75	66.80
DLVR-64	21.91	29.43	83.77	68.50	66.59
DLVR-72	21.39	29.50	84.29	68.00	66.74
DLVR-80	21.39	29.17	83.77	67.62	66.21
DLVR-88	20.62	29.17	83.24	67.25	66.46
DLVR-96	20.36	29.11	83.24	67.50	66.46

Latent Trajectory Perturbations. We perturb the generated trajectories by reversing the token order, randomly shuffling the tokens, or replacing the entire trajectory with one generated from another sample. All DLVR variants receive only the first frame as visual input, while the backbone with full video input provides a reference. As shown in Table 15, reversing and shuffling the latent tokens consistently reduce performance across all categories, with overall accuracy decreasing from 44.81 to 39.22 and 35.93, respectively. These results indicate that the benefits depend on the ordering of latent tokens rather than merely their presence as an unordered collection of features. Replacing the trajectory with one from another sample causes a larger degradation, reducing overall accuracy to 29.22. This further supports the importance of information specific to the input, rather than the trajectory acting as a generic latent prefix.

Temporal Alignment Beyond Cosine Similarity. We further quantify the correspondence between generated latent trajectories and reference trajectories derived from real videos using a metric beyond cosine similarity. For each video, we generate a latent trajectory $\mathbf{z} = [z_1, \ldots, z_N]$ with $N = 16$ and construct a reference trajectory $\mathbf{t} = [t_1, \ldots, t_N]$ from the corresponding real video using MNTS and NATQ. We jointly standardize the two trajectories along each feature dimension, obtaining standardized representations $\tilde{z}_i$ and $\tilde{t}_j$. We then compute the 16×16 Euclidean distance matrix

$$D_{ij} = \left\| \tilde{z}_i - \tilde{t}_j \right\|_2, \qquad i, j \in \{1, \ldots, N\}. \tag{66}$$

The Hungarian algorithm identifies a bijective correspondence that minimizes the total matching distance $\pi^\star = \arg\min_{\pi \in \mathcal{S}_N} \sum_{i=1}^{N} D_{i,\pi(i)}$, where $\mathcal{S}_N$ denotes the set of permutations of $\{1, \ldots, N\}$. The resulting matched reference indices form the sequence $\mathbf{M} = [M_1, \ldots, M_N]$, where $M_i = \pi^\star(i)$ denotes the temporal index of the reference target matched to latent token z_i.

Table 15: **Effect of latent trajectory perturbations on Video-MME.** Bold values indicate the best results under static input. Unpaired replacement uses a trajectory generated from another sample.

Model	Knowledge	Film & Tele.	Sports	Artistic	Life	Multilingual	Overall
Full Video Input							
Qwen2.5-VL-7B	63.33	65.00	60.22	63.06	57.78	53.33	61.37
Static Input (First Frame Only)							
Qwen2.5-VL-7B	33.09	31.67	23.33	41.67	31.11	30.00	31.85
DLVR ($N = 0$)	36.42	33.06	25.56	41.39	31.90	31.11	33.59
DLVR ($N = 16$)	**46.91**	**46.11**	**42.00**	**46.67**	**42.38**	**44.44**	**44.81**
DLVR ($N = 16$) + reverse	41.35	42.78	37.56	43.61	33.17	38.89	39.22
DLVR ($N = 16$) + shuffle	39.38	36.11	31.33	41.39	31.75	34.44	35.93
DLVR ($N = 16$) + unpaired replacement	31.23	28.06	22.00	37.78	28.10	25.56	29.22

For each video, we calculate the Spearman correlation between $\mathbf{M}$ and the canonical latent order $[1, \ldots, N]$. Since both sequences contain distinct ranks, this correlation is given by

$$\rho = 1 - \frac{6 \sum_{i=1}^{N} (M_i - i)^2}{N(N^2 - 1)}. \tag{67}$$

The mean correlation across videos is 0.762, with a 95% confidence interval of $[0.730, 0.791]$. After randomly shuffling $\mathbf{z}$ and repeating the same procedure, the mean correlation decreases to 0.086, with a 95% confidence interval of $[0.068, 0.104]$. The substantial gap supports a positive correspondence between the generated latent order and the temporal stages represented by the reference trajectories, whereas the shuffled control exhibits much weaker correspondence. Together, these analyses provide evidence that different latent positions capture information associated with different stages of dynamic evolution. The generated trajectories encode ordered information specific to the input and align with the temporal progression represented by references derived from real videos.

F Case Studies

As shown in Figure 7, the dynamic cases show that DLVR can actively imagine how a scene will evolve before final answering, **<Latent>** denotes the continuously generated latent embeddings. For physical reasoning tasks such as JENGA and block tower, DLVR infers support relations, center of mass, and collapse directions through latent dynamic rollout. For pattern and real-world dynamic scenes, it continues the visual transformation or anticipates human behavior, such as pedestrians crossing after the signal changes. These examples demonstrate that DLVR does not merely describe the current image, but builds a temporally coherent mental trajectory to support the final decision.

DLVR is also effective on static understanding tasks where the answer depends on fine-grained visual evidence. In real-world static and OCR cases, the model implicitly simulates useful visual operations, such as zooming into a shop sign or highlighting the target menu item, before producing the answer. This suggests that dynamic latent reasoning can serve as an internal visual workspace, helping the model focus on key regions and strengthen perception even when the task itself is not explicitly dynamic. Together, these cases show that DLVR improves both future-oriented reasoning and detail-sensitive static recognition.

G Limitations and Future Work

Following prior latent reasoning works, DLVR adopts a simple and effective fixed latent size for both training and inference [36, 53, 70, 79]. While this design offers stable performance and controllable computational cost, it may not fully match the diverse reasoning demands of different tasks. Some problems require only a few latent steps to capture local visual evidence, whereas others may benefit from deeper dynamic imagination to model long-range temporal evolution, physical interactions, or complex spatial transformations. Future work could explore adaptive latent length selection, allowing the model to dynamically decide when to enter, continue, or stop latent reasoning according to task difficulty and visual complexity. This would further improve the efficiency and flexibility of dynamic latent visual reasoning.

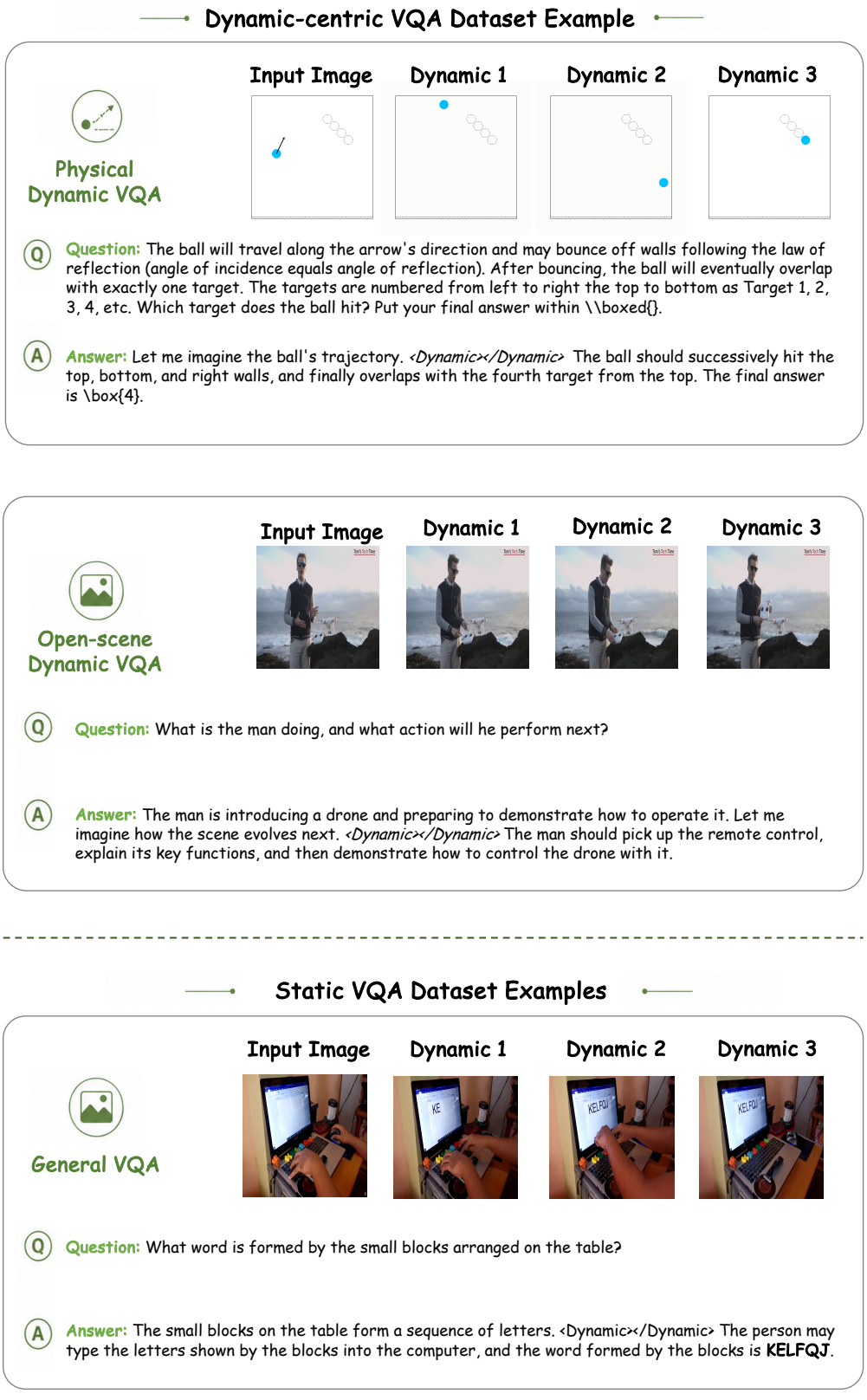


Figure 6: Representative examples from our training dataset.

Case Studies of DLVR Across Diverse Tasks

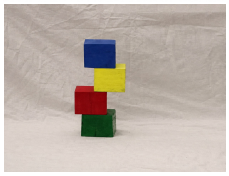
1 JENGA



Question: How many of the red bricks would fall off the table if the dark grey brick wasn't there?

Answer: Let me imagine how the tower would evolve if that support were removed. <Dynamic><Latent></Dynamic> Without the dark grey brick, the tilted red bricks above it would lose their support and start sliding downward, which would also destabilize the upper red bricks resting on them. The right-side stack remains mostly supported by the table, while the upper leaning part collapses outward. Therefore, about three red bricks would fall off the table. The final answer is 3.

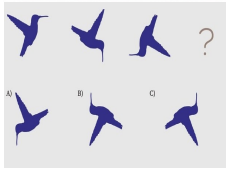
2 Block Tower



Question: : Will this block tower fall?

Answer: The tower is highly unbalanced because the upper blocks are shifted away from the center of support. Let me imagine how the structure evolves under gravity. <Dynamic><Latent></Dynamic> The blue block pushes the top of the tower leftward, while the yellow and red blocks create an offset chain of support. As the center of mass moves outside the green base, the tower should start tipping to the right and then collapse. Therefore, yes, this block tower will fall.

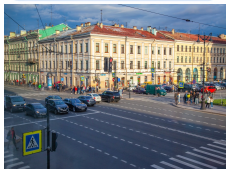
3 BabyVision



Question: : From the given three options, choose the most suitable one to fill in the question mark, so that it presents a certain regularity.

Answer: Let me imagine the next transformation in the pattern. <Dynamic><Latent></Dynamic> The first three silhouettes rotate progressively, so the missing one should continue the same rotational direction and place the beak upward with the wings aligned diagonally. Among the options, B best matches this continuation. The final answer is B.

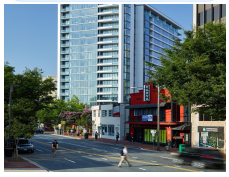
4 RealWorld Dynamic



Question: : What are the pedestrians waiting at the crosswalk likely to do next?

Answer: The pedestrians are waiting at the crosswalk while the pedestrian signal still shows about 2 seconds of red light remaining. Let me imagine how the scene evolves next. <Dynamic><Latent></Dynamic> Once the countdown ends and the pedestrian signal turns green, and the pedestrians will start crossing the street along the zebra crossing.

5 RealWorld Static



Question: : What's the content of the shop sign at the bottom right corner of the picture?

Answer: Let me imagine zooming into that region to inspect the sign more clearly. <Dynamic><Latent></Dynamic> After focusing on the bottom-right shopfront, the enlarged sign reads "International School of Languages." Therefore, the content of the shop sign is International School of Languages.

6 OCR Task



Question: : How much is the Lemon Ricotta Pancakes?

Answer: The question asks for the price of Lemon Ricotta Pancakes on the menu. Let me imagine locating and highlighting the relevant menu entry. <Dynamic><Latent></Dynamic> The item appears in the "Wildflower Signature Pancakes" section, where the corresponding price is circled as $6.99. Therefore, the Lemon Ricotta Pancakes cost $6.99.

Figure 7: Case studies of DLVR.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly summarize the paper's contributions, including the formulation of dynamic latent visual reasoning, the construction of DLVR-SFT-100K and DLVR-RL-4K with MNTS and NATQ, the two-stage supervised training pipeline, CDLPO, and empirical validation on dynamic-centric and general visual reasoning benchmarks, accurately reflecting the scope and claims of the work.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly discuss the limitation in Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results are formally stated with clear assumptions and complete proofs in Appendix B and Appendix B.2.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the detailed hyperparameter settings for reproducing the results in Sec. 4 and Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will release all resources upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all relevant experimental details, including datasets, backbone models, and hyperparameter choices in Sec. 4 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: All experiments use deterministic decoding with do_sample=False. As a result, there is no source of randomness in the evaluation, and error bars or statistical significance measures are not applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We indicate the computation resources in Sec. 4 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work fully complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer:[N/A]

    Justification: This work focuses on algorithmic improvements for reasoning in LVLMs and does not involve deployment in sensitive contexts or affect human subjects directly. Consequently, discussion of broader positive or negative societal impacts is not applicable.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We did not use any unsafe data in this manuscript.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and models used in this work are publicly available and properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We will release our dataset and trained model weights upon acceptance.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Sec. 3.2 describes our use of LLM tools for dataset construction.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.