INTERMEDIATE-ANNOTATION-FREE LATENT VISUAL REASONING VIA MULTI-LEVEL PSEUDO-LABELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent Visual Reasoning (LVR) enables Large Vision-Language Models (LVLMs) to perform intermediate reasoning in continuous representations, avoiding the need to verbalize visual information through discrete text. However, existing LVR methods either rely on costly intermediate annotations or learn latent tokens indirectly from final answer supervision, providing limited guidance on what the latent tokens should encode. We introduce Multi-level Pseudo-labeled Latent Visual Reasoning (MP-LVR), which derives explicit latent supervision directly from the pretrained LVLM's own internal signals without requiring intermediate human annotations or external teacher models. MP-LVR exploits layer-wise query-relevance attention patterns to adaptively construct multiple pseudo-labels, providing distinct supervision to individual latent tokens. In addition, we introduce Stochastic Information-Path Suppression, which reduces direct answer access to image and query tokens during training and encourages answer generation to utilize the latent tokens. Experiments demonstrate consistent improvements across diverse visual reasoning benchmarks and model configurations, while further analyses confirm that the latent tokens are actively utilized during answer generation.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) have shown strong capabilities across various visual reasoning tasks, often relying on Chain-of-Thought (CoT) prompting to generate intermediate reasoning steps before producing a final answer. However, conventional CoT compresses rich visual information into discrete text, potentially losing fine-grained details that are difficult to express textually. This mismatch motivates *Latent Visual Reasoning* (LVR) (Li et al., 2026a;b; Wu et al., 2026b), which performs intermediate reasoning through **latent tokens**, i.e., learnable intermediate tokens inserted between the query and answer tokens. These latent tokens provide continuous intermediate representations that can preserve visual information without requiring complete verbalization.

A central challenge in LVR is how to provide meaningful learning signals to latent tokens. Existing approaches commonly supervise latent tokens using intermediate annotations, such as bounding boxes (Li et al., 2026a; Viveiros et al., 2026; Cui et al., 2026), segmentation masks (Wei et al., 2026), depth maps (Qin et al., 2025; Bigverdi et al., 2025), or other helper images (Yang et al., 2026b; Wang et al., 2026a). While such supervision explicitly guides latent-token representations, it incurs substantial annotation costs, limits scalability, and may introduce errors when the annotations are generated by external models. Moreover, predefined intermediate annotations may not be equally useful across samples, as different problems require different visual evidence. In contrast, recent implicit LVR methods (Li et al., 2026b; Zhang et al., 2026) avoid intermediate annotations and optimize latent tokens only through the final-answer prediction objective. While annotation-free, such indirect supervision provides limited control over what information the latent tokens encode.

These limitations raise a question: *Can we provide explicit supervision to latent tokens without relying on external intermediate annotations?* We explore this possibility by deriving supervision intrinsically from the LVLM itself. Pretrained LVLMs already exhibit visual reasoning capabilities, suggesting that their internal representations encode useful query-conditioned visual information. In particular, we observe that different LVLM layers exhibit distinct query-relevance attention patterns and varying degrees of attention concentration. Such layer-wise diversity provides a natural basis

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

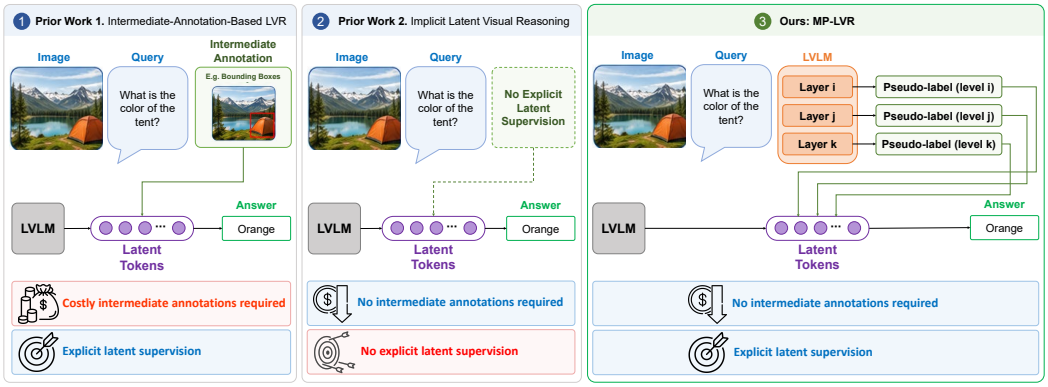


Figure 1: **Comparison of latent visual reasoning paradigms.** Intermediate-annotation-based LVR requires costly intermediate visual annotations, whereas implicit LVR lacks explicit supervision for latent tokens. MP-LVR instead derives multi-level pseudo-labels from the LVLM's internal representations, enabling explicit latent-token supervision without intermediate annotations.

for multi-level supervision, encouraging individual latent tokens to capture complementary query-relevant visual information rather than learning redundant representations across latent tokens.

Based on this insight, we introduce **Multi-level Pseudo-labeled Latent Visual Reasoning (MP-LVR)**, which provides explicit supervision to individual latent tokens using the base LVLM's own layer-wise query-relevance signals, without requiring intermediate human annotations or external teacher models. Specifically, MP-LVR adaptively selects layers with diverse query-relevance concentration levels for each sample and uses their attention patterns to construct distinct pseudo-labels for latent tokens. However, explicitly supervising latent tokens does not guarantee that answer generation will actually utilize them, since the model can still bypass the latent-token pathway through direct access to image and query tokens. To address this issue, we introduce *Stochastic Information-Path Suppression*, which stochastically suppresses direct answer-to-image and answer-to-query attention during training. This encourages greater reliance on the latent-token pathway. Together, these components enable latent tokens to serve as meaningful intermediate states for visual reasoning within a single-stage framework. Figure 1 contrasts MP-LVR with prior LVR paradigms.

Extensive experiments show that MP-LVR achieves strong performance across diverse visual reasoning and hallucination benchmarks, without relying on externally provided intermediate annotations. Analyses further demonstrate that the learned latent tokens encode meaningful and non-redundant representations and are functionally utilized during answer generation.

Our main contributions are summarized as follows:

- We propose **Multi-level Pseudo-labeled Latent Visual Reasoning (MP-LVR)**, which explicitly supervises latent tokens using multi-level pseudo-labels constructed through adaptive layer selection based on attention concentration, without relying on external intermediate annotations.
- We also introduce *Stochastic Information-Path Suppression*, which suppresses the answer token's direct access to the image and query tokens so that answer generation relies more on the latent-token pathway within a single-stage training framework.
- We demonstrate that MP-LVR consistently improves visual reasoning across diverse benchmarks and model configurations, with extensive latent-token analyses and visualizations confirming that the learned latent tokens are distinct and actively utilized during answer generation.

2 Related Works

2.1 Latent Visual Reasoning

Recent Latent Visual Reasoning (LVR) methods can be categorized by how they supervise latent reasoning tokens. One line of work uses explicit intermediate annotations, including bounding

boxes (Li et al., 2026a; Viveiros et al., 2026), subgoal helper images (Wang et al., 2026a; Yang et al., 2026b; Ray et al., 2026), VAE latents (Bigverdi et al., 2025), rendered intermediate images (Wang et al., 2026b), segmentation masks (Wei et al., 2026), depth maps (Qin et al., 2025), sketches (Tong et al., 2025), and region-level semantic descriptions (Cui et al., 2026; Wang et al., 2026c; Huang et al., 2026). These methods constrain latent tokens using predefined intermediate signals, but such signals must be produced by human annotators or external models, incurring substantial cost.

Another line of work avoids intermediate annotations and learns latent tokens implicitly through the final task objective. LIVR (Li et al., 2026b) introduces a visual information bottleneck that completely blocks direct answer-to-image information flow, while CrystaL (Zhang et al., 2026) uses corrupted and original visual inputs to encourage latent-token-mediated reasoning. Since the latent-token representations are not directly supervised, it remains unclear what information the latent tokens encode and whether these representations are actually useful for answer generation.

More recent approaches provide explicit latent-token supervision using auxiliary teacher models. LaViT (Wu et al., 2026b) aligns latent-token representations with a teacher model's hidden states and latent-to-image attention, VaLR (Jeon et al., 2026) aligns latent-token representations with visual features extracted from external vision encoders such as π^3, DINOv3, and SigLIPv2, and LUT (Kang et al., 2026) transfers latent-token representations from a teacher trained with a visual information bottleneck. These methods reduce dependence on external intermediate annotations but still rely on auxiliary teacher models. In contrast, our **MP-LVR** constructs multi-level pseudo-labels from the base LVLM's own layer-wise query-relevance signals, providing explicit supervision to individual latent tokens without intermediate human annotations or external teacher models.

2.2 Layer-wise Visual Representations in LVLMs

Layer-wise analyses of LVLMs reveal substantial layer-dependent differences in how visual information is represented and utilized. Prior work reports progressive refinement of object representations (Neo et al., 2025), visual enrichment and semantic refinement in middle layers (Jiang et al., 2025), and distinct stages of visual grounding, semantic reasoning, and answer preparation (Yu & Lee, 2025). Other studies similarly show that different layers capture distinct aspects of image content and query-relevant information, while visual information is progressively transformed across layers (Zhang et al., 2025b). Beyond visual-token representations, query tokens have also been shown to progressively accumulate visual information across layers (Kaduri et al., 2025), while modality-specific information changes systematically across layers (Wu et al., 2026a). Together, these layer-wise analyses indicate that different LVLM layers provide distinct visual signals.

Such layer-wise properties have also been exploited in practical applications. For efficient visual processing, FastV (Chen et al., 2024), Visual Tokens Withdrawal (Lin et al., 2025), PyramidDrop (Xing et al., 2025), and SparseVLM (Zhang et al., 2025a) leverage layer-dependent changes in visual-token importance or redundancy for visual token pruning. Beyond efficiency, layer-wise visual signals have been used to mitigate hallucination by identifying informative middle-layer visual representations (Jiang et al., 2025), rebalancing perception and reasoning across layers (Lu et al., 2026), and reinforcing intermediate layer visual anchors to correct visual drift in later layers (Yang et al., 2026a). These studies demonstrate that layer-wise internal signals can serve as actionable cues for improving LVLM behavior. In contrast to these applications, MP-LVR exploits layer-wise query-relevant visual focus patterns as supervision signals for learning latent tokens.

3 Methodology

We propose **Multi-level Pseudo-labeled Latent Visual Reasoning (MP-LVR)**, which trains latent tokens as intermediate reasoning states using the base LVLM's own internal signals and encourages the latent-token utilization during answer generation. MP-LVR inserts latent tokens between the query and answer tokens and trains them through two components: *multi-level pseudo-labeling*, which constructs pseudo-labels from layer-wise query-relevance attention patterns to explicitly supervise individual latent tokens, and *stochastic information-path suppression*, which encourages answer generation to rely more on these latent tokens. Figure 2 illustrates the overall framework.

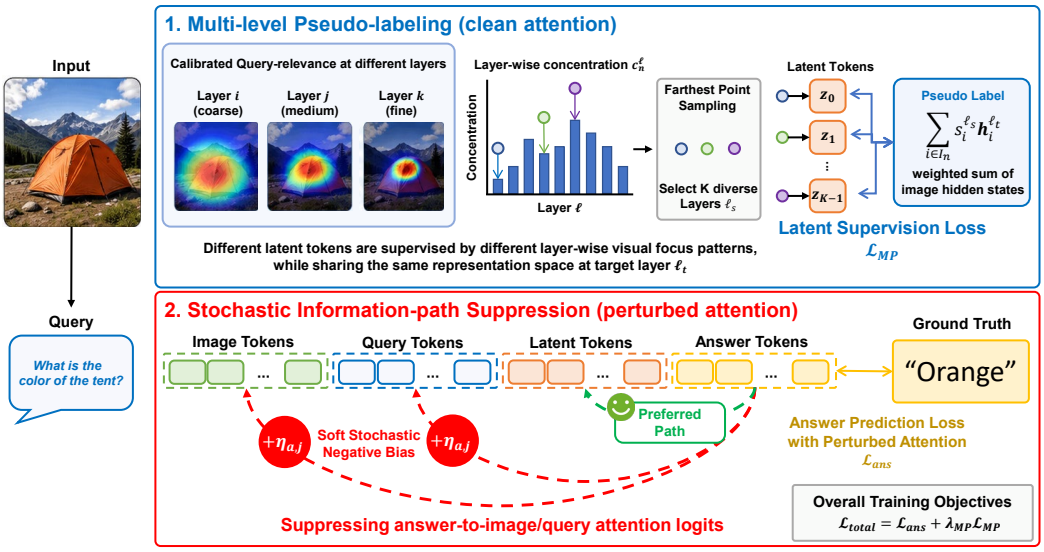


Figure 2: **Overview of MP-LVR.** MP-LVR adaptively selects layers with diverse query-relevance concentration and constructs *multi-level pseudo-labels* to supervise latent tokens. *Stochastic information-path suppression* reduces direct answer access to image and query tokens to promote latent-token utilization. Both components are combined within a single-stage training framework.

3.1 Multi-level Pseudo-labeling

Our goal is to explicitly guide latent tokens toward query-relevant visual representations using the LVLM's own internal signals. Since pretrained LVLMs already exhibit visual reasoning capabilities, we hypothesize that their internal representations contain useful query-conditioned visual information for latent-token supervision. Based on this hypothesis, we examine how such visual information varies across layers, using query-to-image attention as a proxy for query-relevant visual focus, and leverage these layer-wise signals to supervise individual latent tokens.

For sample n and layer ℓ, let $\mathbf{a}_{n,\ell}^q$ denote the head-averaged query-to-image attention, further averaged over query tokens. Since raw query-to-image attention may also reflect query-independent visual saliency induced by positional bias (Wu et al., 2025; Wang et al., 2026d), we calibrate it using $\mathbf{a}_{n,\ell}^v$, the head-averaged attention from the *vision_end* token to image tokens. Because *vision_end* precedes the query and marks the end of the visual sequence, it serves as a query-independent reference. Then, we define the calibrated query-relevance distribution as $\mathbf{s}_n^\ell = \mathcal{P}_+(\mathcal{P}(\mathbf{a}_{n,\ell}^q) - \mathcal{P}(\mathbf{a}_{n,\ell}^v))$, where $\mathcal{P}(\cdot)$ normalizes over image tokens and $\mathcal{P}_+(\cdot)$ clips negative residuals to zero and renormalizes them. Its i-th entry $s_{n,i}^\ell$ denotes the calibrated query-relevance score of image token i.

Using this calibrated distribution, we examine how query-relevant visual focus changes across layers. The top of Figure 3 suggests that attention across layers often remains centered on similar query-relevant objects, while varying in how broadly it covers their surrounding regions. To quantify the concentration of these patterns, we use normalized entropy for image $\mathcal{I}_n$:

$$\mathcal{H}_n^\ell = -\sum_{i \in \mathcal{I}_n} s_{n,i}^\ell \log s_{n,i}^\ell, \qquad c_n^\ell = 1 - \frac{\mathcal{H}_n^\ell}{\log N_n}, \tag{1}$$

where $N_n = |\mathcal{I}_n|$. The concentration score $c_n^\ell \in [0, 1]$ measures how focused the query-relevance distribution is over image tokens, with larger values indicating more localized visual focus. As shown in the bottom of Figure 3, the concentration varies substantially across layers. This variation provides a natural axis for diversifying latent-token supervision. Layers with different concentration levels capture visual evidence ranging from broader contextual regions to fine-grained localized details around query-relevant regions, providing distinct supervision signals that can encourage complementary latent-token representations while reducing redundancy across latent tokens.

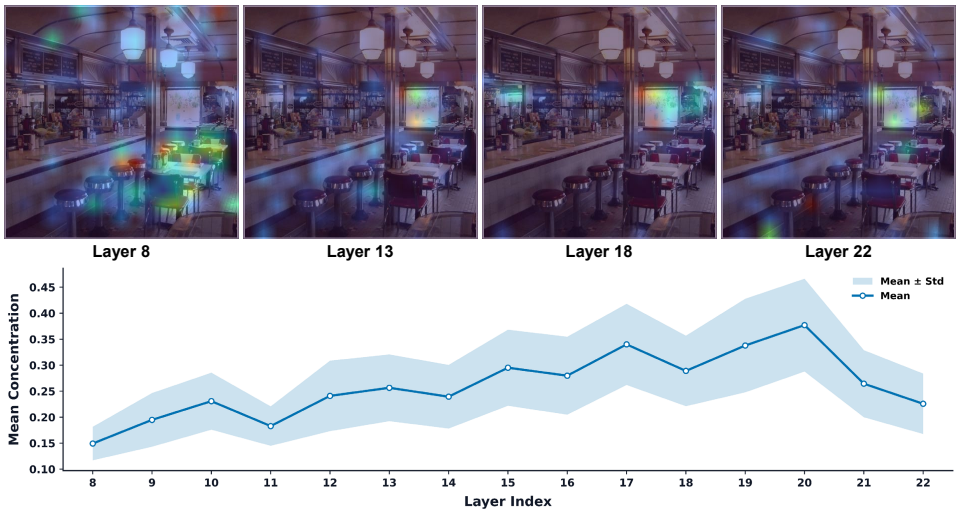


Figure 3: **Layer-wise query-relevance attention patterns.** Top: Query-relevance score maps for "How many windows are in the image?", showing distinct visual focus across layers. Bottom: Mean query-relevance concentration over 100 CV-Bench-2D samples, revealing layer-wise variation.

Based on this motivation, we construct *Multi-level Pseudo-labels* by adaptively selecting layers with diverse concentration levels for each sample. Given K latent tokens, we define the concentration distance as $d_n(\ell_1, \ell_2) = |c_n^{\ell_1} - c_n^{\ell_2}|$ and apply farthest point sampling (FPS) over candidate layers to select K layers. FPS is initialized with the minimum- and maximum-concentration layers, and the selected layers are sorted by depth, with $\ell_{n,k}$ denoting the layer assigned to the k-th latent token.

After selecting a layer for each latent token, we construct its pseudo-label by weighting image representations. The selected layer determines the visual weighting for each latent token, while a fixed target layer provides the hidden states used to construct pseudo-labels in a shared representation space. Specifically, the query-relevance scores from $\ell_{n,k}$ determine the weights over image tokens, while image hidden states are taken from a fixed target layer ℓ_t. This keeps all latent targets in a shared representation space while preserving layer-wise variation in visual focus. Let $\mathbf{h}_{n,k}^{\ell_t}$ and $\mathbf{h}_{n,i}^{\ell_t}$ denote the hidden states of latent token k and image token i, respectively. We optimize

$$\mathcal{L}_{\text{MP}}^{n} = \sum_{k=0}^{K-1} \left[1 - \cos\left(\mathbf{h}_{n,k}^{\ell_t}, \text{sg}\left(\sum_{i \in I_n} s_{n,i}^{\ell_{n,k}} \, \mathbf{h}_{n,i}^{\ell_t} \right) \right) \right], \tag{2}$$

where $\text{sg}(\cdot)$ denotes stop-gradient and $\cos(\cdot, \cdot)$ denotes cosine similarity. The loss is averaged over the mini-batch. Consequently, each latent token is guided toward a layer-conditioned query-relevant visual representation, while all pseudo-labels are constructed in the same target-layer space.

3.2 Stochastic Information-Path Suppression

While explicit supervision guides what latent tokens encode, it does not guarantee their use during answer generation. Prior studies report similar representations across latent tokens or samples (Li et al., 2026c) and limited degradation after removing latent tokens (Guo et al., 2026), suggesting that answer tokens may bypass the latent pathway through direct access to image and query tokens.

To encourage latent utilization, we introduce *Stochastic Information-Path Suppression* (SIPS), which applies stochastic negative perturbations to answer-to-image and answer-to-query attention logits during training, weakening direct information flow. Let $\mathcal{A}_n$, $\mathcal{I}_n$, and $\mathcal{Q}_n$ denote the answer, image, and query token positions. For $a \in \mathcal{A}_n$ and $j \in \mathcal{I}_n \cup \mathcal{Q}_n$, we modify the attention logit as

$$\tilde{S}_{n,a,j}^{\ell} = S_{n,a,j}^{\ell} + \eta_{n,a,j}, \tag{3}$$

where

$$\eta_{n,a,j} = \min(0, \mu_n + \epsilon_{n,a,j}), \qquad \mu_n \sim \mathcal{U}(-10, 0), \quad \epsilon_{n,a,j} \sim \mathcal{N}(0, 1). \tag{4}$$

Table 1: **Comparison across diverse visual reasoning and hallucination benchmarks.** MP-LVR achieves the best overall average performance and consistently strong results across benchmarks.

	Visual Reasoning					Hallucination		
Method	**BLINK**	**CVB2D**	**CVB3D**	**V***	**RQA**	**HallB**	**POPE**	**Avg.**
Qwen2.5-VL-7B	50.87	71.18	73.08	65.97	67.58	64.46	86.34	68.50
CoVT	53.45	74.32	70.50	74.35	69.93	68.03	87.06	71.09
LVR	51.24	73.68	72.08	69.11	66.54	63.62	85.07	68.76
Vision-R1	46.76	71.76	75.83	69.11	67.19	65.30	87.98	69.13
SKILA	51.81	73.79	74.58	70.16	66.14	64.35	87.01	69.98
SFT	53.99 $\pm$ 0.7	75.55 $\pm$ 0.1	82.81 $\pm$ 0.5	72.08 $\pm$ 0.6	69.98 $\pm$ 0.8	54.86 $\pm$ 12.5	86.57 $\pm$ 2.2	70.83 $\pm$ 1.9
LIVR	55.64 $\pm$ 0.4	74.88 $\pm$ 1.7	83.03 $\pm$ 1.4	72.77 $\pm$ 1.6	69.37 $\pm$ 0.3	59.10 $\pm$ 2.6	88.95 $\pm$ 1.1	71.96 $\pm$ 1.1
CrystaL	53.95 $\pm$ 0.1	76.31 $\pm$ 0.2	82.25 $\pm$ 2.2	72.60 $\pm$ 2.0	68.67 $\pm$ 1.9	65.09 $\pm$ 1.0	88.77 $\pm$ 0.1	72.52 $\pm$ 0.3
MP-LVR (Ours)	**54.69** $\pm$ 0.2	**76.24** $\pm$ 0.6	**84.03** $\pm$ 0.7	**74.87** $\pm$ 0.5	**69.02** $\pm$ 0.5	**66.14** $\pm$ 0.2	**89.09** $\pm$ 0.1	**73.44** $\pm$ 0.2

The image and query perturbations are sampled independently. Since $\eta_{n,a,j} \leq 0$, SIPS only reduces direct answer attention to these inputs with stochastic strength, encouraging the model to rely more on the latent-token pathway without completely ignoring image and query. The suppression is applied only during training, while pseudo-labels are constructed from clean attention, preserving stable supervision signals while promoting latent-token utilization within a single-stage framework.

3.3 Overall Training Objective

Overall, MP-LVR jointly optimizes answer prediction and latent-token supervision:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ans}} + \lambda_{\text{MP}} \mathcal{L}_{\text{MP}}, \tag{5}$$

where $\mathcal{L}_{\text{ans}}$ is the next-token prediction loss over answer tokens, $\mathcal{L}_{\text{MP}}$ is the latent supervision loss, and λ_{MP} controls the strength of latent supervision. Both losses are averaged over the mini-batch.

4 Experiments

4.1 Experimental Setup

We evaluate MP-LVR on diverse visual reasoning benchmarks using Qwen2.5-VL-7B-Instruct (Bai et al., 2025) as the primary backbone, and examine generalization across scales and model families with Qwen2.5-VL-3B-Instruct and Gemma3-4B-it (Team et al., 2025). All methods use latent-token configuration for fair comparison. Unlike many prior LVR methods that rely on full fine-tuning, MP-LVR is trained with LoRA (Hu et al., 2021). Additional details are provided in the Appendix.

Baselines. We compare MP-LVR with the base model, CoVT (Qin et al., 2025), LVR (Li et al., 2026a), Vision-R1 (Huang et al., 2026), SKILA (Tong et al., 2025), Supervised Fine-Tuning (SFT), LIVR (Li et al., 2026b), and CrystaL (Zhang et al., 2026). For a fair comparison, we use the same evaluation pipeline for all methods. For CoVT, LVR, Vision-R1, and SKILA, we evaluate their official pretrained checkpoints, whereas SFT, LIVR, and CrystaL are trained on the same train data as MP-LVR. All LVR baselines use $K = 8$ latent tokens, while SFT does not use latent tokens.

Datasets. MP-LVR, SFT, LIVR, and CrystaL are trained on the same CoVT subset. We evaluate all methods on visual reasoning benchmarks: BLINK (Fu et al., 2024), CV-Bench-2D/3D (CVB2D/3D) (Tong et al., 2024), VStarBench (V*) (Wu & Xie, 2024), and RealWorldQA (RQA) (xAI, 2024), as well as hallucination benchmarks: HallusionBench (HallB) (Guan et al., 2024) and POPE (Li et al., 2023). Open-ended VQA results are also reported in the Appendix.

4.2 Main Results

We first compare MP-LVR with existing LVR methods across diverse visual reasoning and hallucination benchmarks. Table 1 shows that MP-LVR achieves the best overall performance, outperforming both intermediate-annotation-based and annotation-free baselines on average. Its gains span multiple visual reasoning benchmarks rather than being driven by a single task, while maintaining strong hallucination robustness. Notably, annotation-based methods generally lag behind annotation-free

Table 2: **Comparison across different model scales and families.** MP-LVR achieves the best average performance on two different backbones, demonstrating consistent generalization across model scales and families. LaViT is evaluated only on Qwen2.5-VL-3B due to checkpoint availability.

Method	BLINK	CVB2D	CVB3D	V*	RQA	Avg.
Qwen2.5-VL-3B-Instruct						
Base	48.92	66.54	60.33	70.68	66.01	62.50
SFT	48.71	71.24	74.08	70.16	66.80	66.20
LIVR	50.29	70.90	73.25	73.82	63.14	66.28
CrystaL	50.29	74.78	70.17	70.68	65.23	66.23
LaViT	48.55	69.99	75.33	70.16	63.66	65.54
MP-LVR (Ours)	**51.03**	**72.26**	**76.42**	**70.68**	**65.10**	**67.10**
Gemma3-4B-it						
Base	43.92	49.08	59.83	39.79	55.56	49.64
SFT	44.77	56.57	59.42	40.84	51.24	50.57
LIVR	42.35	56.98	63.42	34.56	49.54	49.37
CrystaL	45.77	53.09	59.33	38.74	51.76	49.74
MP-LVR (Ours)	**46.24**	**57.46**	**60.00**	**39.27**	**53.99**	**51.39**

Table 3: **Ablation study of MP-LVR.** We investigate the contributions of pseudo-label construction for latent-token supervision and information-path suppression for latent-token utilization.

Method	BLINK	CVB2D	CVB3D	V*	RQA	Avg.
Supervised Fine-Tuning (SFT)	54.66	75.50	82.42	71.73	69.28	70.72
+ Single-level Pseudo-Labeling	53.45	74.66	83.75	72.77	68.50	70.63
+ Multi-level Pseudo-Labeling	53.34	75.66	82.00	74.35	68.89	70.85
+ Adaptive Pseudo-Labeling	54.55	75.60	83.42	74.87	68.10	71.31
+ Image-path Suppression	54.29	75.50	83.75	74.87	69.15	71.51
+ Image/Query-path Suppression	54.92	76.39	84.42	74.87	69.54	72.03

approaches in overall performance, suggesting that predefined intermediate annotations may not provide equally useful cues for every problem. In contrast, MP-LVR achieves these results by deriving sample-specific pseudo-labels from the LVLM's own layer-wise query-conditioned visual signals, providing adaptive latent-token supervision without predefined intermediate annotations.

We further evaluate MP-LVR across model scales and families in Table 2. MP-LVR achieves the highest average performance on each backbone, outperforming SFT, LIVR, and CrystaL overall. On Qwen2.5-VL-3B, it also surpasses LaViT without requiring an external teacher model. Its consistent improvement across smaller models and a different model family indicates that MP-LVR is not specific to the primary backbone, supporting its generalizability across scales and families.

4.3 Analysis

Ablation Study. We conduct ablation studies to investigate effective pseudo-label construction and whether encouraging latent-token utilization improves performance, as reported in Table 3.

We first examine how pseudo-labels should be constructed. Single-level pseudo-labeling does not improve over SFT, suggesting that supervising all latent tokens with the same visual signal is insufficient. Multi-level pseudo-labeling performs better by assigning latent tokens supervision from distinct layer-wise query-relevance patterns, while adaptive layer selection further improves performance by selecting supervision layers adaptively for each input. This allows the pseudo-labels to reflect sample-specific variations in visual focus and reasoning requirements.

We next examine whether encouraging latent-token utilization provides additional performance benefits. Suppressing only the image path provides a modest gain, whereas jointly suppressing the image and query paths yields a larger improvement. Overall, the gains from multi-level supervision, adaptive layer selection, and joint image and query information-path suppression support the complementary roles of MP-LVR's two components in improving visual reasoning performance.

Latent-Token Analysis. We analyze the learned latent tokens in MP-LVR from three perspectives: (1) whether latent tokens are actually utilized during answer generation, (2) whether the latent tokens

Table 4: **Analysis of latent-token utilization** with and without SIPS under various latent-token attention interventions. Here, z_k denotes the k-th latent token, and Late/Early $z_{a:b}$ retains answer access only to latent tokens z_a through z_b, masking access to all other latent tokens. Mask A→L and L→L denote masking all answer-to-latent and latent-to-latent attention connections, respectively.

Method	No Mask	Late $z_{4:7}$	Late $z_{6:7}$	Early $z_{0:3}$	Early $z_{0:1}$	Mask A→L	Mask L→L
MP-LVR	72.03	71.35	40.57	29.53	0.07	0.06	18.48
MP-LVR w/o SIPS	71.31	71.50	70.81	70.99	35.51	4.75	18.72

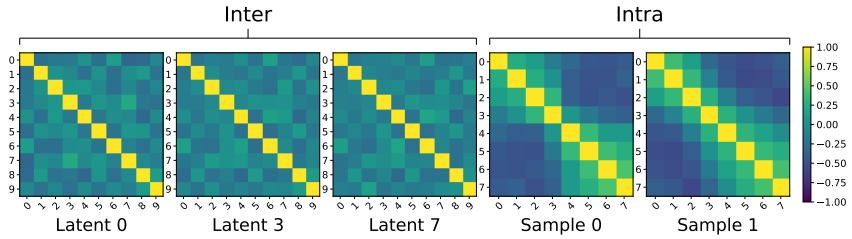


Figure 4: **Inter- and intra-sample latent-token similarity on 10 CV-Bench-2D samples.** Low off-diagonal similarities indicate sample-dependent and token-specific latent representations.

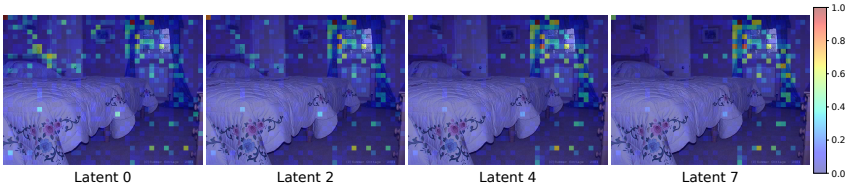


Figure 5: **Latent-wise visualization of latent-to-image attention** for the query "How many curtains are in the image?". Each panel shows the normalized attention from an individual latent token to image tokens, illustrating how different latent tokens attend to distinct regions of the image.

exhibit the causal information integration and refinement expected from an intermediate reasoning process, and (3) what information is encoded in individual latent tokens.

Reliance on Latent Tokens. We first examine whether answer generation actually depends on the latent tokens by intervening on answer-to-latent attention. As shown in Table 4, completely masking this attention collapses performance from 72.03 to 0.06, while progressively restoring access to latent tokens recovers performance (e.g., Early $z_{0:1}$ to $z_{0:3}$ or Late $z_{6:7}$ to $z_{4:7}$). These results indicate that answer generation strongly relies on latent tokens. Consistent with this, Figure A1 shows substantial answer-to-latent attention in the early-to-middle layers. We further compare MP-LVR with a variant trained without SIPS. Without SIPS, the model tolerates partial latent-token restrictions better, retaining 70.81 with access only to latent tokens $z_{6:7}$, compared with 40.57 for MP-LVR, although complete removal causes severe degradation. This indicates that latent tokens are functionally utilized in both variants, but SIPS makes answer generation rely on them more strongly.

Causal Information Integration and Refinement across Latent Tokens. We next investigate whether the latent-token sequence exhibits causal information integration and refinement. Table 4 provides evidence for this behavior in two ways. First, retaining only the last two latent tokens preserves a performance of 40.57, compared with 0.07 when retaining only the first two tokens, suggesting that later latent tokens contain more answer-relevant information. However, the gap from using all latent tokens indicates that complementary information remains distributed across the latent tokens rather than being fully absorbed by later tokens. Second, masking latent-to-latent attention reduces performance from 72.03 to 18.48, showing that later latent tokens critically depend on information propagated from preceding latent tokens. Together, these results suggest that preceding information is integrated and progressively refined into more answer-relevant representations along the latent-token sequence. Overall, the latent-token sequence appears to function as a causal information-propagation pathway rather than as a set of independent intermediate states. Figure A2 visualizes this pathway, showing that later latent tokens attend strongly to their immediately preceding token.

Information Encoded in Latent Tokens. We examine whether the learned latent tokens encode distinct and meaningful visual information. We first analyze inter- and intra-sample representation

Table 5: **Comparison of layer selection strategies for constructing multi-level pseudo-labeling.** Selecting diverse layers is effective, with the adaptive strategy achieving the best performance.

Layer Selection Strategy	BLINK	CVB2D	CVB3D	V*	RQA	Avg.
Random	48.29	74.56	29.33	72.77	68.10	58.61
Fixed-8 [8,8,8,8,8,8,8,8]	53.39	74.85	79.75	73.30	68.89	70.04
Fixed-16 [16,16,16,16,16,16,16,16]	53.13	73.28	83.00	74.87	68.76	70.61
Fixed-21 [21,21,21,21,21,21,21,21]	53.39	74.15	84.25	74.35	68.37	70.90
Forward [8,10,12,14,16,18,20,21]	54.34	76.64	84.08	73.82	69.02	71.58
Reverse [21,20,18,16,14,12,10,8]	54.55	77.56	84.00	74.35	68.24	71.74
Adaptive	**54.92**	**76.39**	**84.42**	**74.87**	**69.54**	**72.03**

similarity. Inter-sample similarity compares the same latent token across different samples, while intra-sample similarity compares different latent tokens within each sample. Figure 4 shows low off-diagonal similarities in both cases, indicating that MP-LVR learns sample-dependent and token-specific representations rather than collapsing latent tokens into similar states. This is consistent with the goal of multi-level pseudo-labeling, which provides distinct supervision signals to encourage complementary latent-token representations. We further visualize latent-to-image attention in Figure 5. Different latent tokens attend to distinct image regions, with earlier tokens showing broader query-relevant focus and later tokens concentrating on localized details, reflecting the intended effect of query-relevance-based multi-level supervision. Together, these results indicate that MP-LVR learns distinct, complementary, query-relevant latent-token representations.

Layer Selection Strategy. We analyze different layer selection strategies for constructing multi-level pseudo-labels. Based on our layer-wise analysis, we restrict candidate layers to the middle range (layers 8–21), as very early and very late layers provide less informative visual signals.

As shown in Table 5, using a single layer performs worse than using multiple layers, suggesting that diverse layer-wise signals are important for effective multi-level pseudo-labels. The comparable results of forward and reverse schedules further indicate that this diversity matters more than layer ordering. Our adaptive strategy selects layers separately for each sample based on query-relevance concentration, enabling sample-specific supervision and achieving the best overall performance.

4.4 Computational Cost.

We compare the computational cost of MP-LVR with implicit LVR baselines in Table 6. MP-LVR requires less training time than LIVR by avoiding its two-stage training and consumes less memory than CrystaL, which jointly processes two images. Overall, MP-LVR offers a favorable trade-off between training time and memory usage through single-stage training.

Table 6: **Comparison of computational cost.**

Method	Batch / Accum.	GPU Time (h)	GPU Memory (GB)
SFT	(8 / 1)	4.27	29.96
LIVR	(8 / 1)	12.37	35.60
CrystaL	(1 / 8)	10.64	46.37
MP-LVR	(8 / 1)	6.30	41.86

5 Conclusion

Latent visual reasoning (LVR) offers a promising alternative to text-based intermediate reasoning, but existing approaches either rely on costly intermediate annotations or provide only indirect supervision to latent tokens. To address these limitations, we introduce **Multi-level Pseudo-labeled Latent Visual Reasoning (MP-LVR)**, which explicitly supervises latent tokens without intermediate human annotations or external teacher models. MP-LVR constructs adaptive multi-level pseudo-labels from the base LVLM's own layer-wise query-relevance signals and assigns them to latent tokens. We further introduce *Stochastic Information-Path Suppression* to encourage answer generation to actively utilize these latent tokens rather than bypassing them through direct image and query pathways. Extensive experiments show that MP-LVR achieves strong performance across diverse visual reasoning benchmarks and generalizes across model scales and families. Our analyses further validate the effectiveness of its components and show that the learned latent tokens are distinct and functionally utilized during answer generation. Overall, MP-LVR opens a new direction for leveraging pretrained LVLMs themselves as a source of supervision for latent visual reasoning.

AI USE STATEMENT

AI tools were used to assist with writing refinement and literature search. For writing, AI was used after the manuscript was drafted to identify and revise grammatical, stylistic, and clarity-related issues. For literature search, AI was used to help identify potentially relevant prior work. All suggested papers were independently verified to exist, read by the authors, and cited only after their relevance and content were confirmed. AI tools were not otherwise involved in the research process, including methodology or experimental design, hypothesis formulation or refinement, method implementation, interpretation of experimental results, or development of the theoretical or modeling framework.

REPRODUCIBILITY STATEMENT

Detailed experimental settings and implementation details are provided in the Appendix, including training configurations, hyperparameters, and evaluation protocols. The source code is also provided with the paper to facilitate reproduction of our experiments.

REFERENCES

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.

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

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

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Jin Cui, Xinyue Long, Xunyong Zhang, Yadong Zhang, Chuanchang Su, Jingye Gan, Boran Zhao, and Pengju Ren. Retrieve, integrate, and synthesize: Spatial-semantic grounded latent visual reasoning. *arXiv preprint arXiv:2605.07106*, 2026.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 2024.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

Jiawei Guo, Yu Chen, Xiang Wang, Shuai Li, Xinpei Zhao, Huaxing Liu, Shuai Dong, Feifei Zhai, and Yu Zhou. Beyond visual memory: Mechanistic diagnostics of latent visual reasoning. *arXiv preprint arXiv:2606.01287*, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Zhe Xu, Yao Hu, Shaohui Lin, et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 63794–63812, 2026.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Byungwoo Jeon, Yoonwoo Jeong, Hyunseok Lee, Minsu Cho, and Jinwoo Shin. Vision-aligned latent reasoning for multi-modal large language model. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=qmvoTiXSfG.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014. IEEE, 2025.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image$f$ a deep-dive into the vision of vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14558. IEEE, 2025.

Jiaxuan Kang, Siyu Chen, Mingda Li, Mingjie Liu, Tianyue Wang, Zhaoyang Wei, Yongheng Zhang, Yanchao Hao, and Zheng Wei. Lut: Latent utility training for visual reasoning. *arXiv preprint arXiv:2608.00743*, 2026.

Benno Krojer, Shravan Nayak, Oscar Mañas, Vaibhav Adlakha, Desmond Elliott, Siva Reddy, and Marius Mosbach. Latentlens: Revealing highly interpretable visual tokens in llms. *arXiv preprint arXiv:2602.00462*, 2026.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 148076–148090, 2026a.

Kelvin Li, Chuyi Shang, Leonid Karlinsky, Rogerio Feris, Trevor Darrell, and Roei Herzig. Latent implicit visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33457–33466, 2026b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

You Li, Chi Chen, Yanghao Li, Fanhu Zeng, Kaiyu Huang, Jinan Xu, and Maosong Sun. Imagination helps visual reasoning, but not yet in latent space. In *Forty-third International Conference on Machine Learning*, 2026c. URL https://openreview.net/forum?id=l1cMErXg1P.

Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5334–5342, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haolang Lu, Bolun Chu, WeiYe Fu, Guoshun Nan, Junning Liu, Minghui Pan, Qiankun Li, Yi Yu, Hua Wang, and Kun Wang. Reallocating attention across layers to reduce multimodal hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4157–4167, 2026.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *Advances in Neural Information Processing Systems*, 37:48224–48255, 2024.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 57172–57189, 2025.

nostalgebraist. Interpreting gpt: the logit lens. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`, 2020.

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.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3505–3506, 2020.

Arijit Ray, Ahmed Abdelkader, Chengzhi Mao, Bryan A Plummer, Kate Saenko, Ranjay Krishna, Leonidas Guibas, and Wen-Sheng Chu. Mull-tokens: Modality-agnostic latent thinking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9477–9488, 2026.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

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

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

André G Viveiros, Nuno Gonçalves, Matthias Lindemann, and André Martins. Lantern: Latent visual structured reasoning. *arXiv preprint arXiv:2603.25629*, 2026.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond image and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12030–12040, 2026a.

Yifan Wang, Shiyu Li, Peiming Li, Xiaochen Yang, Zheng Wei, and Yang Tang. Render-of-thought: Rendering textual chain-of-thought as images for visual latent reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 45236–45253, 2026b.

Yubo Wang, Juntian Zhang, Yichen Wu, Yankai Lin, Nils Lukas, and Yuhan Liu. Forest before trees: Latent superposition for efficient visual reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11272–11288, 2026c.

Ziyang Wang, Mengwei Li, Hao Yin, Wenhao Liu, and Zilei Wang. Posprune: Visual token pruning with positional bias correction for efficient large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10485–10493, 2026d.

Yake Wei, Yuan Wang, Fengyun Rao, Jing Lyu, and Di Hu. Segmentation before answering: Pixel grounding for mllm visual reasoning. *arXiv preprint arXiv:2607.05798*, 2026.

Hongxuan Wu, Yukun Zhang, and Xueqing Zhou. How vision becomes language: A layer-wise information-theoretic analysis of multimodal reasoning. *arXiv preprint arXiv:2602.15580*, 2026a.

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

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094. IEEE, 2024.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the emergence of position bias in transformers. *arXiv preprint arXiv:2502.01951*, 2025.

xAI. Realworldqa: A benchmark for real-world spatial understanding. `https://huggingface.co/datasets/xai-org/RealworldQA`, 2024.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Conical visual concentration for efficient large vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14593–14603. IEEE, 2025.

Chengxu Yang, Siqi Cai, Jingling Yuan, and Chuang Hu. Mitigating hallucination in multimodal large language models via cross-layer visual anchors. *Pattern Recognition*, pp. 114380, 2026a.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33510–33520, 2026b.

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. *arXiv preprint arXiv:2308.02490*, 2023.

Zhuoran Yu and Yong Jae Lee. How multimodal llms solve image tasks: A lens on visual grounding, task reasoning, and answer decoding. In *Second Conference on Language Modeling*, 2025.

Yang Zhang, Danyang Li, Yuxuan Li, Xin Zhang, Tianyu Xie, Mingming Cheng, and Xiang Li. Crystal: Spontaneous emergence of visual latents in mllms. *arXiv preprint arXiv:2602.20980*, 2026.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025a.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19781–19791. IEEE, 2025b.

Appendix

A Experimental Details

We initialize MP-LVR from Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and train it on a 10,116-sample subset of the CoVT training set (Qin et al., 2025). Images are resized to 448×448 pixels, and input sequences are right-padded and truncated to a maximum length of 32,768 tokens. We insert $K = 8$ learnable latent tokens, {<|latent_0|>, ..., <|latent_7|>}, between <think> and </think>. The final answer is enclosed between <answer> and </answer>, which explicitly restrict the answer-generation segment following </think>. Each special token including latent tokens has its own embedding and position-specific Rotary Positional Embedding (RoPE) (Su et al., 2024) index. The new embeddings are initialized using the default vocabulary-resize initialization, and no next-token cross-entropy loss is applied at latent-token positions.

We freeze all pretrained backbone parameters and optimize LoRA (Hu et al., 2021) adapters on all self-attention projections (q_proj, k_proj, v_proj, and o_proj) and MLP projections (gate_proj, up_proj, and down_proj) in every language-model layer, as well as the linear layers of the visual merger. We additionally optimize the input and output embeddings of the newly introduced latent tokens. We use LoRA rank 16, $\alpha = 32$, and dropout 0.01. Training runs for 4,000 optimization steps, corresponding to approximately 3.16 epochs, with a batch size of 8 on a single GPU and no gradient accumulation. We use AdamW with a learning rate of 1×10^{-4}, weight decay 0.1, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and gradient clipping at 1.0. The learning rate is linearly warmed up for the first 5% of training (200 steps) and then decayed using a cosine schedule. Training uses bfloat16 precision, gradient checkpointing, and DeepSpeed ZeRO 2 (Rasley et al., 2020), with TF32 disabled and deterministic training enabled. The main results are reported as the mean and standard deviation over three seeds, $\{0, 42, 50\}$.

For adaptive multi-level pseudo-labeling, we use decoder layers 8–21 for Qwen2.5-VL-7B-Instruct, 9–27 for Qwen2.5-VL-3B-Instruct, and 9–27 for Gemma3-4B-it (Team et al., 2025), out of 28, 36, and 34 decoder layers, respectively. All layer numbers use one-based indexing. For each training sample, we compute clean, head-averaged query-to-image attention over the candidate layers and dynamically select $K = 8$ layers based on query-relevance concentration, as described in Section 3.1. When the number of candidate layers is smaller than K, any remaining latent tokens are assigned to the deepest candidate layer. The fixed target layer is $\ell_t = 21$ for Qwen2.5-VL-7B-Instruct and $\ell_t = 27$ for both Qwen2.5-VL-3B-Instruct and Gemma3-4B-it, where ℓ_t denotes the decoder layer whose output hidden states are used for pseudo-label construction and latent-token supervision. Pseudo-labels are detached from the computation graph, and we set $\lambda_{\text{MP}} = 0.5$.

Stochastic information-path suppression is applied to both answer-to-image and answer-to-query attention for every training example, with perturbations sampled independently for the two paths. We use $\mu \sim \mathcal{U}(-10, 0)$ and $\epsilon \sim \mathcal{N}(0, 1)$ following Section 3.2, where $\mathcal{U}$ and $\mathcal{N}$ denote the uniform and normal distributions, respectively. Perturbations are applied only during the main training forward pass, whereas pseudo-labels are constructed from an unperturbed forward pass. No perturbation is applied during evaluation, and answer-to-latent and latent-to-latent attention remain unchanged.

After training, we merge the LoRA parameters into the backbone and evaluate the resulting model using VLMEvalKit (Duan et al., 2024), following the official evaluation metric for each benchmark. At inference, the model is prefixed with the same eight latent tokens used during training.

All training and evaluation experiments are conducted on a single NVIDIA A100 GPU.

B Additional Experiments

B.1 Hyperparameter Analysis

B.1.1 Effect of Target Layer for Latent-Token Supervision

We analyze the effect of the target layer ℓ_t, whose output hidden states define the representation space for latent supervision. For each setting, adaptive pseudo-labeling selects layers from 8 through

Table A1: **Effect of the target layer ℓ_t for pseudo-label construction and latent supervision on Qwen2.5-VL-7B-Instruct.** For each ℓ_t, adaptive pseudo-labels are constructed from layers 8–ℓ_t.

ℓ_t	BLINK	CVB2D	CVB3D	V*	RQA	Avg.
13	53.29	76.03	83.58	73.82	68.76	71.10
17	54.13	75.47	82.17	72.77	67.97	70.50
21	**54.92**	**76.39**	**84.42**	**74.87**	**69.54**	**72.03**
25	54.23	75.24	78.00	72.77	68.24	69.70
28	52.24	77.31	84.08	73.82	68.50	71.19

Table A2: **Effect of the latent-token supervision weight λ_{MP} on Qwen2.5-VL-7B-Instruct.** Moderate supervision improves overall performance, with the best result achieved at $\lambda_{\text{MP}} = 0.5$.

λ_{MP}	BLINK	CVB2D	CVB3D	V*	RQA	Avg.
0.0	54.08	75.04	82.25	72.77	68.37	70.50
0.1	47.87	75.76	81.92	74.87	63.79	68.84
0.3	54.50	76.52	83.33	73.30	68.10	71.15
0.5	**54.92**	**76.39**	**84.42**	**74.87**	**69.54**	**72.03**
0.7	54.81	76.21	84.17	74.35	69.02	71.71
1.0	54.13	75.86	83.00	74.87	68.63	71.30

Table A3: **Effect of different latent supervision losses**, including mean squared error (MSE), L1, and cosine loss, on Qwen2.5-VL-7B-Instruct. Cosine loss achieves the best overall performance, indicating that directional feature alignment is more effective than matching exact feature values.

Loss Type	BLINK	CVB2D	CVB3D	V*	RQA	Avg.
MSE	47.03	58.76	68.17	64.92	64.97	60.77
L1	52.87	75.29	82.33	73.30	68.50	70.46
Cosine	**54.92**	**76.39**	**84.42**	**74.87**	**69.54**	**72.03**

ℓ_t, while both pseudo-labels and latent-token representations are constructed and supervised at ℓ_t. Table A1 reports performance on five visual reasoning benchmarks using Qwen2.5-VL-7B-Instruct. Performance varies substantially with the target layer and is highest at $\ell_t = 21$, approximately three-quarters of the 28-layer backbone. Earlier target layers may not yet encode sufficiently developed visual information, whereas later layers are increasingly oriented toward answer decoding. Based on this observation, we treat the layer at approximately three-quarters of the network depth as a suitable supervision space and adopt the corresponding target layer for the other backbones.

B.1.2 Effect of Latent-Token Supervision Weight

We analyze the effect of the latent-token supervision weight λ_{MP} on Qwen2.5-VL-7B-Instruct. As shown in Table A2, performance is sensitive to the strength of latent-token supervision. Compared with no latent-token supervision ($\lambda_{\text{MP}} = 0$), a weak supervision weight ($\lambda_{\text{MP}} = 0.1$) degrades overall performance, whereas moderate weights generally improve it. The best performance is achieved at $\lambda_{\text{MP}} = 0.5$, indicating that sufficient latent-token supervision is beneficial for learning effective intermediate representations. Increasing the weight further provides no additional benefit and slightly reduces performance, suggesting that overly strong latent-token supervision may interfere with the answer-prediction objective. These results highlight the importance of balancing latent-token supervision with the task objective.

B.1.3 Effect of Loss Functions for Latent-Token Supervision

We analyze the choice of loss function for latent supervision on Qwen2.5-VL-7B-Instruct, comparing mean squared error (MSE), L1 loss, and cosine losses. As shown in Table A3, cosine loss achieves the best performance across all evaluated benchmarks, while MSE performs substantially

Table A4: **Effect of the number of directly supervised latent tokens on Qwen2.5-VL-7B-Instruct with a fixed total of $K = 8$ latent tokens.** When m tokens are supervised, the first m receive supervision while the remaining $8 - m$ are trained only through answer prediction loss.

m	BLINK	CVB2D	CVB3D	V*	RQA	Avg.
0	54.08	75.04	82.25	72.77	68.37	70.50
2	54.66	76.17	82.42	73.30	68.10	70.93
4	52.18	76.31	82.17	73.82	67.97	70.49
6	54.34	76.42	83.25	74.87	67.71	71.32
8	**54.92**	**76.39**	**84.42**	**74.87**	**69.54**	**72.03**

Table A5: **Effect of the number of latent tokens.** Increasing the latent sequence length does not consistently improve visual reasoning performance, with $K = 8$ achieving the best overall result.

K	BLINK	CVB2D	CVB3D	V*	RQA	Avg.
0 (SFT)	54.66	75.50	82.42	71.73	69.28	70.72
4	54.02	76.20	83.92	73.82	67.84	71.16
8	**54.92**	**76.39**	**84.42**	**74.87**	**69.54**	**72.03**
16	53.55	76.60	85.58	70.68	68.76	71.03

worse and L1 yields intermediate results. Unlike MSE, which penalizes differences in both feature magnitude and direction, cosine loss focuses on directional alignment and is invariant to feature scale. This suggests that aligning the direction of pseudo-label representations is more effective than matching their exact feature values, making cosine loss better suited for latent-token supervision.

B.1.4 Effect of the Number of Supervised Latent Tokens

We investigate how many latent tokens should receive direct supervision on Qwen2.5-VL-7B-Instruct. Motivated by RIS (Cui et al., 2026), which introduces transition tokens learned only through the answer-prediction objective, we vary the number of directly supervised latent tokens while keeping the total number fixed at $K = 8$. When m latent tokens are supervised, the first m tokens receive the latent-token supervision loss, while the remaining $8 - m$ tokens are trained only through answer prediction loss.

As shown in Table A4, partially supervising latent tokens provides limited gains. Performance generally improves as supervision is extended to more latent tokens, with the best overall result obtained when all eight latent tokens are directly supervised. This suggests that, in MP-LVR, leaving some latent tokens to be learned only through answer prediction loss is less effective than providing explicit supervision throughout the latent-token sequence.

B.1.5 Effect of the Number of Latent Tokens

We investigate the effect of the total number of latent tokens on Qwen2.5-VL-7B-Instruct. Unlike the previous Section B.1.4, which varies how many tokens receive supervision while fixing $K = 8$, here we vary the number of latent tokens used for latent visual reasoning. For example, when $K = 4$, only <|latent_0|>,...,<|latent_3|> are inserted and trained, while $K = 0$ corresponds to standard Supervised Fine-Tuning (SFT) without latent tokens.

As shown in Table A5, increasing the number of latent tokens does not monotonically improve performance. Using four latent tokens improves over SFT, while eight latent tokens achieves the best overall performance on five visual reasoning benchmarks. Increasing the number further to 16 slightly reduces average performance despite gains on some benchmarks. These results suggest that selecting an appropriate number of latent tokens is important, as simply increasing the latent-token sequence length does not necessarily improve visual reasoning performance.

Table A6: **Comparison of Shared and Distinct latent tokens on Qwen2.5-VL-7B-Instruct.** Using distinct learnable embeddings for latent tokens consistently improves visual reasoning performance.

Latent Tokens	BLINK	CVB2D	CVB3D	V*	RQA	Avg.
Shared	52.76	75.65	82.92	73.82	68.24	70.68
Distinct	**54.92**	**76.39**	**84.42**	**74.87**	**69.54**	**72.03**

Table A7: **Open-ended VQA performance on Qwen2.5-VL-7B-Instruct.** Among methods trained under the same controlled setting, MP-LVR achieves the strongest performance across benchmarks.

Method	MMVet	LLaVABench	WildVision	Avg.
Qwen2.5-VL-7B	67.3	90.2	63.0	73.5
CoVT	60.7	82.5	53.4	65.5
LVR	42.9	41.9	13.6	32.8
Vision-R1	58.6	56.7	20.0	45.1
SKILA	54.9	62.1	19.0	45.3
SFT	56.8	70.8	30.4	52.7
LIVR	52.6	69.5	33.2	53.4
CrystaL	52.6	55.4	16.6	41.5
MP-LVR	**62.3**	**71.4**	**44.4**	**59.4**

B.1.6 Effect of Latent Token Diversity

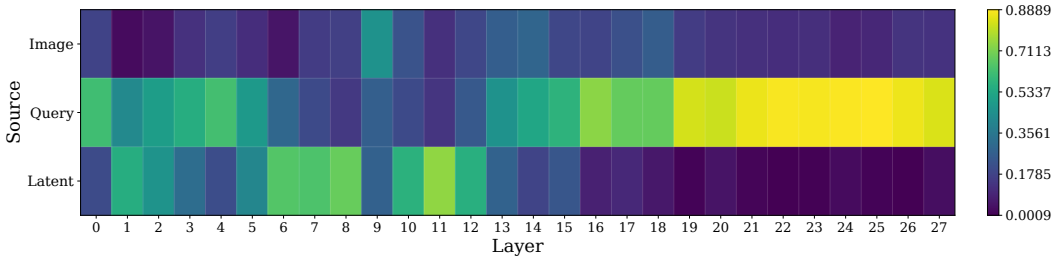
We compare Shared and Distinct latent-token parameterizations on Qwen2.5-VL-7B-Instruct. In the shared setting, all eight latent-token positions use the same learnable token `<|latent|>` and share a single embedding parameter. In the distinct setting, we use eight separate latent tokens, {`<|latent_0|>`, ..., `<|latent_7|>`}, whose embeddings are independently optimized under the same initialization scheme.

As shown in Table A6, using distinct latent tokens consistently improves performance across all evaluated benchmarks. This suggests that sharing a single embedding across all latent-token positions limits the flexibility of latent-token representation learning. In contrast, assigning each latent token an independently optimized embedding allows different latent-token positions to develop differentiated representations, leading to better overall visual reasoning performance.

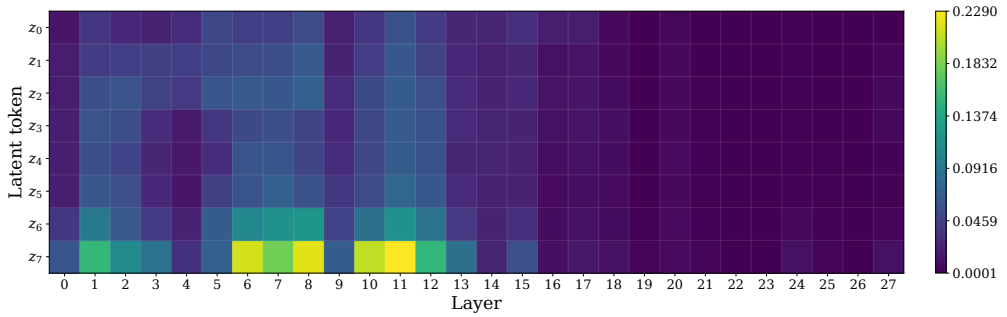
B.2 Open-Ended Visual Question Answering

We examine whether Latent Visual Reasoning (LVR) methods preserve the model's open-ended generation capability, as most LVR baselines and MP-LVR are trained primarily on short-answer examples. We evaluate Qwen2.5-VL-7B-Instruct and the compared methods on representative open-ended Visual Question Answering (VQA) benchmarks, including MMVet (Yu et al., 2023), LLaVABench (Liu et al., 2023), and WildVision (Lu et al., 2024). Following the main experiments, we evaluate the official pretrained checkpoints of CoVT (Qin et al., 2025), LVR (Li et al., 2026a), Vision-R1 (Huang et al., 2026), and SKILA (Tong et al., 2025), while SFT, LIVR (Li et al., 2026b), CrystaL (Zhang et al., 2026), and MP-LVR are trained under the same controlled setting using the same training data and configuration. For evaluation, we use the official GPT-4-based evaluator (Achiam et al., 2023) for MMVet, while LLaVABench and WildVision are evaluated using GPT-4o (Hurst et al., 2024).

As shown in Table A7, LVR methods generally exhibit degraded open-ended VQA performance relative to the base model, suggesting that training predominantly on short-answer data may weaken the original open-ended generation capability. Nevertheless, among the methods trained under the same controlled setting, MP-LVR achieves the strongest performance across all three benchmarks, outperforming SFT, LIVR, and CrystaL. These results suggest that although short-answer fine-tuning introduces a trade-off in open-ended generation, MP-LVR better preserves this capability than the other methods under the same training setting.

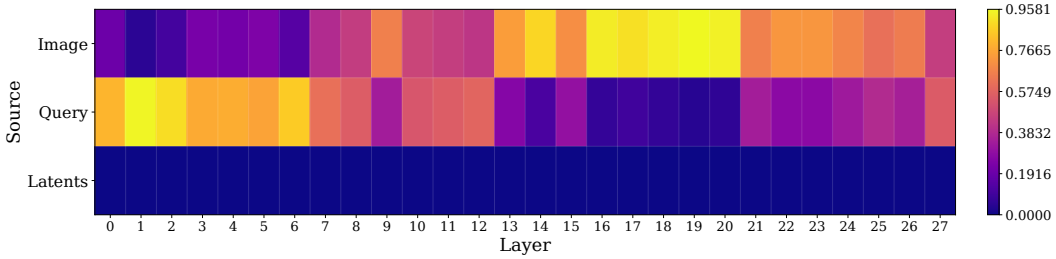


(a) Layer-wise attention from the <answer> token to image, query, and latent tokens.



(b) Layer-wise attention from the <answer> token to individual latent tokens.

Figure A1: **Attention patterns of the <answer> token across layers over 100 CV-Bench-2D samples.** The model attends to latent tokens in the early-to-middle layers and shifts toward query tokens in later layers, while later latent tokens receive stronger attention than earlier ones.



(a) Layer-wise attention from the <|latent_0|> token to image, and query tokens.

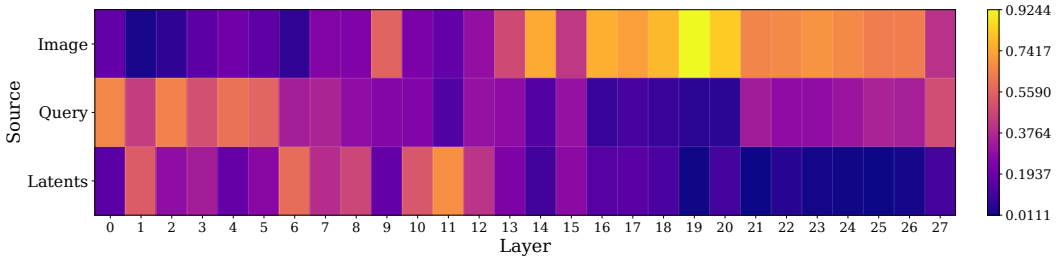
C Additional Analysis

C.1 Attention Visualization

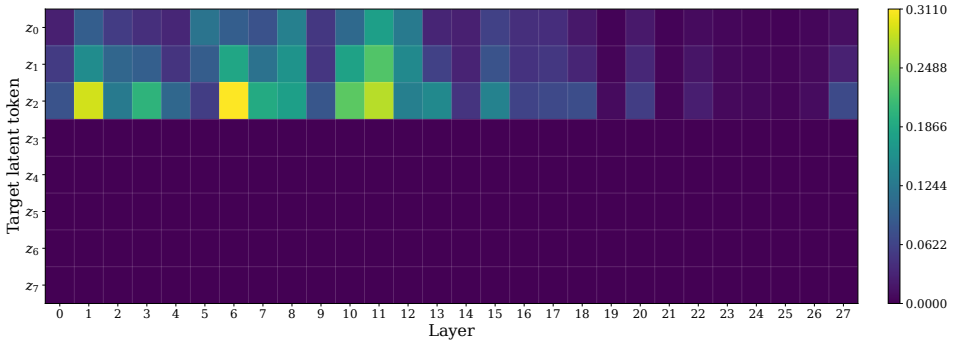
C.1.1 Answer Attention to Image, Query, and Latent Tokens

We analyze how information from different input sources is utilized immediately before answer generation. Specifically, we measure the attention of the <answer> token, whose output predicts the first answer token, to image, query, and latent tokens at each layer. For each source, we sum the attention weights over all corresponding tokens. The visualization is obtained by aggregating attention patterns over 100 samples from CV-Bench-2D.

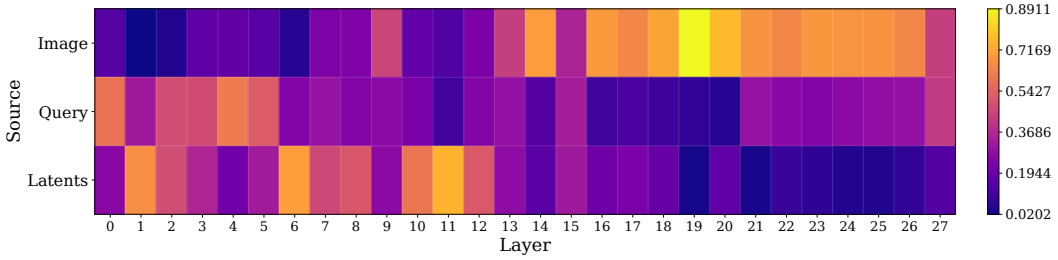
As shown in Figure A1a, the <answer> token attends more strongly to latent tokens than to image tokens across much of the early and middle layers. In the later layers, however, attention to query tokens becomes increasingly dominant, revealing a clear transition from latent-focused intermediate processing to query-focused textual answer preparation. Direct attention to image tokens remains present across layers, but is generally weaker than latent-token attention in the early-to-middle layers and query attention in the later layers.



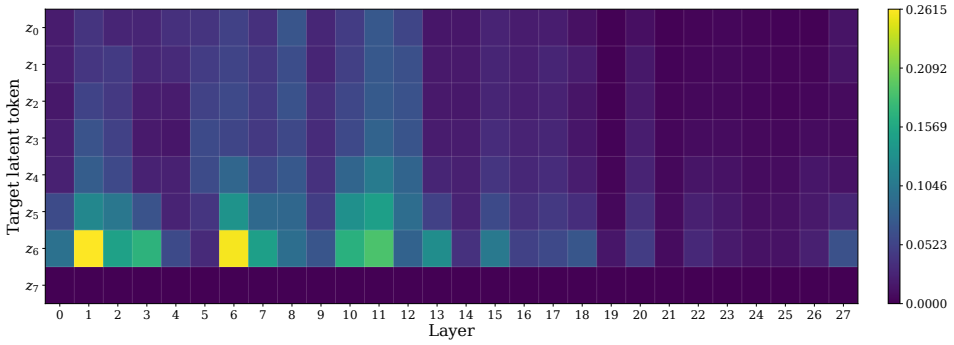
(b) Layer-wise attention from the <|latent_3|> token to image, query, and latent tokens.



(c) Layer-wise attention from the <|latent_3|> token to individual latent tokens.



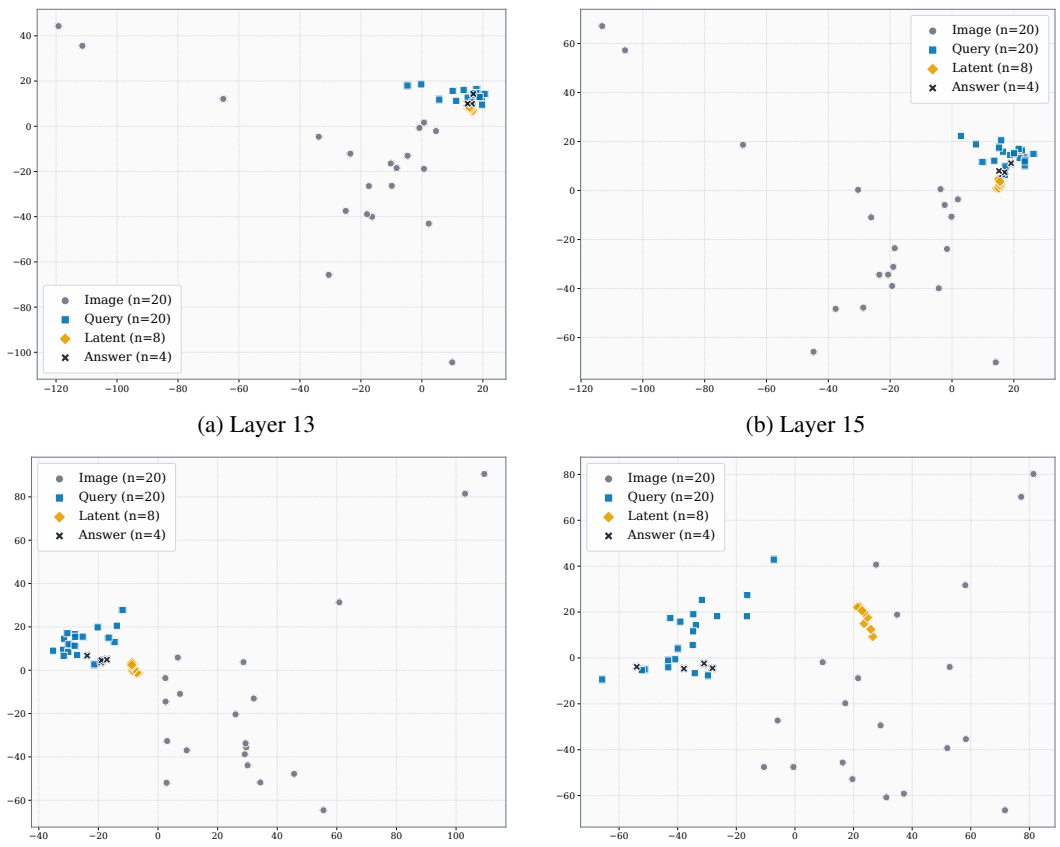
(d) Layer-wise attention from the <|latent_7|> token to image, query, and latent tokens.



(e) Layer-wise attention from the <|latent_7|> token to individual latent tokens.

Figure A2: **Attention patterns of three latent tokens across layers over 100 CV-Bench-2D samples.** We visualize attention from latent tokens to image, query, and preceding latent tokens. Later latent tokens show a strong tendency to attend to the immediately preceding latent token.

Figure A1b further decomposes attention across individual latent tokens. Later latent tokens, which are closer to the <answer> token in the sequence, receive substantially stronger attention, with <|latent_7|> (z_7) showing the most pronounced attention and <|latent_6|> (z_6) also receiving relatively high attention. This pattern is consistent with the intervention results in Table 4,



(a) Layer 13 (b) Layer 15

(c) Layer 17 (d) Layer 19

Figure A3: **PCA visualization of image, query, latent, and answer-token representations across layers.** Latent tokens gradually shift from query representations toward an intermediate region between query and image representations, and move closer to image representations in later layers.

where later latent tokens preserve more answer-relevant information. Together, these results provide further evidence that the learned latent tokens are actively utilized during answer generation.

C.1.2 Latent-to-Image, Query, and Latent Attention

We analyze how latent tokens integrate information from image, query, and preceding latent tokens across layers. To visualize this behavior compactly, we select three representative latent tokens, <|latent_0|>, <|latent_3|>, and <|latent_7|>, corresponding to early, middle, and late positions in the latent-token sequence. For each latent token, we sum the attention weights over image, query, and preceding latent tokens, respectively, excluding attention to itself to isolate interactions with preceding latent tokens.

As shown in Figure A2, <|latent_0|> can attend only to image and query tokens due to causal attention. In the early layers, it attends predominantly to the query, while attention becomes more balanced between image and query tokens in the early-to-middle layers.

The later latent tokens, <|latent_3|> and <|latent_7|>, can additionally access preceding latent tokens. In the early-to-middle layers, they attend substantially to both query and preceding latent tokens, while this latent-to-latent attention diminishes in deeper layers. The per-latent attention maps further show a strong local sequential pattern: <|latent_3|> attends most strongly to <|latent_2|>, while <|latent_7|> primarily attends to <|latent_6|>. This tendency is particularly pronounced for later latent tokens.

These observations suggest that information propagation across latent tokens occurs primarily in the early-to-middle layers, where latent tokens integrate preceding latent tokens with the original image and query inputs to form an intermediate pathway for progressive information refinement.

C.1.3 PCA Analysis of Latent-Token Representations

We examine how latent-token representations evolve across layers using Principal Component Analysis (PCA). For a CV-Bench-2D sample, we jointly project 20 sampled image tokens, 20 sampled query tokens, all eight latent tokens, and four answer tokens into a two-dimensional space at different layers.

As shown in Figure A3, the relative positions of latent tokens in the PCA space change noticeably with layer depth. In the earlier layers, latent-token representations are located close to the query token representations. As layer depth increases, the latent tokens gradually shift toward an intermediate region between the query and image tokens, before moving closer to the image token representations in later layers. This transition suggests that latent tokens progressively integrate information from both modalities. Notably, MP-LVR's adaptive layer selection exploits middle-layer representations ranging from those closer to query representations, through intermediate states, to those closer to image representations, providing diverse supervision for latent tokens.

D Limitation

Our current framework uses a fixed number of latent tokens for all inputs, without adapting the latent reasoning budget to the complexity of each sample. Dynamically determining the number of latent tokens is an interesting direction for future work.

Although our analyses show that the learned latent tokens are functionally utilized and encode distinct visual representations, their precise lexical semantics remain difficult to interpret. Preliminary analyses using vocabulary-projection-based methods, such as Logit Lens (nostalgebraist, 2020) and Latent Lens (Krojer et al., 2026), did not reveal clear or stable lexical interpretations for individual latent tokens. Developing interpretation methods better suited to continuous latent reasoning representations remains an important direction for future work.