FUSED MULTI-LATENT REASONING: UNIFIED MULTIMODAL REASONING IN LATENT SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language reasoning often unfolds through intermediate steps that mix textual reasoning, visual evidence, and cross-modal grounding. Existing latent reasoning methods attempt to move such processes into latent space, but they typically rely on unimodal or weakly coupled supervision, leaving latent units poorly aligned with the multimodal reasoning steps they replace and prone to losing task-relevant visual evidence or semantic intent during compression. Our key insight is that high-dimensional vision-language embeddings can compress interleaved textual-visual reasoning sequences into dense representations that, after adaptation, serve as informative latent units for multimodal reasoning. In this work, we propose **Fused Multi-Latent Reasoning (FMLR)**, a latent-space reasoning framework that learns latent units supervised by fused textual-visual step representations. By distilling step-level multimodal embeddings into latent units, FMLR enables the model to reason over implicit multimodal states. To further improve latent reasoning and control its length, we introduce **Minimum-Sufficient Latent GRPO (MSL-GRPO)**, which combines answer correctness, threshold-based early sufficiency, and a per-unit latent cost to learn concise reasoning and explicit stopping decisions. Experiments on challenging vision-language reasoning benchmarks show that FMLR improves accuracy while producing substantially fewer output tokens than explicit-reasoning baselines. Code and datasets will be released upon acceptance.

1 INTRODUCTION

Multimodal large language models (MLLMs) Alayrac et al. (2022); Li et al. (2023); Zhu et al. (2023); Liu et al. (2023; 2024); Peng et al. (2023); Bai et al. (2023); Team et al. (2023); Chen et al. (2024b;a); Xu et al. (2025a); Team et al. (2025); Yan et al. (2026) have recently advanced rapidly, significantly improving the integration of visual perception and language understanding. These models have demonstrated strong performance on a wide range of tasks, including visual question answering, image captioning, and multimodal reasoning. However, as tasks become more complex, requiring fine-grained perception and multi-step reasoning, existing MLLMs still struggle to reliably ground intermediate reasoning in visual evidence, leading to unfaithful or inconsistent predictions.

Recent "thinking with images" Zheng et al. (2025); Fan et al. (2025); Xu et al. (2025b); Su et al. (2025a;b) approaches enable MLLMs to perform multi-step reasoning by explicitly interleaving textual thoughts with visual grounding operations. As illustrated in Fig. 1(a), this paradigm provides interpretable reasoning traces, but it also introduces long textual chains, repeated cross-modal interactions, and substantial inference overhead. Latent reasoning methods Li et al. (2025b); Wang et al. (2025a); Li et al. (2026c); Zhang et al. (2026) reduce this cost by moving intermediate reasoning into hidden space, allowing models to reason with latent states instead of fully verbalized token sequences, as shown in Fig. 1(b). Despite this progress, existing latent reasoning methods still leave open how each latent unit should be formed. Most approaches obtain latent states through visual alignment, intermediate distillation, or sequential latent generation, which improves efficiency but does not directly define what reasoning content a latent unit should preserve. In multimodal reasoning, each intermediate step usually contains semantic intent, relevant visual evidence, and their cross-modal relation. A high-dimensional vision-language embedding provides a natural carrier for these components, since it can condense an interleaved textual-visual sequence into a single dense representation. Recent studies suggest that latent states can carry more reasoning information than explicit token-level traces Tan et al. (2026), further supporting the use of such

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.

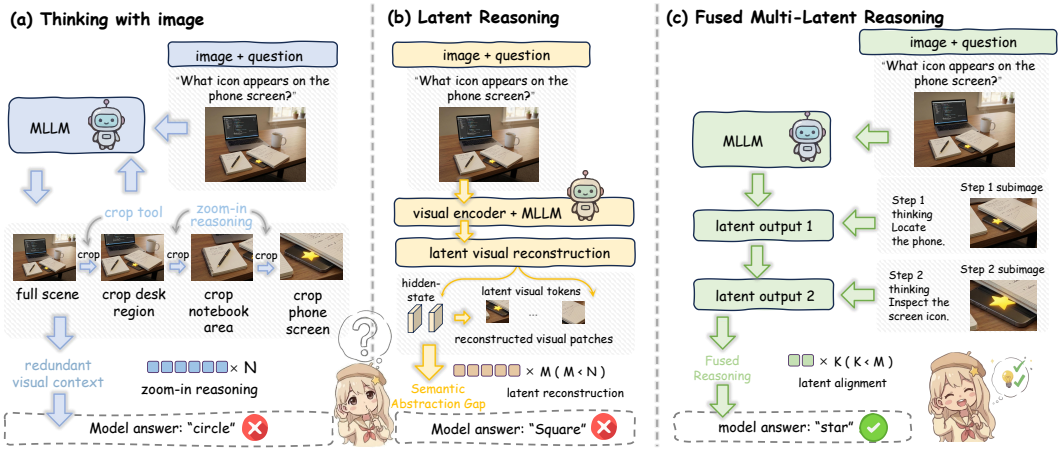


Figure 1: Comparison of visual reasoning paradigms. (a) "Thinking with images" on repeated visual-text interactions over cropped regions. (b) Latent Reasoning reconstructs multi-token visual or textual features. (c) Fused Multi-Latent Reasoning fuses multimodal reasoning into compact latent embeddings and aligns them with model hidden states for unified reasoning.

information-rich representations as reasoning units. This motivates a fused latent reasoning paradigm, where condensed step-level vision-language embeddings are adapted into latent units for subsequent multimodal reasoning.

Building on this motivation, we propose **Fused Multi-Latent Reasoning (FMLR)**, which uses fused vision-language representations to supervise latent reasoning slots (Fig. 1(c)). Stage 1 aligns hidden states with targets derived from paired reasoning text and visual evidence. Stage 2, **Minimum-Sufficient Latent GRPO (MSL-GRPO)**, optimizes a stochastic policy over full latent vectors, stop/continue decisions, and answer tokens. Its reward combines answer correctness, early attainment of a sufficiency threshold, and a cost for each generated latent action. This formulation encourages confident intermediate states and concise trajectories while explicitly optimizing stopping decisions.

Our contributions are summarized as follows:

- We propose **Fused Multi-Latent Reasoning (FMLR)**, a latent-space reasoning framework that converts fused vision-language step representations into latent reasoning units carrying semantic intent, visual evidence, and cross-modal grounding.
- We introduce **Minimum-Sufficient Latent GRPO (MSL-GRPO)** to jointly optimize answer correctness, early sufficiency, and latent reasoning length.
- We construct two datasets, **FMLR-400K-SFT** for multimodal reasoning traces and **FMLR-510K-EMB** for fused embedding adaptation, supporting the learning of vision-language latent reasoning units.

2 Related Work

Thinking with Images. A major line of multimodal reasoning makes visual evidence an explicit part of the reasoning process by introducing intermediate images, visual operations, or image-text interleaving during inference. COGCOM Qi et al. (2025), V* Wu & Xie (2024), and VisionReasoner Liu et al. (2026) treat reasoning as a sequence of grounded visual actions, such as cropping, searching, OCR, or region-level inspection, so that the model can progressively acquire missing evidence instead of relying on a single static view. GRIT Fan et al. (2025), DeepEyes Zheng et al. (2025), MLLMs Know Where to Look Zhang et al. (2025b), PixelReasoner Su et al. (2025a) and Chain-of-Focus Zhang et al. (2025c) further improve this paradigm with adaptive region selection or zoom-in policies, enabling the model to refine attention over small or ambiguous visual details. Visual-RFT Liu et al. (2025) and related reinforcement-based approaches improve the faithfulness of these explicit reasoning traces by encouraging closer alignment between intermediate reasoning

steps and visual evidence. Beyond direct image manipulation, StruVis Lyu et al. (2026) uses structured visual states as lightweight intermediate representations, while ThinkMorph Gu et al. (2026) studies interleaved image-text reasoning and shows that explicitly alternating modalities can provide complementary reasoning steps. Although these methods offer strong grounding and interpretability, they usually depend on repeated visual operations or explicit multimodal traces at test time, which increases inference cost and makes reasoning tightly coupled to the design of external or visible intermediate states.

Implicit Latent-Space Reasoning. Recent work suggests that reasoning is more fundamentally driven by latent-state trajectories than by surface chain-of-thought alone Wang (2026). In contrast to explicit visual reasoning, latent reasoning methods shift part of the reasoning process into hidden representations, aiming to improve efficiency while avoiding repeated explicit visual interaction. In LLMs, COCONUT Hao et al. (2025), SIM-CoT Wei et al. (2025), and related latent CoT methods Deng et al. (2026); Tan et al. (2026) show that intermediate reasoning can be compressed into continuous hidden states, while adaptive variants such as System-1.5 Reasoning Wang et al. (2025c), Hybrid Latent Reasoning Yue et al. (2025), and Inner Thinking Transformer Chen et al. (2025) explore when and how latent computation should be combined with token-space reasoning. Extending this idea to multimodal settings, LVR Li et al. (2025b) introduces latent visual reasoning by reconstructing task-relevant visual semantics from hidden states and integrating them into autoregressive inference. Follow-up work enriches this latent interface from different angles: Monet Wang et al. (2025a) uses staged supervision and latent-level reinforcement learning, Machine Mental Imagery Yang et al. (2025) encourages internal visual imagination, Latent Sketchpad Zhang et al. (2025a) provides a sketch-like latent workspace, and Reasoning Within the Mind Liu et al. (2026) studies dynamic multimodal interleaving inside latent space. Mull-Tokens Ray et al. (2025) proposes modality-agnostic latent thinking tokens, while CoVT Qin et al. (2025) preserves dense perceptual cues such as depth, edges, and segmentation in continuous visual thought tokens. RoT Wang et al. (2026) and OneLatent Lv et al. (2026) show that latent reasoning can also be supervised or compressed through more structured bottlenecks. At the same time, CapImagine Li et al. (2026c) questions whether current latent visual tokens are always causally effective, suggesting that the key challenge is not only to hide reasoning in latent space, but to make latent representations semantically informative, stable, and reusable across different reasoning queries.

3 METHOD

3.1 OVERVIEW

Figure 2 shows the construction of the multimodal supervision: interleaved traces are generated, checked, and refined. Stage 1 aligns latent hidden states with fused targets derived from this supervision. Stage 2 reuses that checkpoint and optimizes latent actions and stopping with MSL-GRPO on the ZebraCoT portion, using answer correctness, threshold-based early sufficiency, and actual latent count. Figure 3 summarizes the training procedure.

3.2 DATA CONSTRUCTION

The data construction pipeline constructs a reasoning-trace resource and a fused-embedding resource. These support the reused Stage-1 checkpoint; Stage 2 draws its prompts from the ZebraCoT portion of the same fused-embedding resource. For SFT we start from Visual CoT Shao et al. (2024), where each raw sample is $x = (I, q, a)$ with image I, question q, and answer a. As shown in Fig. 2, the SFT-stage construction has three steps: CoT generation, verification, and filtering with refinement.

Step 1: CoT Generation. Given (I_i, q_i), we prompt an MLLM to generate a multimodal reasoning chain $T_i = \{s_{i,1}, \ldots, s_{i,M_i}\}$, where each reference step $s_{i,j}$ pairs a bounding box with a reasoning statement. Here M_i counts reference reasoning steps, not generated latent actions. The output is a reasoning trace and a final answer.

Step 2: Verification. The generated CoT may contain noisy box references, weak steps, or answer–reasoning mismatches, so we run a bidirectional verification. We first crop every box from the image, which gives step-wise grounded visual evidence. *Forward verification* asks the MLLM to predict

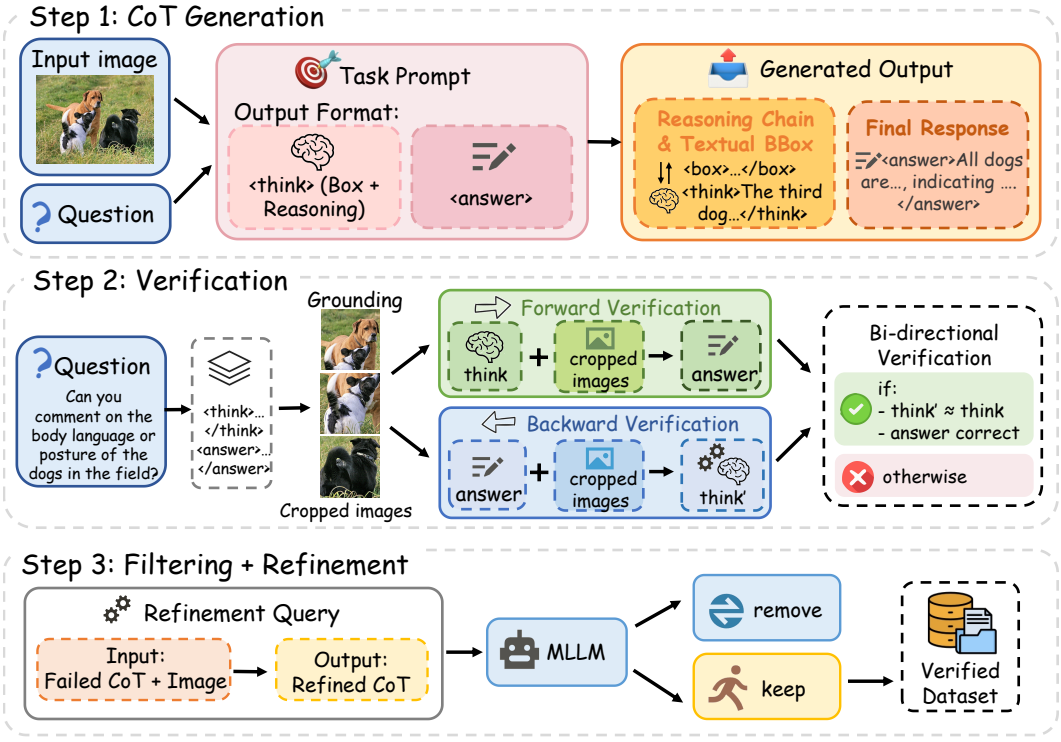


Figure 2: Overview of the data construction pipeline, which consists of three steps: multimodal CoT generation, verification, and sample filtering with reasoning refinement.

the answer from the question, the reasoning chain, and the crops, testing whether they support the answer. *Backward verification* asks it to reconstruct the chain from the question, the answer, and the crops, testing whether the trace is recoverable. We keep a sample only if an LLM judge finds both the forward answer correct and the reconstructed trace consistent with the original. All verification uses Qwen3-VL-235B Bai et al. (2025a); prompts are in Appendix C.

Step 3: Filtering and Refinement. Failed samples are refined rather than discarded: we regenerate their boxes and reasoning steps with the same MLLM and re-verify. We refine about 71,635 problematic samples this way, yielding the verified SFT dataset **FMLR-400K-SFT**, which keeps for each sample the image, question, answer, reasoning chain, and its crops. Examples are in Appendix D.

Embedding Dataset Construction. For the fused embedding dataset we use verified Visual CoT traces plus 117,450 ZebraCoT Li et al. (2025a) samples (after removing 2D-only data). Each reference step's text and paired visual evidence are jointly encoded by Qwen3-VL-Embedding-8B Li et al. (2026a), yielding step representations $\{e_{i,j}^{\text{step}}\}_{j=1}^{M_i}$. Each example supplies an ordered list of latent targets $E_i = \{e_{i,n}\}_{n=1}^{N_i}$ and matching placeholder positions. N_i counts supervised latent slots, while M_i counts reference steps. The resulting **FMLR-510K-EMB** provides fused targets for Stage-1 alignment.

3.3 Stage 1: Fused-Latent Alignment

We reuse the Qwen3-VL-8B Stage-1 checkpoint trained on the Visual CoT portion of FMLR-510K-EMB.

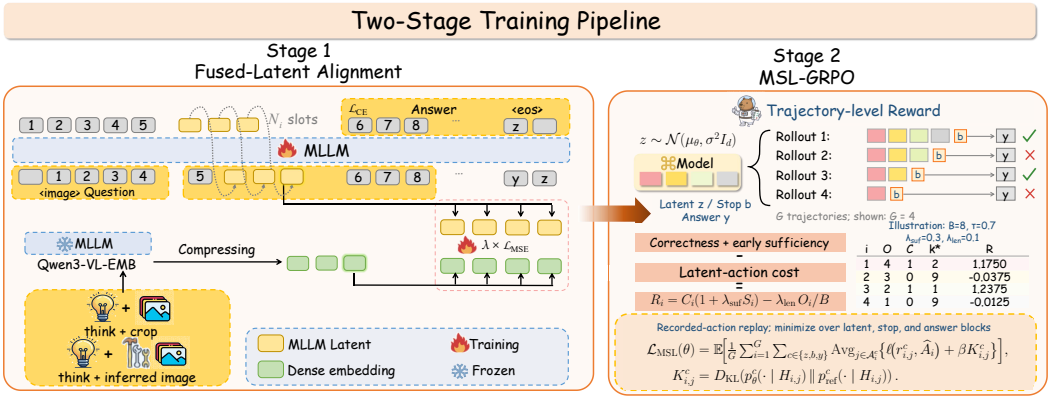


Figure 3: Overview of the Fused Multi-Latent Reasoning training pipeline.

For example i, let N_i count its latent targets, $h_{i,n}$ be the hidden state preceding the n-th latent placeholder, and $e_{i,n}$ be the paired fused textual-visual target. Training combines answer-token cross-entropy with latent alignment:

$$\mathcal{L}_{\text{align},i} = \frac{1}{N_i d} \sum_{n=1}^{N_i} \|h_{i,n} - e_{i,n}\|_2^2, \qquad \mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{ans}} + 0.1\, \mathbb{E}_i[\mathcal{L}_{\text{align},i}]\,. \tag{1}$$

In Eq. (1), $\mathcal{L}_{\text{ans}}$ is cross-entropy on answer tokens, averaged over the same batch as the alignment term. Intermediate reasoning text and latent placeholders are excluded from answer supervision. The hidden dimension is $d = 4096$.

The alignment and answer objectives serve complementary purposes. The fused target provides a step-level training signal that combines the reasoning statement with the visual evidence it refers to. Answer supervision links these intermediate states to the final task. This separates the choice of a compact representation from the question of how much latent computation an input needs: Stage 1 supplies the initial representation, while Stage 2 learns to use and terminate the resulting latent trajectory.

3.4 Stage 2: Minimum-Sufficient Latent GRPO

MSL-GRPO starts from the frozen Stage-1 weights and trains language-backbone LoRA adapters. It runs on the ZebraCoT portion of FMLR-510K-EMB. Given an image-question pair x, the old policy samples G trajectories, each containing full latent actions $z_i^{(1:O_i)}$, non-forced binary stopping decisions b_i, and answer tokens $y_i^{(1:T_i^y)}$. The model receives the image together with its question or options.

MSL-GRPO rewards correct answers, identifies early sufficient prefixes with a confidence threshold τ, and charges for every generated latent action.

Continuous actions with a common support. Let $\mathcal{H}_{i,k}$ contain the input and all previously recorded actions, and let $\mu_\theta(\mathcal{H}_{i,k}) \in \mathbb{R}^d$ be the last language-model hidden state before latent action k. We use a fixed noise scale $\sigma = 0.5$ and sample a full-dimensional action:

$$\begin{aligned} z_i^{(k)} &= \mu_\theta(\mathcal{H}_{i,k}) + \sigma \varepsilon_{i,k}, \qquad \varepsilon_{i,k} \sim \mathcal{N}(0, I_d), \\ \log p_\theta^z(z \mid \mathcal{H}) &= -\frac{\|z - \mu_\theta(\mathcal{H})\|_2^2}{2\sigma^2} - \frac{d}{2}\log(2\pi\sigma^2). \end{aligned} \tag{2}$$

Eq. (2) gives every policy the common support $\mathbb{R}^d$ at a fixed history. The sampled latent vectors are retained for likelihood evaluation and replay.

Stopping and answer policies. After the minimum of one latent action, the language-model logits for the existing latent and latent-end tokens define

$$q_\theta(\mathcal{H}) = \text{softmax}\left([u_\theta^{\text{latent}}(\mathcal{H}), u_\theta^{\text{end}}(\mathcal{H})]\right), \qquad b \sim \text{Categorical}(q_\theta), \tag{3}$$

where $b = 0$ continues and $b = 1$ stops. Eq. (3) uses the existing LM head. We require at least one latent action and stop at budget $B = 8$. After exit, a shared latent-end boundary and answer prefix condition answer generation. Tokens are sampled from the full vocabulary at temperature 1, until EOS or 64 tokens. Policy probabilities and KL terms cover sampled actions, excluding forced boundaries and the fixed answer prefix.

Correctness, sufficiency, and latent cost. A deterministic scorer returns $C_i \in \{0, 1\}$ by comparing the parsed final answer with the reference choice. An unambiguous correct label or uniquely matching option text receives one; malformed or conflicting answers receive zero. For each latent prefix, an independent answer branch of the old policy scores the canonical choice labels after the shared latent-end boundary and answer prefix. Let $p_i^{(k)}$ be the correct label's probability normalized over these choices. With a fixed budget B,

$$k_i^* = \min\Big(\{1 \le k \le O_i : p_i^{(k)} \ge \tau\} \cup \{B + 1\}\Big), \qquad S_i = \max\left(1 - \frac{k_i^*}{B}, 0\right). \tag{4}$$

Here k_i^* is the first threshold crossing, with $B+1$ denoting no crossing and $S_i = 0$. The probe scores in Eq. (4) and resulting rewards are fixed during each policy update. We combine early sufficiency with correctness and generated latent count:

$$R_i = C_i\big(1 + \lambda_{\text{suf}} S_i\big) - \lambda_{\text{len}} \frac{O_i}{B}. \tag{5}$$

We use $B = 8$, $\tau = 0.7$, $\lambda_{\text{suf}} = 0.3$, and $\lambda_{\text{len}} = 0.1$. In Eq. (5), O_i counts the latent actions actually generated. Correct trajectories receive a larger bonus when they reach the threshold earlier. Once k_i^* is fixed, appending m steps at unchanged correctness reduces reward by $\lambda_{\text{len}} m/B$. Correct trajectories score at least 0.9, whereas incorrect trajectories receive only the length penalty.

For example, a correct trajectory with $k_i^* = 2$ receives reward 1.200 when it stops after two latent actions and 1.175 when it continues to four. Thus, the stopping policy can benefit from retaining a sufficient prefix while avoiding additional latent actions that leave the answer unchanged.

Replay and clipped optimization. We record the sampled actions, old log-probabilities, and rewards. During optimization, the current and reference policies evaluate the same recorded actions and histories. Let $\mathcal{A}_i^z$, $\mathcal{A}_i^b$, and $\mathcal{A}_i^y$ index the latent, sampled stopping, and answer actions. For action j in block $c \in \{z, b, y\}$,

$$r_{i,j}^c = \frac{p_\theta^c(a_{i,j}^c \mid \mathcal{H}_{i,j})}{p_{\theta_{\text{old}}}^c(a_{i,j}^c \mid \mathcal{H}_{i,j})}, \qquad \widehat{A}_i = \frac{R_i - \overline{R}}{\max(\text{std}_{\text{pop}}(R_1, \ldots, R_G), 10^{-6})}. \tag{6}$$

The group advantage $\widehat{A}_i$ of Eq. (6) is shared by all three action blocks; all-equal rewards yield zero policy advantage. We sample $G = 4$ trajectories per prompt, allowing three or more reward levels to carry graded advantages. The reward weights set component trade-offs; group normalization remains invariant to positive affine reward transformations whenever its numerical floor is inactive. Define $\text{Avg}_{j \in \mathcal{A}}$ as the arithmetic mean for a nonempty block and zero for an empty block. The joint trajectory density factorizes over the recorded free actions as

$$p_\theta(\mathcal{T}_i \mid x) = \prod_{c \in \{z,b,y\}} \prod_{j \in \mathcal{A}_i^c} p_\theta^c(a_{i,j}^c \mid \mathcal{H}_{i,j}). \tag{7}$$

Eq. (7) factorizes over sampled actions, with deterministic boundaries and the fixed answer prefix forming part of the conditioning context. We apply clipping to each action ratio:

$$\ell(r, A) = -\min\{rA, \text{clip}(r, 1 - \epsilon, 1 + \epsilon)A\}. \tag{8}$$

Let $K_{i,j}^c = D_{\text{KL}}(p_\theta^c(\cdot \mid \mathcal{H}_{i,j}) \| p_{\text{ref}}^c(\cdot \mid \mathcal{H}_{i,j}))$ be the conditional KL for a recorded history. With $\epsilon = 0.2$ and $\beta = 0.01$, the minimized loss combines the clipped term of Eq. (8) with these conditional KL regularizers:

$$\mathcal{L}_{\text{MSL}} = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} \sum_{c \in \{z,b,y\}} \text{Avg}_{j \in \mathcal{A}_i^c} \left\{\ell(r_{i,j}^c, \widehat{A}_i) + \beta K_{i,j}^c\right\}\right]. \tag{9}$$

The reference is the same frozen Stage-1 model with the LoRA adapters disabled. The latent KL is exactly

$$D_{\mathrm{KL}}(p_\theta^z \| p_{\mathrm{ref}}^z) = \frac{\|\mu_\theta - \mu_{\mathrm{ref}}\|_2^2}{2\sigma^2}. \tag{10}$$

Together with the latent KL in Eq. (10), we use the exact categorical KL for stopping and answer actions at the same recorded history. The positive KL term in Eq. (9) regularizes each action block toward the frozen reference.

At inference, latent actions equal their means, stopping selects the larger logit within the same budgets, and answer decoding is greedy. The sufficiency threshold supplies a training reward; inference uses the learned stopping policy independently of ground-truth labels.

4 Experiments

4.1 Experiment Setup

Training Setup. The benchmark experiments use Qwen3-VL-8B Bai et al. (2025b). Stage 1 uses the Visual CoT portion of FMLR-510K-EMB, and Stage 2 uses a 0.05 ZebraCoT subset (approximately 5k examples). Each stage uses one epoch on 8 NVIDIA A100 GPUs.

Benchmarks. The evaluation covers 8 vision-centric benchmarks. V* Wu & Xie (2024), HRBench-4K Wang et al. (2025b), and HRBench-8K Wang et al. (2025b) evaluate high-resolution visual question answering and fine-grained visual search. MME-RealWorld Zhang et al. (2025d) focuses on real-world image understanding. MMVP Tong et al. (2024) tests visually challenging perception cases. AI2D Kembhavi et al. (2016) evaluates diagram understanding, while TextVQA Singh et al. (2019) measures text-rich visual question answering. VisualPuzzles Song et al. (2025) assesses visual puzzle-solving and reasoning ability.

Baselines. We compare FMLR with a diverse set of multimodal models. For closed-source models, we include GPT-4o Microsoft (2024), several Gemini models, GPT-5 variants, and Claude models. For open-source base VLMs, we compare against representative Qwen3-VL Bai et al. (2025a). We also compare with tool-use or explicit visual reasoning methods, including PixelReasoner Su et al. (2025a), DeepEyes Zheng et al. (2025), DeepEyesV2 Hong et al. (2025). Finally, for latent reasoning baselines, we compare against representative methods such as LVR Li et al. (2025b), Monet Wang et al. (2025a), and CapImagine Li et al. (2026c).

4.2 Main Results

Table 1 compares FMLR-8B with representative multimodal models on 8 vision-centric benchmarks. Against the strongest Qwen3-VL-8B variant, FMLR gains on all of them, including +3.6 on MMVP, +6.8 on MME-RW, and +3.7 on V*, so fused latent reasoning improves the base model beyond both direct answering and explicit thinking. Without any explicit visual operation it stays competitive with tool-use models, outperforming PixelReasoner on all shared benchmarks and DeepEyes on most. Among latent reasoning methods it surpasses LVR-7B, Monet-7B, and CapImagine-7B on most shared benchmarks. Re-implemented on the same Qwen3-VL-8B initialization, LVR-8B and Monet-8B still lose on all eight, by 8.1 and 4.8 on MME-RW, so the advantage does not come from a stronger backbone. FMLR also emits about nine output tokens per question (Table 3) rather than the hundreds used by explicit thinking, and stays competitive with strong open-source and closed-source models despite its 8B backbone.

Appendix B examines the same training recipe on InternVL3.5-8B and Qwen3-VL-4B, reusing the fused embedding targets. FMLR improves InternVL3.5-8B on all eight benchmarks and Qwen3-VL-4B on seven, with a tie on HR-4K. The gains include 8.7 points on HR-8K for InternVL3.5-8B and 9.8 points on MME-RW for Qwen3-VL-4B. These results extend the comparison beyond the primary backbone to another model family and a smaller model.

Table 1: Comparison of multimodal models across 8 vision-centric benchmarks. LVR-8B and Monet-8B denote our re-implementation of these latent reasoning methods on the same Qwen3-VL-8B backbone as FMLR, for a controlled comparison.

Model	MMVP	HR-4K	HR-8K	AI2D	MME-RW	V*	TextVQA	VisPuz
Closed-source Models								
GPT-4o Microsoft (2024)	72.0	59.0	55.5	84.6	52.0	67.5	77.4	41.3
Gemini-2.0-Flash Comanici et al. (2025)	–	–	–	–	–	–	–	45.0
Gemini-2.5-Pro (Thinking) Comanici et al. (2025)	–	87.3	85.4	90.9	–	83.8	–	60.9
GPT-5 high Singh et al. (2025)	–	–	–	89.7	–	72.8	–	57.3
GPT-5 mini high Singh et al. (2025)	–	78.6	74.4	88.2	–	78.6	–	59.3
Claude 3.5 Sonnet Anthropic (2024)	45.5	–	–	81.2	51.6	–	74.1	42.4
Open-source Base VLMs								
Qwen3-VL-8B (Thinking) Bai et al. (2025b)	76.3	68.9	63.5	79.4	45.8	77.0	76.7	31.0
Qwen3-VL-8B (Instruct) Bai et al. (2025b)	76.7	75.6	69.1	81.3	51.6	80.6	81.7	36.1
Tool-Use Models								
PixelReasoner Su et al. (2025a)	67.0	72.9	66.9	–	49.7	80.1	–	–
DeepEyes Zheng et al. (2025)	–	75.1	72.6	37.0	53.2	83.3	40.4	–
DeepEyesV2 Hong et al. (2025)	–	77.9	**73.8**	–	**64.9**	81.8	–	–
Latent Reasoning Methods								
LVR-7B Li et al. (2025b)	71.7	70.8	63.0	77.3	50.6	81.7	75.1	29.4
Monet-7B Wang et al. (2025a)	–	71.0	68.0	–	55.5	83.3	–	35.0
CapImagine Li et al. (2026b)	–	74.1	70.7	–	54.8	**85.9**	–	–
LVR-8B Li et al. (2025b)	77.7	70.0	63.8	81.7	50.3	82.7	80.6	34.4
Monet-8B Wang et al. (2025a)	78.3	76.2	70.0	82.3	53.6	83.2	80.1	36.9
FMLR-8B	**80.3**	**78.8**	71.9	**83.3**	58.4	84.3	**81.9**	**37.2**
Δ vs. Qwen3-VL-8B	+3.6	+3.2	+2.8	+2.0	+6.8	+3.7	+0.2	+1.1

Table 2: Ablation of training data, supervision, latent representation, and optimization on Qwen3-VL-8B. Accuracy (%); best per column in **bold**.

Setting	**MMVP**	**HR-4K**	**HR-8K**	**AI2D**	**MME-RW**	**V***	**TextVQA**	**VisPuz**
Base models								
Qwen3-VL-8B-Instruct	76.7	75.6	69.1	81.3	51.6	80.6	81.7	36.1
Qwen3-VL-8B-Thinking	76.3	68.9	63.5	79.4	45.8	77.0	76.7	31.0
Explicit CoT supervision (same image-question pairs)								
+ Visual CoT SFT	77.3	72.3	66.4	81.5	49.5	83.8	83.5	31.8
+ Auto-CoT	76.0	68.4	63.7	77.3	42.0	72.8	75.5	33.7
+ Refine CoT	75.3	75.6	70.2	81.4	51.4	82.7	81.9	**37.5**
Matched-data explicit supervision (adds the same ZebraCoT subset as FMLR)								
+ Visual CoT SFT + ZebraCoT SFT	72.0	69.0	58.8	78.5	41.3	69.6	75.2	31.7
+ Refine CoT SFT + ZebraCoT SFT	78.7	74.0	67.6	81.7	49.0	83.2	**83.6**	33.6
+ Refine CoT SFT + ZebraCoT GRPO	76.3	74.0	67.4	81.2	49.0	80.1	82.4	34.2
Latent form and optimization (RL stage controlled)								
Non-latent CoT + GRPO	71.7	74.4	69.9	76.5	54.9	82.2	81.7	29.0
+ Separate Latent (SFT)	79.0	76.2	69.9	82.3	54.6	82.5	81.6	33.3
+ Separate Latent + MSL-GRPO	79.7	77.5	71.4	82.3	50.7	82.7	81.2	34.2
+ Fused Latent (SFT)	78.3	73.9	71.8	79.2	55.4	83.2	78.0	33.3
+ Fused Latent + GRPO	78.3	77.9	67.1	82.3	55.8	83.8	80.1	36.4
FMLR-8B (Fused Latent + MSL-GRPO)	**80.3**	**78.8**	**71.9**	**83.3**	**58.4**	**84.3**	81.9	37.2

4.3 Ablation Studies

Ablation setup. Table 2 compares variants built on Qwen3-VL-8B. The upper blocks vary training data and supervision; the lower block compares reasoning forms and optimization on matched data.

Data quality and supervision form. Auto-CoT reduces V* accuracy to 72.8 and MME-RW to 42.0; verification and refinement recover these scores to 82.7 and 51.4. FMLR outperforms the explicit-supervision and text-space GRPO variants on most benchmarks under the matched data budget.

Latent form. Fusing the step-level text and image embedding into one supervision target pays off once latent-level RL is applied. With supervised alignment alone the fused and separate variants are comparable, with fused ahead on three benchmarks and separate ahead on four. A non-latent baseline that keeps explicit text reasoning and applies GRPO on the same data stays clearly below both, for example 76.5 on AI2D and 29.0 on VisualPuzzles. Under the same MSL-GRPO, the fused model surpasses the separate-latent model on all eight benchmarks.

Optimization stage. MSL-GRPO improves fused latent supervision on all eight benchmarks, including HR-4K from 73.9 to 78.8, MME-RW from 55.4 to 58.4, and VisualPuzzles from 33.3 to 37.2. Its effect on separate latents is mixed, with gains on high-resolution benchmarks and a decline on MME-RW from 54.6 to 50.7.

With fused latent supervision fixed, MSL-GRPO also improves over standard GRPO on all eight benchmarks, with the largest gains on HR-8K (+4.8) and MME-RW (+2.6). These results support the combined use of fused supervision and MSL-GRPO.

Accuracy–Token Analysis Table 3 compares FMLR with explicit-reasoning and latent baselines on accuracy and average output length. FMLR has the highest accuracy on all four benchmarks while emitting about nine tokens, whereas Qwen3-VL-8B Thinking and Refine CoT need hundreds. It also beats LVR-7B on both axes, with higher accuracy and fewer tokens everywhere. Visual CoT SFT is shorter still on MMVP and V*, but is less accurate on all four. Latent reasoning therefore improves accuracy without long explicit traces.

Compared with Visual CoT SFT on HR-4K and HR-8K, FMLR improves accuracy by 6.5 and 5.5 points while reducing average output length to nine tokens, improving both accuracy and output length on these high-resolution benchmarks.

Table 3: Comparison of accuracy and average inference output tokens.

	MMVP		**HR-4K**		**HR-8K**		**V***	
Model	Acc↑	Tok↓	Acc↑	Tok↓	Acc↑	Tok↓	Acc↑	Tok↓
Qwen3-VL-8B (Instruct)	76.7	91.6	75.6	168.0	69.1	169.9	80.6	4.1
Qwen3-VL-8B (Thinking)	76.3	302.2	68.9	389.0	63.5	422.9	77.0	182.8
Qwen3-VL-8B + Visual CoT SFT	77.3	**7.6**	72.3	31.7	66.4	25.8	83.8	**2.2**
Qwen3-VL-8B + Refine CoT SFT	75.3	217.4	75.6	197.4	70.2	208.3	82.7	156.4
LVR-7B	71.7	24.9	70.8	29.1	63.0	29.9	81.7	20.0
FMLR-8B	**80.3**	9.9	**78.8**	**9.0**	**71.9**	**9.0**	**84.3**	9.0

5 Conclusion

We propose Fused Multi-Latent Reasoning (FMLR), a latent-space reasoning framework that uses fused vision-language embeddings to construct informative multimodal reasoning units. Stage 1 adapts fused step-level representations into latent states, preserving semantic intent, visual evidence, and cross-modal grounding. Stage 2 uses Minimum-Sufficient Latent GRPO (MSL-GRPO) to optimize answer correctness, early sufficiency, and latent reasoning length, with explicit stopping and a cost for each generated latent unit. We construct FMLR-400K-SFT and FMLR-510K-EMB to support multimodal reasoning supervision and fused embedding adaptation. Experiments on vision-centric benchmarks show stronger reasoning accuracy with fewer output tokens than explicit-reasoning baselines. Under the same Qwen3-VL-8B backbone and a matched data budget, fused supervision outperforms separate latent supervision, highlighting the value of combining semantic and visual evidence within each reasoning unit. Experiments on another model family and a smaller backbone further examine the transferability of this training recipe. The framework depends on the quality of its fused supervision; future work can explore more robust adaptation across embedding models, backbone families, and scales.

Ethics Statement

This work studies latent multimodal reasoning and does not involve human subjects or personally identifying data. All training and evaluation data come from publicly available vision-language benchmarks. Latent reasoning is less transparent than explicit chains of thought, so we discuss the resulting need for careful evaluation and human oversight in the Broader Impacts section of the appendix.

Reproducibility Statement

The main text presents the method and experimental setup. Code and datasets will be released upon acceptance.

AI Use Statement

We used AI-based assistants for minor coding support and for language polishing. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final text, claims, and results.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Anthropic. Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet, 2024. URL https://assets.anthropic.com/m/1cd9d098ac3e6467/original/Claude-3-Model-Card-October-Addendum.pdf. Accessed: 2025-03-07.

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.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

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*, 2025b.

Yilong Chen, Junyuan Shang, Zhenyu Zhang, Yanxi Xie, Jiawei Sheng, Tingwen Liu, Shuohuan Wang, Yu Sun, Hua Wu, and Haifeng Wang. Inner thinking transformer: Leveraging dynamic depth scaling to foster adaptive internal thinking. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28241–28259, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1369. URL https://aclanthology.org/2025.acl-long.1369/.

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, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Jingcheng Deng, Liang Pang, Zihao Wei, Shicheng Xu, Zenghao Duan, Kun Xu, Yang Song, Huawei Shen, and Xueqi Cheng. Llm latent reasoning as chain of superposition, 2026. URL https://arxiv.org/abs/2510.15522.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images. *arXiv preprint arXiv:2505.15879*, 2025.

Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning, 2026. URL https://arxiv.org/abs/2510.27492.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space, 2025. URL https://arxiv.org/abs/2412.06769.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 235–251. Springer, 2016.

Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. Zebra-cot: A dataset for interleaved vision language reasoning, 2025a. URL https://arxiv.org/abs/2507.16746.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*, 2025b.

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*, pp. 19730–19742. PMLR, 2023.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026a. URL https://arxiv.org/abs/2601.04720.

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. *arXiv preprint arXiv:2602.22766*, 2026b.

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, 2026c. URL https://arxiv.org/abs/2602.22766.

Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyue Wei, Sheng Liu, and Xin Eric Wang. Reasoning within the mind: Dynamic multimodal interleaving in latent space, 2026. URL https://arxiv.org/abs/2512.12623.

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

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*, pp. 26296–26306, 2024.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2034–2044, October 2025.

Bo Lv, Yasheng Sun, Junjie Wang, and Haoxiang Shi. Onelatent: Single-token compression for visual latent reasoning, 2026. URL https://arxiv.org/abs/2602.13738.

Yuanhuiyi Lyu, Kaiyu Lei, Ziqiao Weng, Xu Zheng, Lutao Jiang, Teng Li, Yangfu Li, Ziyuan Huang, Linfeng Zhang, and Xuming Hu. Struvis: Enhancing reasoning-based text-to-image generation via thinking with structured vision, 2026. URL https://arxiv.org/abs/2603.06032.

Microsoft. Introducing gpt-4o-2024-08-06 api with structured outputs on azure. https://techcommunity.microsoft.com/blog/azure-ai-services-blog/introducing-gpt-4o-2024-08-06-api-with-structured-outputs-on-azure/4232684, 2024. Accessed: 2025-03-07.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, and Jie Tang. Cogcom: A visual language model with chain-of-manipulations reasoning, 2025. URL https://arxiv.org/abs/2402.04236.

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, 2025. URL https://arxiv.org/abs/2511.19418.

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, 2025. URL https://arxiv.org/abs/2512.10941.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge, 2025. URL https://arxiv.org/abs/2504.10342.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025a.

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*, 2025b.

Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Ruihua Song, and Jian Luan. Think silently, think fast: Dynamic latent compression of llm reasoning chains, 2026. URL https://arxiv.org/abs/2505.16552.

5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghua Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao, Hongwei Liu, Hongxi Yan, Huan Liu, Huilong Chen, Ji Li, Jiajing Zhao, Jiamin Ren, Jian Jiao, Jiani Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao, Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai Zhi, Junli Gao, Ke Wang, Lekang Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghuan Xu, Mingming Zhao, Mingshu Zhai, Pengfan Du, Qian Dong, Shangde Lei, Shangqing Tu, Shangtong Yang, Shaoyou Lu, Shijie Li, Shuang Li, Shuang-Li, Shuxun Yang, Sibo Yi, Tianshu Yu, Wei Tian, Weihan Wang, Wenbo Yu, Weng Lam Tam, Wenjie Liang, Wentao Liu, Xiao Wang, Xiaohan Jia, Xiaotao Gu, Xiaoying Ling, Xin Wang, Xing Fan, Xingru Pan, Xinyuan Zhang, Xinze Zhang, Xiuqing Fu, Xunkai Zhang, Yabo Xu, Yandong Wu, Yida Lu, Yidong Wang, Yilin Zhou, Yiming Pan, Ying Zhang, Yingli Wang, Yingru Li, Yinpei Su, Yipeng Geng, Yitong Zhu, Yongkun Yang, Yuhang Li, Yuhao Wu, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yuxuan Zhang, Zezhen Liu, Zhen Yang, Zhengda Zhou, Zhongpei Qiao, Zhuoer Feng, Zhuorui Liu, Zichen Zhang, Zihan Wang, Zijun Yao, Zikang Wang, Ziqiang Liu, Ziwei Chai, Zixuan Li, Zuodong Zhao, Wenguang Chen, Jidong Zhai, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models, 2025. URL https://arxiv.org/abs/2508.06471.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. URL https://arxiv.org/abs/2401.06209.

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. *arXiv preprint arXiv:2511.21395*, 2025a.

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, pp. 7907–7915, 2025b.

Wenshuo Wang. Llm reasoning is latent, not the chain of thought, 2026. URL https://arxiv.org/abs/2604.15726.

Xiaoqiang Wang, Suyuchen Wang, Yun Zhu, and Bang Liu. System-1.5 reasoning: Traversal in language and latent spaces with dynamic shortcuts, 2025c. URL https://arxiv.org/abs/2505.18962.

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

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. Sim-cot: Supervised implicit chain-of-thought. *arXiv preprint arXiv:2509.20317*, 2025.

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*, pp. 13084–13094, 2024.

Tianrun Xu, Guanyu Chen, Ye Li, Yuxin Xi, Zeyu Mu, Ruichen Wang, Tianren Zhang, Haichuan Gao, and Feng Chen. Ouro: A self-bootstrapped framework for enhancing multimodal scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18240–18251, 2025a.

Tianrun Xu, Haoda Jing, Ye Li, Yuquan Wei, Jun Feng, Guanyu Chen, Haichuan Gao, Tianren Zhang, and Feng Chen. Defacto: Counterfactual thinking with images for enforcing evidence-grounded and faithful reasoning. *arXiv preprint arXiv:2509.20912*, 2025b.

Shilin Yan, Jintao Tong, Hongwei Xue, Xiaojun Tang, Yangyang Wang, Kunyu Shi, Guannan Zhang, Ruixuan Li, and Yixiong Zou. Act wisely: Cultivating meta-cognitive tool use in agentic multimodal models. *arXiv preprint arXiv:2604.08545*, 2026.

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

Zhenrui Yue, Bowen Jin, Huimin Zeng, Honglei Zhuang, Zhen Qin, Jinsung Yoon, Lanyu Shang, Jiawei Han, and Dong Wang. Hybrid latent reasoning via reinforcement learning, 2025. URL https://arxiv.org/abs/2505.18454.

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

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. *arXiv preprint arXiv:2502.17422*, 2025b.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv preprint arXiv:2505.15436*, 2025c.

Yang Zhang, Danyang Li, Yuxuan Li, Xin Zhang, Tianyu Xie, Mingming Cheng, and Xiang Li. Crystal: Spontaneous emergence of visual latents in mllms, 2026. URL https://arxiv.org/abs/2602.20980.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans?, 2025d. URL https://arxiv.org/abs/2408.13257.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A Appendix A: Latent Representation Analysis

The representation and perturbation analyses inspect two selected latent positions in each sampled Stage-1 trajectory. Section A.3 examines the sufficiency threshold used by MSL-GRPO.

A.1 Representation Collapse

Table 4 reports low cosine similarity between the two selected latent states within each question, indicating that they remain distinct.

Table 4: Intra-question adjacent cosine similarity between the two latent states of the same question. Low values show that the two latent steps stay distinct.

MMVP	HR-4K	HR-8K	AI2D	MME-RW	V*	TextVQA	VisPuz
0.35	0.29	0.29	0.27	0.36	0.32	0.25	0.37

Figure 4 compares within- and cross-dataset cosine similarities after removing the shared mean direction at each latent position.

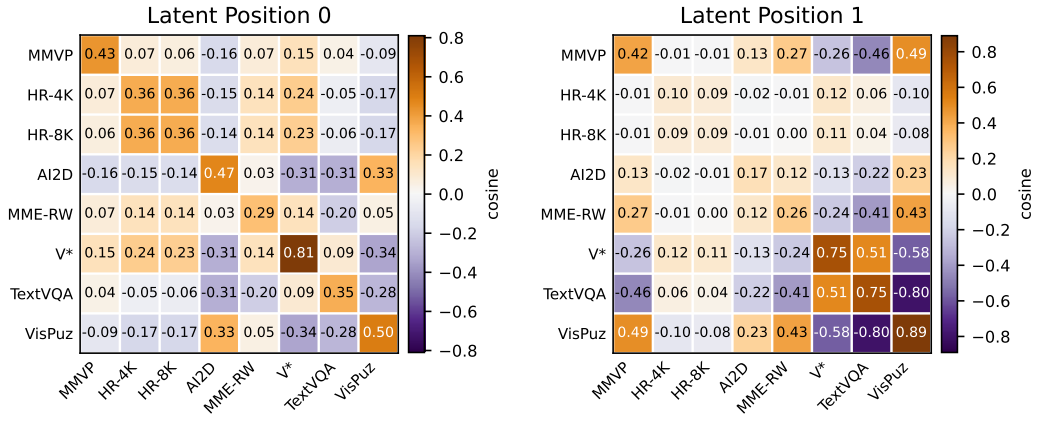


Figure 4: Cross-dataset centered cosine similarity of the two latent positions, with the diagonal giving the within-dataset similarity.

A.2 Effect of Latent Perturbations

We inject random noise into latent states at inference on 100 samples per benchmark. Perturbing one or two states reduces macro accuracy from 68.9 to 57.4 or 51.3, respectively (Table 5), indicating that predictions depend on the latent representations.

Table 5: Effect of perturbing latent states with random noise (accuracy %, 100 samples per benchmark).

Benchmark	No perturbation	Noise on 1 latent	Noise on 2 latents
MMVP	87	75	75
V*	81	58	40
HR-4K	81	73	58
HR-8K	78	68	54
MME-RW	58	51	49
MMStar	63	47	56
VisPuz	34	30	27
Macro	68.9	57.4	51.3

A.3 SUFFICIENCY-THRESHOLD SENSITIVITY

Table 6 compares sufficiency thresholds $\tau \in \{0.2, 0.7, 0.8\}$ in Eq. (4). The default $\tau = 0.7$ achieves the highest accuracy on most benchmarks; its FMLR-8B row repeats the main result for comparison.

Table 6: Sensitivity of FMLR-8B to the sufficiency threshold τ. Accuracy (%); best per column in **bold**.

Setting	MMVP	HR-4K	HR-8K	AI2D	MME-RW	V*	TextVQA	VisPuz
$\tau = 0.2$ (loose)	**80.3**	78.1	71.5	83.0	**58.6**	**84.3**	**82.1**	36.1
$\tau = 0.8$ (strict)	79.6	76.4	71.4	81.3	58.2	80.6	81.7	36.9
FMLR-8B ($\tau = 0.7$, default)	**80.3**	**78.8**	**71.9**	**83.3**	58.4	**84.3**	81.9	**37.2**

A.4 FAILURE-MODE ANALYSIS

Table 7 shows gains in inductive and spatial puzzles and both MME-RealWorld categories, but lower accuracy on algorithmic, analogical, and deductive puzzles than the base model. These differences motivate future work on combining latent and explicit reasoning.

Table 7: Failure-mode analysis (accuracy %; best per column in **bold**). Left: VisualPuzzles by reasoning category (Alg.: algorithmic; Ana.: analogical; Ded.: deductive; Ind.: inductive; Spa.: spatial). Right: MME-RealWorld-lite split into perception and reasoning.

	VisualPuzzles (by reasoning category)						**MME-RealWorld-lite**		
Setting	Overall	Alg.	Ana.	Ded.	Ind.	Spa.	Perception	Reasoning	Overall
Qwen3-VL-8B-Inst.	36.13	39.31	**45.50**	49.50	23.44	26.22	54.34	47.33	51.56
+ Visual CoT SFT	31.85	38.55	27.01	50.00	22.97	23.08	54.32	42.00	49.50
+ Auto-CoT	33.73	40.84	42.65	43.50	25.36	19.93	45.43	36.66	42.01
+ Refine CoT	**37.50**	**43.51**	43.60	**53.00**	24.40	26.22	56.54	43.33	51.38
FMLR-8B	37.24	35.50	37.91	47.50	**33.97**	**33.57**	**60.91**	**54.40**	**58.36**

B Appendix B: Generalization across Model Families and Scales

These experiments apply the two-stage FMLR recipe, with the same fused embedding targets, to a backbone from a different model family (InternVL3.5-8B) and to a smaller backbone within the Qwen family (Qwen3-VL-4B). The fused targets are reused without re-embedding, and the embedding model is used only offline during training. Tables 8 and 9 show that FMLR improves over each base model on most benchmarks, so the gains are not tied to a single backbone or scale.

Table 8: Generalization to a different model family (InternVL3.5-8B). Accuracy (%); best per column in **bold**.

Setting	MMVP	HR-4K	HR-8K	AI2D	MME-RW	V*	TextVQA	VisPuz
InternVL3.5-8B (w/o think)	75.3	61.9	53.2	74.8	48.5	71.2	73.9	31.1
+ Visual CoT SFT	73.3	65.1	55.2	79.6	47.9	72.3	74.5	33.0
+ Refine CoT SFT	74.7	64.1	57.5	80.0	48.9	68.1	75.3	30.8
+ Refine CoT + GRPO	76.0	66.4	59.2	**80.4**	50.3	71.2	**75.4**	30.3
FMLR-InternVL3.5-8B	**78.3**	**68.4**	**61.9**	80.2	**54.0**	**73.3**	75.1	**35.1**
Δ vs. w/o think	+3.0	+6.5	+8.7	+5.4	+5.5	+2.1	+1.2	+4.0

Table 9: Scaling to a smaller backbone (Qwen3-VL-4B). Accuracy (%); best per column in **bold**.

Setting	MMVP	HR-4K	HR-8K	AI2D	MME-RW	V*	TextVQA	VisPuz
Qwen3-VL-4B-Instruct	78.7	**78.8**	74.8	80.2	46.5	80.6	81.0	30.5
+ Visual CoT SFT	79.0	75.8	68.5	80.4	55.8	80.6	81.8	27.7
+ Refine CoT SFT	76.0	75.6	69.5	80.7	52.2	80.6	**82.7**	32.1
+ Refine CoT + GRPO	76.7	76.9	72.4	80.8	54.0	80.6	82.2	33.3
FMLR-4B	**80.3**	**78.8**	**75.6**	**82.2**	**56.3**	**82.7**	81.6	**36.4**
Δ vs. Instruct	+1.6	+0.0	+0.8	+2.0	+9.8	+2.1	+0.6	+5.9

C Appendix C: Verification and Refinement Prompts

This appendix lists the system prompts for forward verification, backward verification, and reasoning-trace refinement.

System Prompt for Forward Verification

You are a vision-language model used for dataset verification.
Your task is to verify whether a multimodal reasoning chain is sufficient to support the final answer to a given question.
The input consists of:

1. an original image,
2. a question about the image,
3. a multi-step reasoning chain,
4. a set of cropped image regions corresponding to the referenced visual evidence in each reasoning step.

Your goal is to judge whether the provided reasoning chain is visually grounded, logically coherent, and sufficient for deriving the final answer.
Verification principles:

1. Each reasoning step should be grounded in the corresponding cropped visual region.
2. The reasoning chain should form a coherent progression toward the answer.
3. The final answer must be supported by the reasoning chain and the visual evidence.
4. Do not assume any information that is not visible in the original image or the cropped regions.
5. If the reasoning is incomplete, unsupported, or inconsistent with the visual evidence, the sample should be regarded as invalid.

Instructions:

- Read the question carefully.
- Examine the reasoning chain step by step.
- Use the cropped image regions as evidence for checking whether each reasoning step is grounded.
- Determine whether the final answer can be reliably derived from the reasoning chain and the provided visual evidence.
- Output only the final verified answer.
- If the reasoning chain is invalid or insufficient, output the token `INVALID`.

Output format:

- If valid: output a single final answer only.
- If invalid: output exactly `INVALID`.

System Prompt for Backward Verification

You are a vision-language model used for dataset verification.
Your task is to reconstruct a plausible multi-step reasoning chain from a given question, answer, and grounded visual evidence.
The input consists of:

1. an original image,
2. a question about the image,
3. a final answer,
4. a set of cropped image regions corresponding to the relevant visual evidence.

Your goal is to generate a reasoning chain that is consistent with the answer and grounded in the provided visual evidence.
Reconstruction principles:

1. Each reasoning step must be supported by the original image and the corresponding cropped region.

2. The reasoning should explain how the answer is derived from the visual evidence.
3. The reconstructed reasoning chain should be semantically consistent with the answer.
4. Do not invent visual details or reasoning steps that are not supported by the image.
5. The reasoning chain should be concise, stepwise, and visually grounded.

Instructions:

- Read the question and the given answer carefully.
- Examine the cropped image regions and identify the relevant visual evidence.
- Reconstruct a multi-step reasoning chain that could lead to the answer.
- Each step should include:
  - a short textual reference to the relevant visual region,
  - a reasoning statement grounded in that region.
- Keep the reasoning focused on the evidence necessary for deriving the answer.

Output format:

```
<think>
Step 1: [Region description] Reasoning statement
Step 2: [Region description] Reasoning statement
...
</think>
```

Do not output the final answer again. Output only the reconstructed reasoning chain.

System Prompt for Reasoning Trace Refinement

You are a vision-language model used for dataset refinement.
Your task is to repair a failed multimodal reasoning chain when possible, or determine that the sample should be removed if the reasoning cannot be reliably corrected.
The input consists of:

1. an original image,
2. a question about the image,
3. a failed multi-step reasoning chain,
4. optionally, the associated answer and the failure signal from verification.

Your goal is to examine the failed reasoning chain, identify whether the failure comes from incorrect visual grounding, weak reasoning, missing evidence, or answer inconsistency, and then either:

- generate a corrected reasoning chain, or
- decide that the sample is too unreliable and should be removed.

Refinement principles:

1. Preserve the original question and remain faithful to the image.
2. Correct region references if they are inaccurate, ambiguous, or visually unsupported.
3. Revise reasoning steps if they are logically weak, redundant, or inconsistent with the visual evidence.
4. Keep the reasoning concise and stepwise.
5. Do not invent visual details that are not supported by the image.
6. If the question cannot be answered reliably from the image, or if the reasoning trace cannot be repaired with high confidence, mark the sample as `REMOVE`.

Instructions:

- Read the question and inspect the image carefully.
- Examine the failed reasoning chain step by step.
- Identify whether each step is visually grounded and logically necessary.
- If the chain can be fixed, rewrite it into a corrected multi-step reasoning trace.
- If the chain is fundamentally unreliable, contradictory, or unsupported by the image, output `REMOVE`.

Output format:

- If refinable, output:

```
<think>
Step 1: [Region description] Reasoning statement
Step 2: [Region description] Reasoning statement
...
</think>
<answer>final answer</answer>
```

- If not refinable, output exactly:

```
REMOVE
```

D APPENDIX D: DATASET VISUALIZATION

The following examples compare original Visual CoT annotations with their refined FMLR-400K-SFT counterparts.

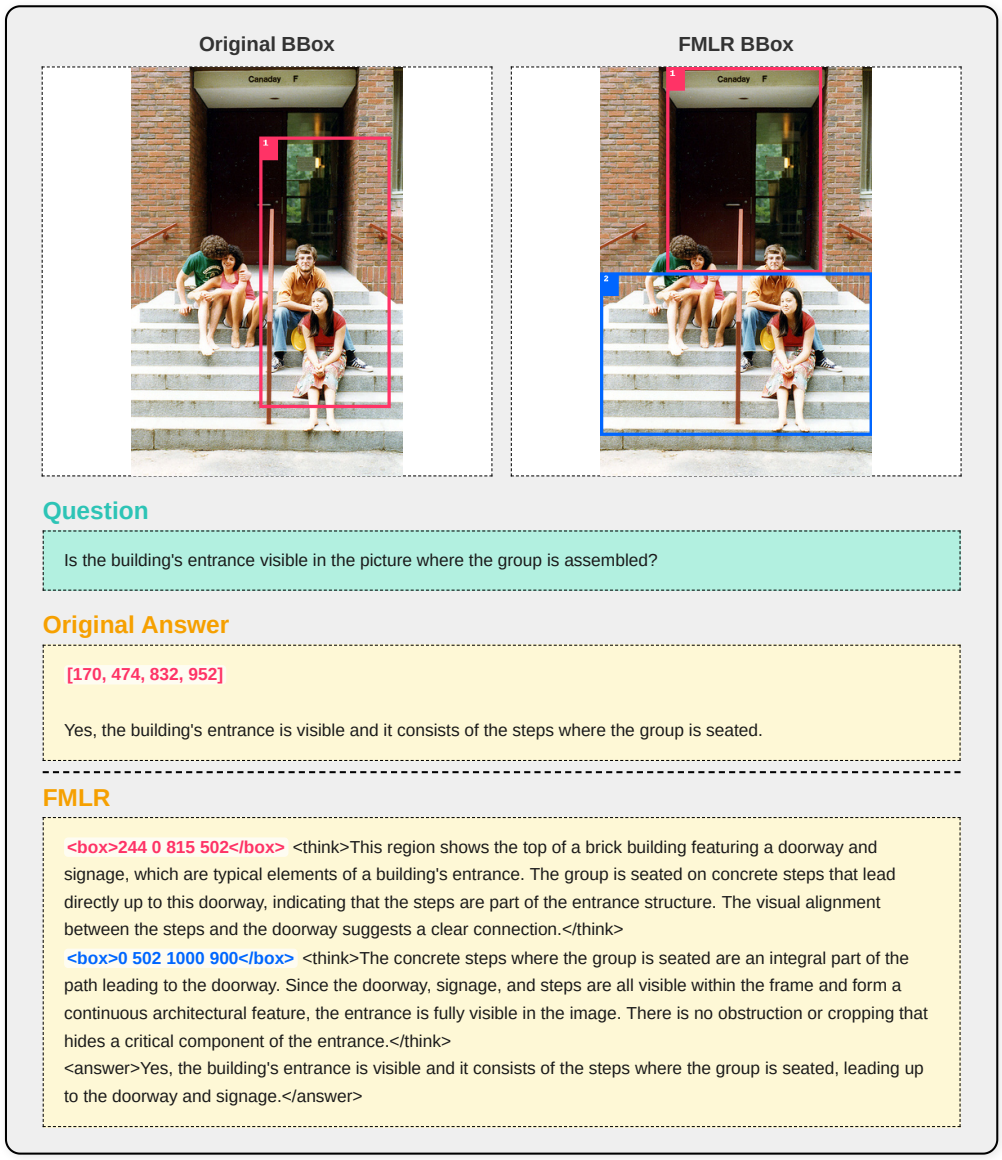


Figure 5: Comparison between the original Visual CoT annotation and the refined FMLR annotation.

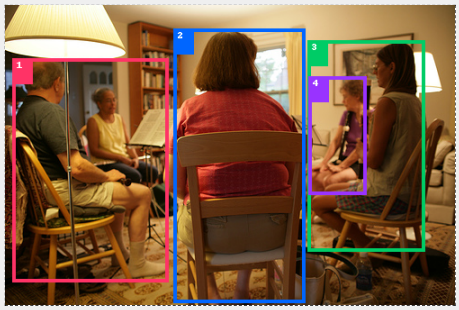


Figure 6: Comparison between the original Visual CoT annotation and the refined FMLR annotation.

Original BBox



FMLR BBox



Question

Are the musicians gathered in a particular formation and if so, what is it?

Original Answer

[178, 348, 348, 574]

Yes, the musicians are gathered in a circular formation.

FMLR

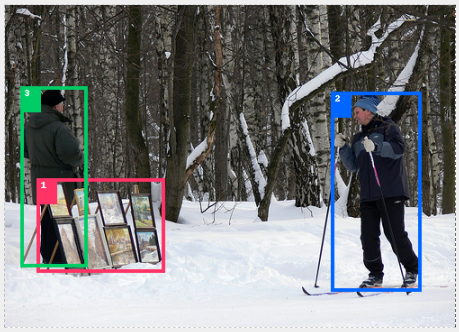
<box>17 179 363 924</box> <think>This region shows two seated individuals facing inward, suggesting they are oriented toward a central point. This initial observation indicates a possible group arrangement with a shared focus, but more context is needed to determine the overall formation.</think>
<box>374 79 666 991</box> <think>Another individual is seated with their back to the camera, yet their body orientation and posture suggest they are facing toward the same central area as the others. This reinforces the idea of a collective focal point, consistent with a circular or semi-circular setup, though the full layout remains unclear.</think>
<box>673 118 933 821</box> <think>A third musician is seated and clearly turned toward the center, actively holding an instrument. Their directional focus aligns with the others, continuing a consistent pattern of inward-facing positioning, which supports the hypothesis of a circular formation.</think>
<box>675 239 803 633</box> <think>The fourth individual is also seated and oriented toward the group's center, with musical equipment arranged around them. Their placement completes a coherent pattern of musicians seated around a central space, all facing inward. Given the consistent angular positions and mutual orientation, the arrangement strongly indicates a circular formation.</think>
<answer>Yes, the musicians are gathered in a circular formation.</answer>

Figure 7: Comparison between the original Visual CoT annotation and the refined FMLR annotation.

Original BBox



FMLR BBox



Question

Are the pictures being observed by the skier up for sale, and if so, where are they presented?

Original Answer

[108, 614, 180, 740]

Yes, there is artwork presented in the snow that seems to be offered for sale.

FMLR

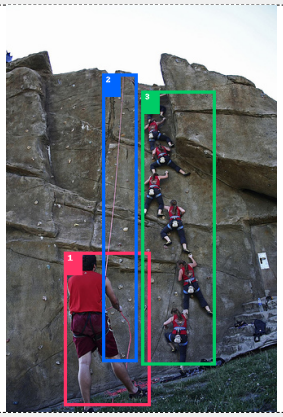
<box>70 540 354 830</box> <think>The image shows multiple framed items standing upright in the snow, arranged in a deliberate manner that suggests a formal display. Their positioning and visibility in an outdoor, public-appearing setting indicate they are likely part of an exhibition or market setup, possibly for commercial purposes.</think>
<box>724 271 925 889</box> <think>The skier is facing the displayed items and appears to be observing them closely, which may indicate interest. While observation alone does not confirm a sale, it aligns with the behavior of someone evaluating items that are available for purchase.</think>
<box>34 253 182 814</box> <think>Another individual is standing near the display, positioned in a way that suggests oversight or stewardship—potentially as a vendor or attendant. This presence supports the interpretation that the display is organized and possibly transactional, especially in a context like an outdoor art market or pop-up sale.</think>
<answer>Yes, there is artwork presented in the snow that seems to be offered for sale.</answer>

Figure 8: Comparison between the original Visual CoT annotation and the refined FMLR annotation.

Original BBox



FMLR BBox



Question

What is the person standing on the ground holding onto?

Original Answer

[404, 162, 542, 976]

The person standing on the ground is holding onto a rope, which is attached to the climbers for their safety.

FMLR

<box>216 606 533 984</box> <think>This region shows a person standing on the ground, wearing a red shirt, and holding an object in their hands. The object extends upward toward the climbers, suggesting a physical connection. This area is crucial for identifying the object being held, as it captures the person's interaction with it.</think>
<box>357 168 485 877</box> <think>This region reveals a long, thin, red-colored object extending vertically from the person on the ground toward the climbers above. The object appears taut and linear, consistent with a rope or cord under tension. Its continuity from the person's hands upward indicates they are actively holding and managing it.</think>
<box>500 210 772 885</box> <think>This region shows multiple climbers ascending a rock wall, each equipped with climbing harnesses. The red object from the previous region is visually connected to the climbers, running from the person on the ground to at least one climber's harness. This confirms the object is a safety rope used in belaying, where the person on the ground controls the rope to protect the climbers from falling.</think>
<answer>The person standing on the ground is holding onto a rope, which is attached to the climbers for their safety.</answer>

Figure 9: Comparison between the original Visual CoT annotation and the refined FMLR annotation.

Original BBox



FMLR BBox



Question

What kind of weather or outdoor conditions can be inferred from what the group of people are doing in the park?

Original Answer

[196, 538, 256, 608]

The group of people are enjoying a beautiful day in the park, indicating that the weather is likely pleasant and conducive to outdoor activities.

FMLR

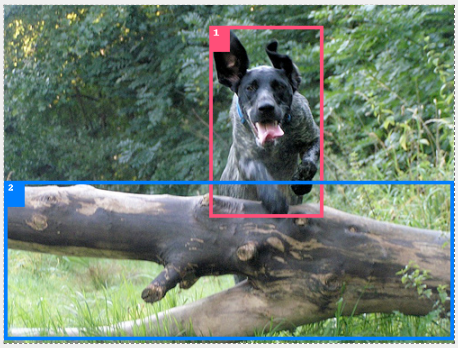
<box>0 0 999 550</box> <think>The sky is visible through the tree canopy and appears bright and clear, with no signs of rain or dense cloud cover, indicating favorable outdoor conditions and suggesting that the weather is dry and likely sunny.</think>
<box>107 540 999 878</box> <think>Multiple people are seated at picnic tables, some wearing hats and light clothing, which implies they are dressed appropriately for warm, comfortable weather and are engaged in a prolonged outdoor gathering.</think>
<box>254 447 366 563</box> <think>A colorful inflatable play structure is set up in the background, which is typically used in dry, mild to warm conditions, further supporting the inference that the weather is suitable for outdoor recreational activities, especially for children.</think>
<box>470 708 562 920</box> <think>A large beverage dispenser and multiple drink containers are present on the table, indicating that hydration is a priority, which is common during warm or sunny weather when people are more likely to perspire and need fluids.</think>
<answer>The group of people are enjoying a beautiful day in the park, indicating that the weather is likely pleasant and conducive to outdoor activities.</answer>

Figure 10: Comparison between the original Visual CoT annotation and the refined FMLR annotation.

Original BBox



FMLR BBox



Question

Is there any specific part of a tree that the dog interacts with in the photo?

Original Answer

[2, 522, 1000, 874]

The dog is interacting with a log, which is part of a fallen tree.

FMLR

<box>456 63 708 629</box> <think>This region shows the dog in mid-air with its paws extended toward a large, horizontal, wooden object on the ground. The object has visible bark, knots, and a cylindrical shape, which are characteristic of a tree trunk or large branch. Given the dog's posture and proximity, it is likely engaged in an interaction such as jumping over or landing on this object.</think>
<box>0 520 999 991</box> <think>The object the dog is interacting with is a thick, fallen wooden structure with clear features of a tree trunk, including bark texture and natural irregularities like knots. Its position on the ground and the dog's motion above it confirm that the dog is interacting with a log—a segment of a fallen tree. This constitutes a specific part of a tree, namely a trunk section, now detached and lying on the ground.</think>
<answer>The dog is interacting with a log, which is part of a fallen tree.</answer>

Figure 11: Comparison between the original Visual CoT annotation and the refined FMLR annotation.

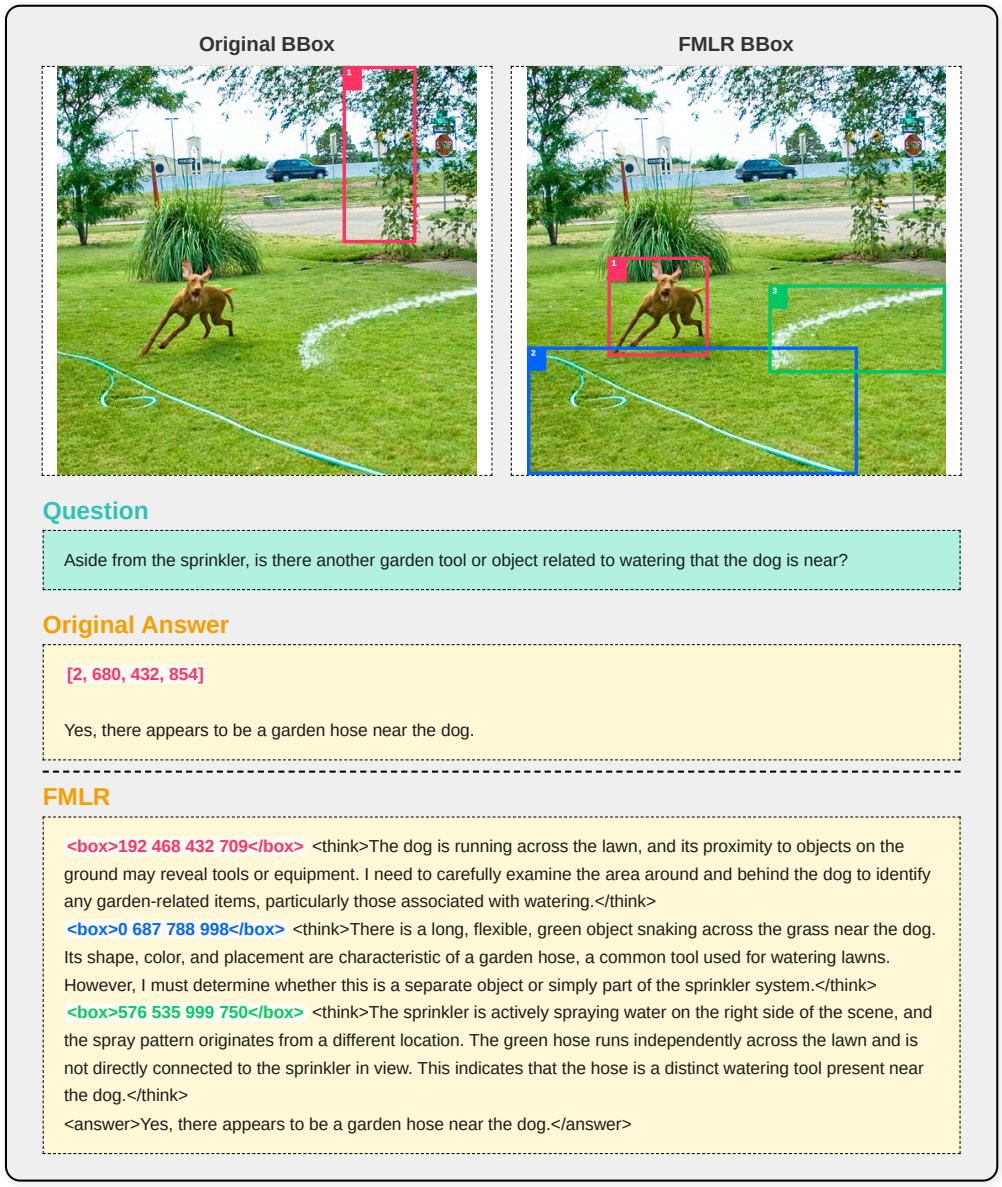


Figure 12: Comparison between the original Visual CoT annotation and the refined FMLR annotation.

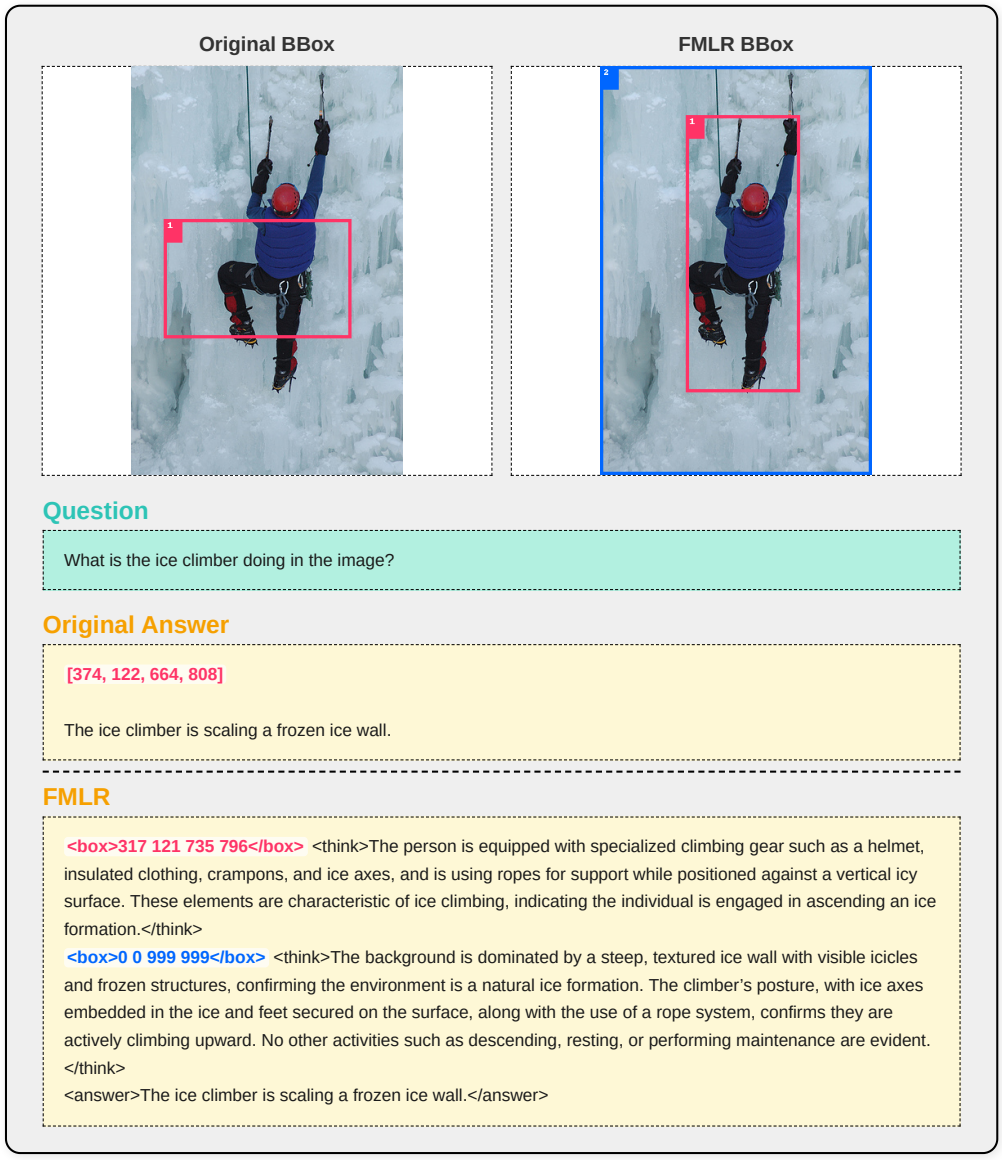


Figure 13: Comparison between the original Visual CoT annotation and the refined FMLR annotation.

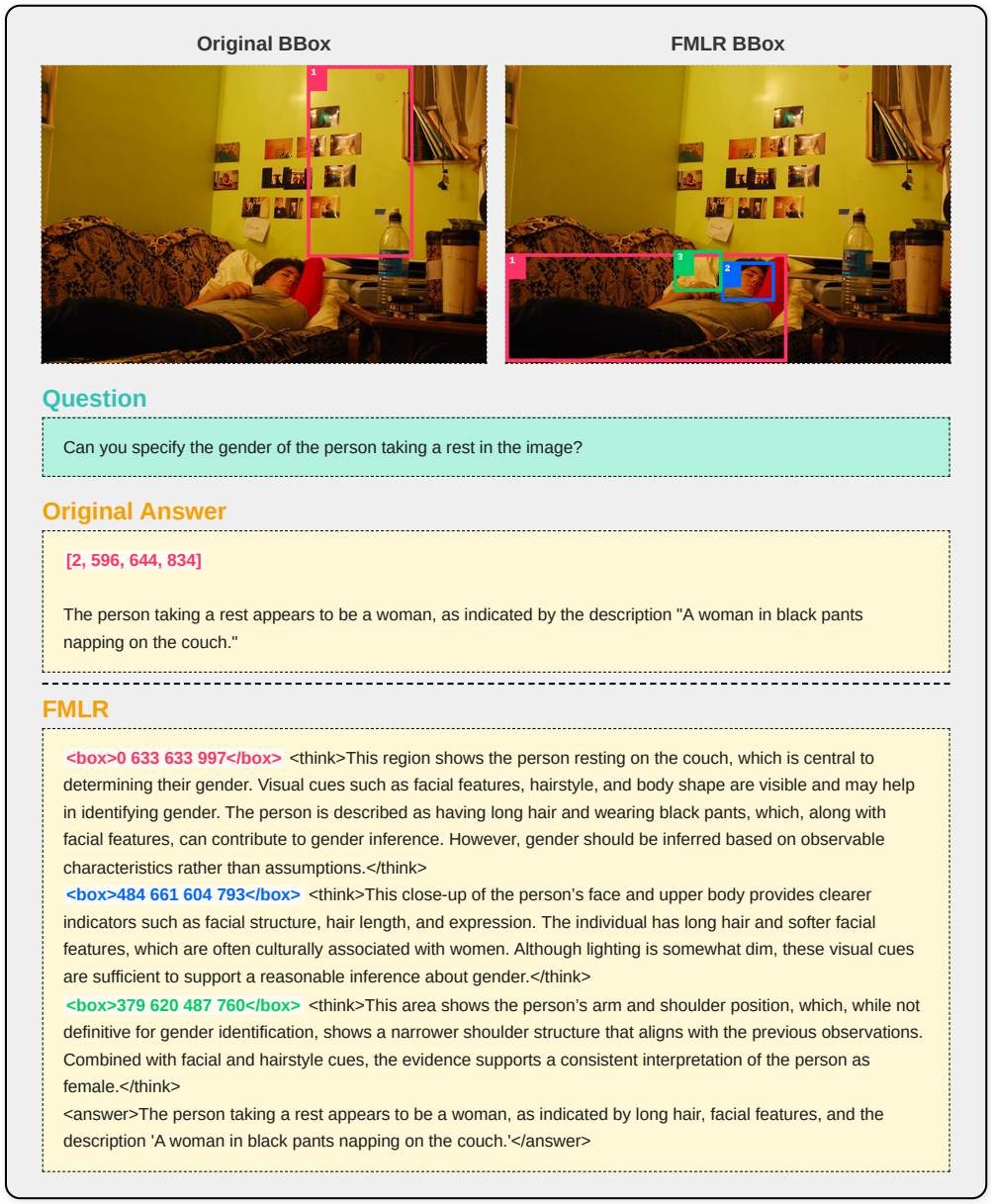


Figure 14: Comparison between the original Visual CoT annotation and the refined FMLR annotation.

E Broader Impacts

FMLR may support multimodal applications with shorter explicit reasoning traces. Its latent states are difficult to inspect, motivating careful evaluation and human oversight when outputs inform consequential decisions.