UNIFYING HETEROGENEOUS VISUAL REASONING IN EXPRESSIVE LATENT SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual reasoning increasingly relies on heterogeneous intermediate representations, including textual rationales, regions, scene graphs, segmentation, depth, and visual manipulations. Because these representations differ substantially in structure and modality, they are typically modeled with separate reasoning interfaces, which scale poorly as more reasoning forms are incorporated. This raises a fundamental question: *can heterogeneous forms of visual reasoning be unified in a single shared latent space?* We investigate this question through HVLR (Heterogeneous Visual Latent Reasoning), which unifies eight distinct reasoning forms within a common autoregressive latent interface. To reconcile heterogeneous supervision, we first render all reasoning representations into a shared pixel space and encode them with the same vision encoder. We then use a two-stage distillation procedure to translate these representations into autoregressive latent states and internalize them into a closed-book model that requires no explicit intermediate representation at inference time. We conduct a systematic analysis with HVLR and uncovers three main findings. First, heterogeneous reasoning forms can be effectively unified: combining all eight improves Qwen2.5-VL-7B from 64.66% to 69.97% average accuracy across nine benchmarks. Second, the shared latent space consistently benefits from reasoning diversity. Under a fixed 40K-example budget, increasing the number of reasoning forms from 1 to 3, 5, and 8 improves performance from 66.65% to 67.67%, 68.61%, and 69.97%, respectively, with positive trends across all nine benchmarks. Third, latent reasoning outperforms explicit intermediate decoding while reducing intermediate computation from roughly 177 generated tokens to only 8 latent steps, suggesting that continuous latent representations provide a more expressive and efficient medium for heterogeneous visual reasoning.

1 INTRODUCTION

Visual reasoning has been approached through increasingly diverse intermediate representations. Textual chains of thought support compositional reasoning (Zhang et al., 2023; Xu et al., 2025; Yao et al., 2026); regions and crops focus computation on task-relevant evidence (Shao et al., 2024; Wu & Xie, 2024); scene graphs expose object relations (Mitra et al., 2024; Batra et al., 2025); segmentation and depth provide fine-grained spatial structure (Lai et al., 2024; Chen et al., 2024; Liao et al., 2024); and visual manipulation enables reasoning directly over perceptual states (Hu et al., 2024; Li et al., 2025b; Zhang et al., 2025a). These representations capture complementary aspects of visual reasoning, yet their heterogeneous structures have largely led them to be modeled separately, with different systems specialized to different reasoning forms. This raises a fundamental question: *how can a single model jointly acquire and exploit heterogeneous forms of visual reasoning?* Modeling them explicitly within the same system requires dedicated prediction heads or decoders and scales poorly as the number of reasoning forms increases.

In this work, we ask: *can heterogeneous forms of visual reasoning be unified in a single shared latent space?* such unification poses two key challenges. First, heterogeneous reasoning signals occupy incompatible representation spaces and therefore cannot directly provide uniform supervision. We address this by rendering all reasoning forms into a common pixel space and encoding them with the same vision encoder, yielding a shared supervision interface without dedicated prediction heads. Second, these encoded auxiliary representations must be transferred into latent states that the model can generate and exploit autoregressively. We therefore adopt a two-stage distillation

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.

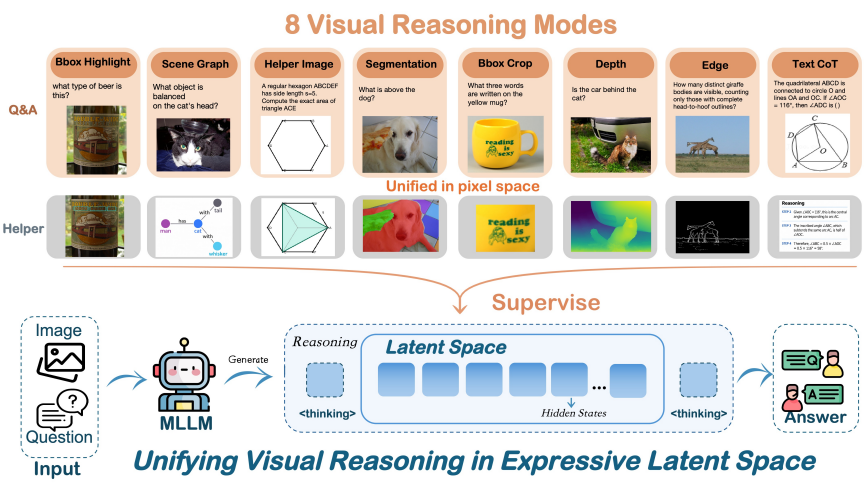

Figure 1: **Overview of HVLR.** *Top (training):* Eight heterogeneous reasoning forms are rendered as helper images and internalized into latent slots. *Bottom (inference):* Given only the image and question, our model generates latent states for reasoning and then the answer.

procedure. Stage 1 trains a latent teacher to transform auxiliary reasoning signals into task-relevant autoregressive latent states, while preventing answer prediction from directly accessing the auxiliary input. Stage 2 removes the auxiliary representation and distills these states into a closed-book student that regenerates them from the original image and question alone. Together, these designs lead to HVLR (Heterogeneous Visual Latent Reasoning), which unifies eight distinct reasoning forms within a common autoregressive latent interface and requires no explicit intermediate representation at inference time (Fig. 1).

Using HVLR, we systematically study whether heterogeneous visual reasoning can be effectively unified in a shared latent space. Our experiments yield three main findings. **First, heterogeneous reasoning forms can be jointly internalized into a single latent interface.** Unifying eight forms improves Qwen2.5-VL-7B from 64.66 to 69.97 average accuracy across nine benchmarks. **Second, greater reasoning diversity consistently improves the shared latent reasoner.** Under a fixed 40K-example supervision budget, increasing the number of reasoning forms from 1 to 3, 5, and 8 raises average performance from 66.65 to 67.67, 68.61, and 69.97, respectively, with positive trends across all nine benchmarks. **Third, the shared latent space provides a more expressive and efficient medium for intermediate reasoning than explicit decoding.** For reasoning forms that admit textual serialization, latent reasoning achieves higher accuracy while reducing intermediate computation from roughly 177 generated tokens to only 8 latent steps, suggesting that continuous latent representations can preserve useful reasoning information without constraining it to an explicit output format. These effects further generalize across model scales and an alternative latent-learning recipe.

Together, our results show that heterogeneous visual reasoning forms need not remain separated by dedicated interfaces. Instead, they can provide complementary supervision to a shared latent reasoning space, enabling a single model to integrate diverse reasoning capabilities through a common internal interface.

2 Method

2.1 Collection of Heterogeneous Visual Reasoning Data

Constructing unified supervision across heterogeneous reasoning forms is challenging because their native representations differ substantially in structure and modality. To provide a common supervision interface, we rasterize every reasoning representation into an image and process it with the same vision encoder. This allows linguistic, symbolic, spatial, and perceptual reasoning signals to supervise the same latent space without dedicated prediction heads.

Table 1: **Details of our data.** Our dataset covers eight reasoning forms, all rendered into images.

Reasoning Form	Data Source	Count	Visual Manipulation	Curation Steps
Text CoT	Mulberry (Yao et al., 2026) LLaVA-CoT (Xu et al., 2025) ScienceQA (Saikh et al., 2022)	5,000	Rationale rendered on canvas	Remove long or misleading rationales
Scene Graph	Visual Genome (Krishna et al., 2017)	5,000	Object–relation graph	Remove misleading graphs; rewrite unsuitable questions
BBox Crop	Visual CoT (Shao et al., 2024) ThinkMorph (Gu et al., 2026) ArxivQA (Li et al., 2024b)	5,000	Crop question-relevant regions	Filter by box number and area
BBox Highlight	Visual CoT (Shao et al., 2024) ThinkMorph (Gu et al., 2026) ArxivQA (Li et al., 2024b)	5,000	Highlight question-relevant regions	Filter by box number and area; remove misleading cases
Segmentation	Visual CoT (Shao et al., 2024)	5,000	Overlay SAM 2.1 masks (Ravi et al., 2025)	Retain task-relevant, non-misleading masks
Depth	Visual CoT (Shao et al., 2024) Visual Genome (Krishna et al., 2017)	5,000	Depth map (Yang et al., 2024)	Retain or construct depth-relevant questions
Edge	Visual CoT (Shao et al., 2024) Visual Genome (Krishna et al., 2017)	5,000	Edge map (Xie & Tu, 2015)	Retain or construct edge-relevant questions
Interleaved	V-Interaction-400K (Qiao et al., 2025) Zebra-CoT (Li et al., 2025a) ThinkMorph (Gu et al., 2026)	5,000	1–3 intermediate reasoning images	Remove duplicates, input copies, and misleading steps

We collect eight representative reasoning forms. **Text CoT** renders textual rationales on a blank canvas; **Scene Graph** visualizes objects and relations as node–edge diagrams; **BBox Crop** and **BBox Highlight** isolate or emphasize question-relevant regions; **Segmentation**, **Depth**, and **Edge** expose complementary dense visual structure; and **Interleaved** reasoning provides one to three intermediate images from multi-step visual reasoning traces. Each training example is represented as $(v, q, \{I_h^{(k)}\}_{k=1}^{N_h}, a)$, where v is the input image, q the question, $I_h^{(k)}$ the k-th rendered reasoning helper, and a the answer.

Data curation. Table 1 summarizes the data sources, visual transformations, and form-specific curation. We further apply a common quality pipeline across all forms: removing corrupted, low-resolution, duplicated, and evaluation-overlapping samples; filtering invalid visual transformations; and discarding examples in which the helper degrades answer correctness. For forms requiring targeted supervision, we additionally select or construct questions whose solution depends on the corresponding representation. After curation, the full corpus contains 1.31M examples. Unless otherwise specified, all experiments use the same balanced 40K subset, with 5K examples each form.

2.2 Unified Autoregressive Latent Interface

Our goal is to internalize all reasoning forms within a common latent interface. During training, each helper is compressed into a block of L latent slots. For N_h helpers, the sequence contains $M = N_h L$ slots $h_{1:M}$. Unlike vocabulary tokens, each latent slot directly reuses the preceding last-layer hidden state as the next input embedding, bypassing the language-model head:

$$h_\ell = f_\theta(c, h_{<\ell}), \qquad \ell = 1, \ldots, M, \qquad h_\ell \in \mathbb{R}^d, \tag{1}$$

where c denotes the preceding context and $d = 3584$ for the 7B backbone. The resulting slots therefore evolve autoregressively in the model's continuous hidden space.

Directly supervising these slots with encoded helper features is suboptimal because helper features and autoregressively generated states occupy different representation spaces. We therefore first translate heterogeneous helper supervision into autoregressive latent targets, and then distill these targets into a model that can reproduce them without access to the helpers.

2.3 Stage 1: Latent Teacher

Stage 1 converts each explicit helper representation into an autoregressively generated latent target. Each training step uses two forward passes (Fig. 2). In the first, the teacher observes the helper

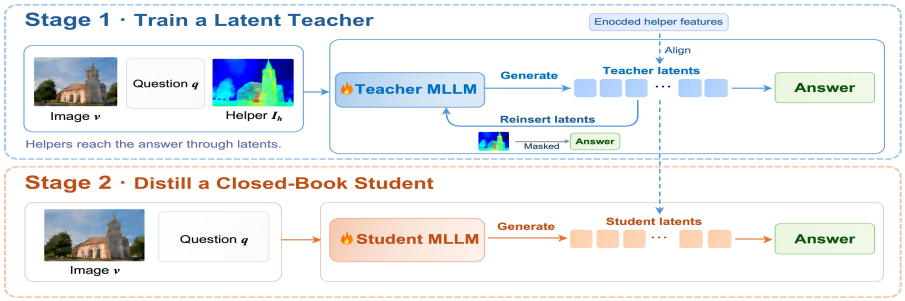

Figure 2: **Training overview.** Stage 1 trains a teacher that summarizes the helper images into latent slots. Stage 2 trains a student to generate these latent slots from the image and question alone, supervised by the features exported from the Stage-1 teacher.

images and autoregressively generates the latent slots $h_{1:M}$. In the second, these slots are reinserted into the sequence, while an attention mask prevents the answer positions from directly attending to the helpers. Consequently, information from the helpers must be mediated through the latent slots.

To preserve the helper content, we align the latent states with representations extracted from the same helpers by a frozen base model. Let $H^{(n)} \in \mathbb{R}^{M\times d}$ denote the hidden states at the latent positions at layer n, including the input embeddings for $n = 0$, and let $Z^{(n)}$ denote the corresponding helper representations, average-pooled to L vectors per helper. Stage 1 minimizes

$$\mathcal{L}_1 = \mathcal{L}_{\text{CE}} + \lambda\, \mathcal{L}_{\text{align}}(H, Z), \qquad \lambda = 2.0, \tag{2}$$

where

$$\mathcal{L}_{\text{CE}} = -\frac{1}{|\mathcal{S}|}\sum_{t\in\mathcal{S}} \log p_\theta(x_t \mid x_{<t}), \qquad \mathcal{L}_{\text{align}}(H, Z) = \frac{1}{|\mathcal{V}|}\sum_{(n,j)\in\mathcal{V}} \left[1 - \cos\left(H^{(n)}_{:,j}, Z^{(n)}_{:,j}\right)\right]. \tag{3}$$

Here, $\mathcal{S}$ denotes the supervised answer positions and $\mathcal{V}$ contains layer–channel pairs with non-zero columns. The alignment gradient reaches the model only through the latent slots (Appendix A.2). Stage 1 therefore produces targets that both preserve the helper information and lie in the model's autoregressive latent space.

2.4 Stage 2: Closed-Book Latent Internalization

Stage 2 removes the helpers and trains the model to reconstruct the Stage-1 latent states from the original input alone. A student f_ϕ, initialized from the base model, autoregressively generates

$$\tilde{h}_\ell = f_\phi(v, q, \tilde{h}_{<\ell}). \tag{4}$$

The frozen Stage-1 teacher provides precomputed targets $H^{\star(n)}$ at each latent position and layer. With $\tilde{H}^{(n)}$ denoting the corresponding student states, we optimize

$$\mathcal{L}_2 = \mathcal{L}_{\text{CE}} + \lambda\, \mathcal{L}_{\text{align}}\left(\tilde{H}, H^\star\right), \tag{5}$$

where

$$\mathcal{L}_{\text{CE}} = -\frac{1}{|\mathcal{S}|}\sum_{t\in\mathcal{S}} \log p_\phi(x_t \mid x_{<t}), \qquad \mathcal{L}_{\text{align}}(\tilde{H}, H^\star) = \frac{1}{|\mathcal{V}|}\sum_{(n,j)\in\mathcal{V}} \left[1 - \cos\left(\tilde{H}^{(n)}_{:,j}, H^{\star(n)}_{:,j}\right)\right]. \tag{6}$$

At inference, the student receives only (v, q), generates the latent reasoning states, and then decodes the answer. No helper image, rationale text, or dedicated reasoning decoder is required; inference therefore incurs only the fixed latent-reasoning overhead (Appendix A.7).

3 Experiments Setup

Implementation Details. We adopt Qwen2.5-VL-7B-Instruct (Bai et al., 2025b) as the base model and use L=8 latent slots throughout. Training uses the 40K subset, AdamW, a linear learning-rate

Table 2: **Main results** on nine benchmarks. **Bold** = best per column among same-protocol rows. Frontier-model rows (gray) are *reported* numbers taken from published papers.

Model	V*	HR4K	HR8K	MME-RW	BLINK	POPE	RWQA	CV-B	QSp++	Avg.
base (Qwen2.5-VL-7B)	75.92	68.88	65.25	45.49	55.65	86.58	69.15	73.40	41.58	64.66
Latent Reasoning										
Monet-7B (Wang et al., 2026)	80.85	70.38	68.88	47.06	41.56	87.75	61.05	51.64	31.68	60.09
LVR-7B (Li et al., 2026a)	81.18	70.75	66.12	49.56	53.45	84.76	68.10	75.09	24.75	63.75
CoVT-7B (Qin et al., 2025)	78.53	71.00	69.50	50.34	56.35	84.58	70.07	78.75	53.47	68.07
SkiLa-7B (Tong et al., 2025)	**82.72**	71.12	64.50	46.74	54.87	87.39	69.93	76.19	40.59	66.01
Mirage-VSP (Yang et al., 2026b)	79.06	70.12	66.88	48.31	56.78	85.29	69.93	75.07	42.57	66.00
Open-source VLM										
InternVL3-8B (Zhu et al., 2025a)	70.16	69.00	59.00	47.68	55.18	89.00	70.07	**80.98**	53.47	66.06
LLaVA-OneVision-7B (Li et al., 2024a)	72.25	65.12	58.13	48.41	50.55	88.20	70.07	76.32	51.49	64.50
Tool Using										
DeepEyes-7B (Zheng et al., 2026)	82.20	**73.62**	**70.88**	52.42	44.29	87.74	69.41	78.59	44.55	67.08
PixelReasoner-7B (Su et al., 2026)	78.53	70.62	64.50	45.75	45.08	85.97	69.67	79.09	45.54	64.97
Frontier Close-Source Models										
GPT-5‡	74.87	74.6	71.62	57.0	71.0	—	82.8	—	63.37	—
GPT-4o‡	62.8	68.0	63.9	51.6	60.0	86.9	75.4	79.4	61.06	–
Gemini-2.5-Pro‡	79.2	87.3	83.62	58.30	70.6	—	78.0	86.3	40.59	—
Ours	81.70	71.75	**70.88**	**54.30**	**56.85**	**89.19**	**72.03**	78.53	**54.46**	**69.97**
Δ vs base	+5.78	+2.87	+5.63	+8.81	+1.20	+2.61	+2.88	+5.13	+12.88	+5.31

schedule with 10 warmup steps, an effective batch size of 128, and alignment weight λ=2.0 at all layers. For the $K \leq 5$ diversity study in Sec. 4.1, all 25 models use one epoch per stage with a learning rate of 1×10^{-5}; the K=8 result uses the default configuration. Full training details are provided in Appendix A.

Evaluation. We evaluate on nine benchmarks: V* (Wu & Xie, 2024), HR-Bench 4K/8K (Wang et al., 2025a), MME-RealWorld-Lite (Zhang et al., 2025b), RealWorldQA (AI, 2024), BLINK (Fu et al., 2024a), POPE (Li et al., 2023), CV-Bench (Tong et al., 2024), and QSpatial++ (Liao et al., 2024). All methods are evaluated using VLMEvalKit (Duan et al., 2024) with temperature 0. *Avg.* denotes the average score, the unweighted macro-mean across all nine benchmarks. For CV-Bench, we report the mean over its 2D and 3D subsets, with the 2D score averaged across its two image sources. For QSpatial++, we parse the predicted distance and unit, convert the value to meters, and count predictions within a factor of two of the ground truth ($\delta \leq 2$) as correct.

4 Unifying Heterogeneous Visual Reasoning in Latent Space

Our experiments yield three findings. First, under a fixed budget the latent reasoner becomes stronger as it internalizes more reasoning forms (Sec. 4.1). Second, keeping reasoning in the latent space is more accurate and far cheaper than decoding it as text (Sec. 4.2). Third, these gains transfer across backbones and training recipes (Sec. 4.4).

4.1 More Forms, Stronger Model: Capability Scales with Latent Diversity

Finding ❶. *The capability of latent reasoning scales with the* diversity *of the content the space must host: packing more forms of thought in one latent space yields a stronger model.*

Design. We fix the budget at 40K and vary *only* the number of forms it spans: K=1 gives the whole budget to a single form (8 models, one per form); K=3 and K=5 each draw 8 *random* combinations with 40K examples in total; K=8 is the model of Table 2. Table 10 in the appendix lists the form combination and all nine scores of every model.

Capability rises with more forms. The tier-mean average score rises strictly with K, as does nearly every benchmark column (Table 8). In the average-score ranking (Fig. 3b), the tiers overlap only slightly, and all 25 ablations outperform base, including the weakest single form (scene graph, +0.14 over base). Adding forms at a fixed budget reduces the examples per form, yet performance still improves, indicating that the gain stems more from the diversity of reasoning forms than from

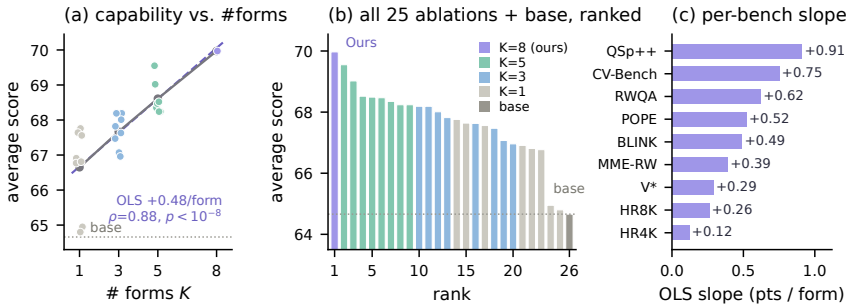

Figure 3: **Modality-count ablation**. **(a)** average score vs. number of forms K: tier means (solid) rise monotonically; OLS $+0.483$/form, Spearman ρ=0.88. **(b)** All 25 ablations + base ranked by average score: ours ranks #1, the eight random K=5 subsets occupy #2–9, and base ranks last. **(c)** Refitting the slope on each benchmark separately: all nine are positive.

per-form data volume. More forms also reduce variance across form choices: the within-tier std decreases from 1.09 at K=1 to 0.42 at K=5 (Table 8).

The upward trend is statistically robust. We test the trend in four ways (Table 9 in the appendix). A least-squares fit of the average score on K gives $+0.483 \pm 0.079$ points per added form (R^2=62.2%, p=2.8×10^{-6}). Rank correlations, which drop the linearity assumption, agree (Spearman ρ=0.879, Kendall τ=0.765, both p<10^{-5}). A Jonckheere–Terpstra test for the ordered groups $K=1 < 3 < 5 < 8$ finds 205 of 216 cross-tier pairs in the predicted order (z=4.79, p=8.2×10^{-7}). In exact pairwise Mann–Whitney tests every K=5 model outscores every K=3 and every K=1 model (64/64 pairs, p=7.8×10^{-5}, the smallest p attainable at these sample sizes).

The count of forms explains the trend rather than their identity. We next show that the gain comes from the diversity of forms rather than from the identity of any particular form: larger mixtures might otherwise score higher merely because they are more likely to contain one strong form. We compare the count model above with an *identity* model that gives each of the eight forms its own coefficient and nests the count model. Moving from the count model to the identity model raises R^2 from 62.2% to 72.4%, but the nested F-test finds this gain no larger than seven extra parameters would achieve by chance ($F(7, 16)$=0.85, p=0.57; Table 9). The identity model's coefficients, the marginal gain of adding each form to a mixture, are all non-negative and do not follow the single-form ranking: bbox-crop, the strongest form alone, ranks sixth, while the interleaved helper, 21st of 25 alone, contributes the most.

Robust across benchmarks, and not yet saturated. The trend holds on every benchmark: fitting each benchmark's score against K separately gives a positive slope on all nine (Fig. 3c), and recomputing the average with any single benchmark removed keeps the slope between 0.43 and 0.53 points per form. The gain also shows no sign of saturating up to K=8: if each added form helped less than the previous one, a fit in $\log K$ would describe the data better than a straight line in K, yet it fits worse (R^2=58.3% against 62.2%).

4.2 Latent Reasoning Is More Effective and More Efficient than Decoded Text

Finding ❷. *Keeping intermediate thought in the latent space dominates decoding it into text: carried in vectors, the same reasoning reaches higher accuracy at a fraction of the inference cost.*

Text CoT and scene graphs can be written out faithfully as text, which lets us test whether a form is better internalized in the latent space or emitted as text.

Design. From the 40K eight-form corpus of Sec. 4.3 we train three variants in which the text-CoT portion, the scene-graph portion, or both are supervised as *explicit decoded text*. Every other form stays latent, and the data, budget, backbone and schedule are held fixed.

Externalizing a form changes its training objective. Write D for the set of forms re-expressed as text. An example whose helper form $f \notin D$ is trained exactly as in Sec. 2. For $f \in D$ the helper is

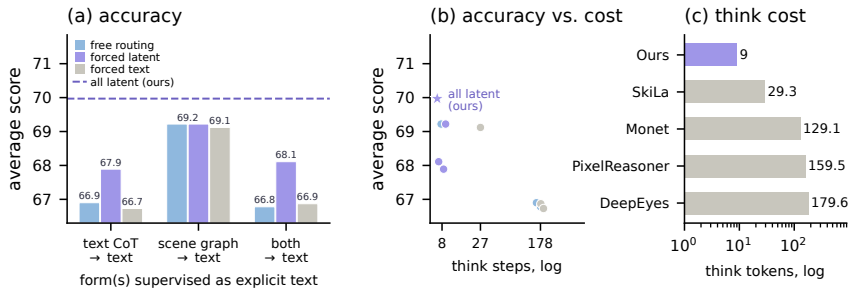

Figure 4: **Latent versus explicit decoded-text reasoning.** Three variants supervise the text-CoT portion, the scene-graph portion, or both as decoded text while all other forms stay latent; each is evaluated with free routing, forced latent or forced text decoding. The dashed line is our all-latent model. **(a)** Average score over the nine benchmarks. **(b)** The same runs against think steps. **(c)** Think tokens per question across methods, measured as in Fig. 5.

serialized into a token string $r = (r_1, \ldots, r_T)$ that the model must write before the answer; such an example has no thinking markers and no latent slots, so its alignment term is zero and r joins the answer in a single token-averaged cross-entropy. In stage $s \in \{1, 2\}$ the objective is

$$\mathcal{L}^{(s)} = -\frac{1}{|\mathcal{S}_f|} \sum_{t \in \mathcal{S}_f} \log p(x_t \mid x_{<t}) + \mathbf{1}[f \notin D]\, \mathcal{L}^{(s)}_{\text{align}},$$

$$\mathcal{S}_f = \begin{cases} \mathcal{S} \text{ of Eq. (3)} & f \notin D, \\ \text{all response tokens: } r, \text{ the answer and the template tokens} & f \in D, \end{cases}$$

where p is the model of stage s and $\mathcal{L}^{(s)}_{\text{align}}$ is the λ-weighted alignment term of Eq. (2) for $s{=}1$ and of Eq. (5) for $s{=}2$. One model thus learns both paths (the latent block from the forms outside D, the text path from those inside it), which makes its routing between them observable at inference.

At inference every variant receives the same prompt and differs only in how its response begins:

$$\underbrace{v,\ q}_{\text{prompt}} \ \rightarrow \ \begin{cases} \texttt{<thinking>}\ \tilde{h}_1, \ldots, \tilde{h}_L\ \texttt{</thinking>}\ \ a & \textit{latent}, \text{ LM head bypassed} \\ r_1, \ldots, r_T\ \ a & \textit{text}, \text{ via the LM head} \end{cases}$$

The first token of the response selects the continuation: the opening marker starts the latent block, and any other token starts the text.

Results. No variant reaches the all-latent model under any inference mode (Fig. 4a; full scores in Appendix A.6), and externalizing the scene graph incurs the smallest drop. For the text-CoT and both-form variants, forced latent outperforms forced text on the same weights by 1.16 and 1.24 points of average score, respectively; under free routing both of these models nonetheless select the weaker text form (the text-CoT model on every question) and score below their own forced-latent runs. The scene-graph model always selects the latent path.

Efficiency. Inference cost follows the same pattern: the latent block takes 8 think steps per question against roughly 177 for a decoded text CoT (Fig. 4b). To compare with other methods, we measured token counts, latency and FLOPs for eight methods (Fig. 5; Appendix A.7). Our method is the fastest and the most accurate (Fig. 5c) and spends under a third of the decode FLOPs of the base model answering directly. The sketch- and tool-based methods emit more think tokens (Fig. 4c) and the tool-using methods additionally repeat prefill over a context that grows with every round.

4.3 Unification Leads to Better Performance

Table 2 evaluates three model families. Our model improves on its base and attains the highest average score among all same-protocol rows, +1.9 above the best latent-reasoning baseline (CoVT-7B). It also exceeds the best tool-using agent, DeepEyes, with a single closed-book pass instead of a multi-pass zoom-in loop.

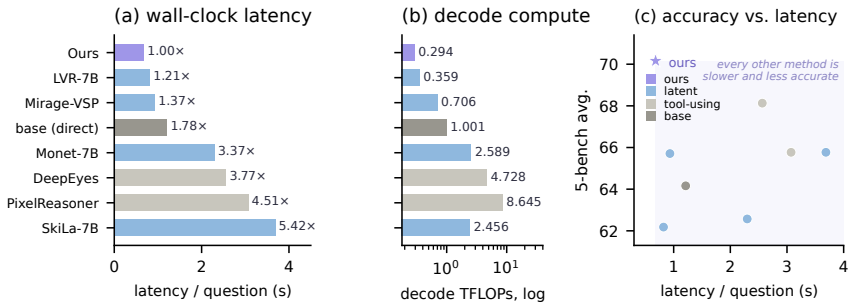

Figure 5: **Inference efficiency** on V*, HR-Bench 4K/8K, RealWorldQA and QSpatial++. **(a)** Latency per question, with the ratio to ours. **(b)** Decode FLOPs computed from token counts; all methods share one backbone. **(c)** Mean accuracy over the five benchmarks against latency; the shaded region marks methods that are both slower and less accurate than ours.

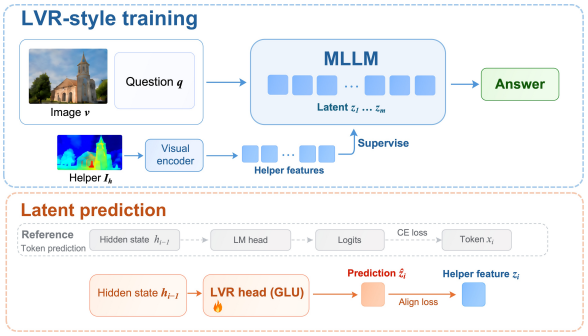

Figure 6: **LVR-style training pipeline.** *Top (Overview):* Helper image is encoded by visual encoder as ground truth to supervise latent. *Bottom (Details):* latent is predicted from preceding hidden state like a text token.

Table 3: **Cross-backbone transfer.** Other backbones trained with the same data and training pipeline also improve over the base models.

Qwen2.5-VL-3B

	CV-Bench	MME-RW	QSpatial++	POPE
base	59.22	42.31	52.48	86.31
ours	69.91	48.93	58.42	87.96
Δ	+10.69	+6.62	+5.94	+1.65

Qwen3-VL-8B

	QSpatial++	MME-RW	V*	HR8K
base	51.49	49.82	80.10	70.25
ours	64.36	55.97	83.77	71.88
Δ	+12.87	+6.15	+3.67	+1.63

4.4 The Gains Transfer across Backbones and Recipes

Finding ❸. *The benefit of heterogeneous latent supervision transfers across model scales, backbones, and training recipes.*

The same data and training stages also improve a smaller (Qwen2.5-VL-3B) and a newer (Qwen3-VL-8B) backbone (Table 3), as well as a modified LVR-style recipe (Table 4).

We adapt LVR (Li et al., 2026a) into a pipeline that shares only the data and the latent interface with ours (Fig. 6). Training is a *single* teacher-forced stage without distillation: the rendered helper's visual features fill the latent slots as inputs, and a trainable prediction head predicts each slot from the hidden state at the preceding position, trained with an alignment loss, the answer cross-entropy and a contrastive term that keeps the latents sample-specific. At inference the helper is absent and the head generates the latent slots autoregressively. Trained on 500K examples of our corpus, this pipeline gains 3.47 points of average score over its base and surpasses the publicly released LVR checkpoint by 5.03 points, most notably on QSpatial++ (Table 4).

5 Ablation Study

Both training stages are necessary. We ablate the two components of the training procedure (Table 5). Removing the AR-generated teacher and aligning the closed-book student directly to the pooled helper states Z of Stage 1 (*w/o distillation*) lowers accuracy consistently and reduces the average score by 1.95 points. This corroborates the analysis of Sec. 2: alignment targets must lie in the same autoregressively generated latent space as the student's outputs. Retaining the distillation

Table 4: **Cross-recipe transfer.** A modified LVR-style recipe trained on our dataset, against its base and against the publicly released LVR checkpoint.

	V*	MME-RW	BLINK	POPE	RWQA	CV-Bench	QSpatial++	**Avg.**
base (Qwen2.5-VL-7B)	75.92	45.49	55.65	86.58	69.15	73.40	41.58	63.97
LVR-7B (Li et al., 2026a)	81.18	**49.56**	53.45	84.76	68.10	75.09	24.75	62.41
LVR-style + our data	**82.72**	49.40	**56.13**	**87.21**	**70.85**	**78.26**	**47.52**	**67.44**
Δ vs. base	+6.80	+3.91	+0.48	+0.63	+1.70	+4.86	+5.94	+3.47
Δ vs. LVR-7B	+1.54	−0.16	+2.68	+2.45	+2.75	+3.17	+22.77	+5.03

Table 5: **Ablation of the two-stage procedure.** *w/o distillation* aligns the student directly to the encoded helper states Z, removing the Stage 1 teacher; *w/o Stage 1* retains the distillation objective but uses the untrained base model as the teacher.

Variant	MME-RW	BLINK	RWQA	CV-Bench	HR8K	**Avg.**
Ours (two-stage)	**54.30**	**56.85**	**72.03**	**78.53**	**70.88**	**69.97**
w/o distillation	50.50	53.13	70.07	75.56	67.50	68.02
w/o Stage-1 (base teacher)	52.42	52.13	69.93	76.49	69.50	68.50

objective with the untrained base model as the teacher (*w/o Stage 1*) also degrades performance (average score -1.47), indicating that an effective teacher must first internalize the helper signal through Stage-1 training.

Table 6: **Ablation of the latent-block size.** Relative to the default $L{=}8$, both a smaller ($L{=}4$) and a larger ($L{=}12$) block reduce accuracy on the benchmarks shown.

Slots	V*	HR4K	HR8K	MME-RW	BLINK	POPE	RWQA	CV-Bench
$L{=}4$	80.10	71.38	69.62	53.05	53.60	88.47	68.76	75.08
$L{=}8$ (default)	**81.70**	**71.75**	**70.88**	**54.30**	**56.85**	**89.19**	**72.03**	**78.53**
$L{=}12$	79.58	70.62	68.25	50.65	51.50	89.08	68.76	68.92

The default block size $L{=}8$ performs best. Reducing the block to $L{=}4$ lowers accuracy on every benchmark shown, consistent with a block too small to retain the helper content. Enlarging it to $L{=}12$ degrades performance further; we tentatively attribute this to the weaker compression constraint, which spreads the alignment signal over more slots and thereby dilutes it (Table 6).

6 Related Work

Visual reasoning methods use diverse explicit intermediate representations, from textual rationales (Xu et al., 2025; Yao et al., 2026) and region crops (Shao et al., 2024; Wu & Xie, 2024) to scene graphs (Mitra et al., 2024), sketches (Hu et al., 2024) and tool-based interaction (Zheng et al., 2026). Latent reasoning replaces such rationales with continuous hidden states (Hao et al., 2024; Shen et al., 2025), and its multimodal variants each internalize a single predefined form (Li et al., 2026a; Yang et al., 2026b; Wang et al., 2026; Qin et al., 2025). We instead unify heterogeneous forms in one shared latent space; We refer to Appendix B for a full related work.

7 Conclusion

We study whether heterogeneous forms of visual reasoning can be unified within a single shared latent space. HVLR internalizes eight categorically different reasoning forms into one compact autoregressive latent interface that requires no explicit intermediate representation at inference. The resulting latent space proves highly expressive: it accommodates all forms without any of them hurting the others, grows stronger as more forms are added under a fixed budget, and outperforms explicit decoding at a fraction of the cost. These findings suggest that a shared latent space can serve as a common medium for diverse visual reasoning.

AI USE STATEMENT

We used vision–language models for dataset construction and curation: Qwen3-VL-235B generated or rewrote questions for depth maps, edge maps, and scene graphs; Qwen3-VL-32B selected among segmentation, depth, and edge helpers; and Qwen2.5-VL-7B supported filtering by comparing answers with and without helper images. The data sources and filtering procedures are described in Table 1 and Appendix A.1.

ETHICS STATEMENT

All authors have read and adhere to the ICLR Code of Ethics. This work involves no human subjects and no new collection of data from people. All training images, questions and answers come from publicly released research datasets (Table 1), used for research purposes under their respective licenses, and the helper images are rendered with publicly available models; part of the questions were generated or rewritten by Qwen3-VL models (Appendix A.1) and may contain errors. The method improves general-purpose visual reasoning and is not designed for any specific high-risk application.

REPRODUCIBILITY STATEMENT

Section 2 specifies the latent interface and two-stage training procedure, and Table 1 identifies the data sources and helper representations. Appendix A.1 describes data construction and filtering. Section 3 and Appendix A.2 document optimization settings, training schedules, the random seed, visual-token budgets, target extraction, and attention masking. The evaluation benchmarks and metrics are specified in Section 3 and Appendix A.3. Appendices A.5 and A.6 provide detailed ablation and routing results, and Appendix A.7 describes the token, FLOPs, and latency measurements. Appendix A.4 identifies the sources and protocol differences for externally reported frontier-model results.

REFERENCES

X AI. Grok-1.5 vision preview. *https://x. ai/blog/grok-1.5 v*, 2024.

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*, 2025a.

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. *arXiv preprint arXiv:2502.13923*, 2025b.

Hunar Batra, Haoqin Tu, Hardy Chen, Yuanze Lin, Cihang Xie, and Ronald Clark. Spatialthinker: Reinforcing 3d reasoning in multimodal llms via spatial rewards. *arXiv preprint arXiv:2511.07403*, 2025.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14455–14465, 2024.

Zeren Chen, Xiaoya Lu, Zhijie Zheng, Pengrui Li, Lehan He, Yijin Zhou, Jing Shao, Bohan Zhuang, and Lu Sheng. Geometrically-constrained agent for spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38689–38699, 2026.

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

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.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023.

Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*, 2024.

Hao Ding, Zhichuan Yang, Weijie Ge, Ziqin Gao, Chaoyi Lu, and Lei Zhao. Tikart: Stabilizing aperture-guided fine-grained visual reasoning with reinforcement learning. *arXiv preprint arXiv:2602.14482*, 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, 2024a.

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, 2024b.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *Advances in Neural Information Processing Systems*, 38: 41340–41391, 2026.

Jiawei Gu, Yunzhuo Hao, Huichen Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 141405–141447, 2026.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

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. *arXiv preprint arXiv:2412.06769*, 2024.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 139348–139379, 2024.

Ailin Huang, Chengyuan Yao, Chunrui Han, Fanqi Wan, Hangyu Guo, Haoran Lv, Hongyu Zhou, Jia Wang, Jian Zhou, Jianjian Sun, et al. Step3-vl-10b technical report. *arXiv preprint arXiv:2601.09668*, 2026.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589. IEEE, 2024.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. In *International Conference on Learning Representations*, volume 2026, pp. 76722–76746, 2026.

Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, et al. Zebra-cot: A dataset for interleaved vision language reasoning. *arXiv preprint arXiv:2507.16746*, 2025a.

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.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025b.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14369–14387, 2024b.

Qingxiao Li, Lifeng Xu, QingLi Wang, Yudong Bai, Mingwei Ou, Shu Hu, and Nan Xu. S1-vl: Scientific multimodal reasoning model with thinking-with-images. *arXiv preprint arXiv:2604.21409*, 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.

Huizhi Liang, Yichao Shen, Yu Deng, Sicheng Xu, Zhiyuan Feng, Tong Zhang, Yaobo Liang, and Jiaolong Yang. Hispatial: Taming hierarchical 3d spatial understanding in vision-language models. *arXiv preprint arXiv:2603.25411*, 2026.

Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17028–17047, 2024.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14420–14431. IEEE, 2024.

OpenCompass Contributors. OpenVLM leaderboard. `https://huggingface.co/spaces/opencompass/open_vlm_leaderboard`, 2024. Hugging Face Space; VLMEvalKit evaluation results collection. Accessed 2026-07.

Runqi Qiao, Qiuna Tan, Minghan Yang, Guanting Dong, Peiqing Yang, Shiqiang Lang, Enhui Wan, Xiaowan Wang, Yida Xu, Lan Yang, et al. V-thinker: Interactive thinking with images. *arXiv preprint arXiv:2511.04460*, 2025.

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.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: a novel resource for question answering on scholarly articles: T. saikh et al. *International Journal on Digital Libraries*, 23(3):289–301, 2022.

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.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 677–693, 2025.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. *Advances in Neural Information Processing Systems*, 38:8222–8251, 2026.

DiJia Su, Hanlin Zhu, Yingchen Xu, Jiantao Jiao, Yuandong Tian, and Qinqing Zheng. Token assorted: Mixing latent and text tokens for improved language model reasoning. *arXiv preprint arXiv:2502.03275*, 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.

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, 2026.

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, 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.

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.

Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 1395–1403, 2015.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2087–2098. IEEE, 2025.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024.

Wenhao Yang, Yu Xia, Jinlong Huang, Shiyin Lu, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, Yuanyu Wan, and Lijun Zhang. Deep but reliable: Advancing multi-turn reasoning for thinking with images. *arXiv preprint arXiv:2512.17306*, 2025.

Wenhao Yang, Yu Xia, Jinlong Huang, Shiyin Lu, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, Yuchen Zhou, Xiaobo Xia, et al. Walk the talk: Bridging the reasoning-action gap for thinking with images via multimodal agentic policy optimization. *arXiv preprint arXiv:2604.06777*, 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.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *Advances in Neural Information Processing Systems*, 38:29918–29952, 2026.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025a.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, volume 2025, pp. 89655–89701, 2025b.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, and Xin Wang. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *Advances in Neural Information Processing Systems*, 38:168990–169012, 2026.

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

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a.

Rui-Jie Zhu, Tianhao Peng, Tianhao Cheng, Xingwei Qu, Jinfa Huang, Dawei Zhu, Hao Wang, Kaiwen Xue, Xuanliang Zhang, Yong Shan, et al. A survey on latent reasoning. *arXiv preprint arXiv:2507.06203*, 2025b.

A Appendix

A.1 Data cleaning

The raw pool is cleaned in six steps. (i) At construction, images must decode and answers must parse; bounding-box samples keep at most three boxes covering 1–80% of the image (at most 25% for the first highlighted box), text-CoT rationales are capped at three rendered pages, and duplicates and questions from our held-out split are removed. (ii) Qwen2.5-VL-7B answers each sample with and without its helper, and samples on which the helper reliably turns a correct answer into a wrong one are removed. (iii) Qwen3-VL-32B splits the original visual-representation samples, which pair each question with a depth, an edge and a segmentation map, keeping only the most useful map and discarding samples for which none helps. (iv) Samples without a helper image and samples carried over from an earlier corpus version are removed. (v) Qwen3-VL-235B writes a map-dependent question for new Visual Genome depth and edge maps. (vi) Qwen3-VL-235B keeps scene graphs that contain the information their answer requires and replaces the other questions with two graph-dependent ones.

A.2 Training and Implementation Details

This section gives the full configuration used for every model reported in the paper, beyond the summary in Sec. 3.

Optimization. We use fused AdamW (β_1=0.9, β_2=0.999, ϵ=10^{-8}) with weight decay 0.01, gradient clipping at 1.0, and bf16 precision. The learning-rate schedule is linear with 10 warmup steps, and each stage is a separate run whose learning rate decays to zero at its final step. Training uses DeepSpeed ZeRO-2 without optimizer offload, a maximum sequence length of 4096, and random seed 42. The effective batch size is 128 in every run; its split into per-device batch size, gradient accumulation and number of GPUs depends on the hardware (e.g., $1\times16\times8$). For the main model, Stage 1 trains for two epochs (626 steps) at a peak learning rate of 1×10^{-5} and Stage 2 for one epoch (313 steps) at 5×10^{-6}; the reported model is the final Stage-2 checkpoint. The modality-count ablations with $K\leq5$ train each stage for one epoch at 1×10^{-5}.

The alignment objective. Both stages evaluate the alignment loss of Eq. (3) at every layer ($N+1$=29 hidden-state entries, including the input embedding) in fp32, with the cosine clipped to $[-1, 1]$; a (layer, channel) pair enters $\mathcal{V}$ when both of its columns have norm above 10^{-6}. All M slots of an example enter each cosine jointly, so the loss is normalized by $|\mathcal{V}|$ rather than by M, and is then averaged over the examples in the batch. Each training step runs two forward passes with gradients enabled. The first generates the slots with a key–value cache; the vector fed back as the next slot's input is detached, while the generated $h_{1:M}$ keep their computation graph. The second writes $h_{1:M}$ into the slot positions and evaluates both losses. In Stage 1 the alignment term of Eq. (2) is back-propagated only through the generated slots: the implemented objective is

$$\mathcal{L}_1 = \mathcal{L}_{\mathrm{CE}} + \mu \sum_{\ell=1}^{M} \Big\langle \mathrm{sg}\big[\nabla_{h_\ell}\, \lambda\, \mathcal{L}_{\mathrm{align}}(H, Z)\big],\ h_\ell \Big\rangle, \qquad \mu{=}2.0,$$

where $\mathcal{L}_{\mathrm{CE}}$ is Eq. (3) and $\mathrm{sg}[\cdot]$ is the stop-gradient operator. Its parameter gradient in the second term is $\mu\lambda \sum_\ell \left(\partial\mathcal{L}_{\mathrm{align}}/\partial h_\ell\right)^\top \partial h_\ell/\partial\theta$, with the first factor taken through the second pass and the second through the first; the effective weight is therefore $\mu\lambda$=4.0, and the direct gradient of the alignment term through the second pass is discarded. Stage 2 applies no such routing: it minimizes $\mathcal{L}_2$ of Eq. (5) with coefficient λ=2.0, and the emphasis factor is not used. The training script also exposes a cross-entropy emphasis factor; it is inactive in our text-free pipeline, because the token set it would emphasize is empty once the explicit text stream is removed. It is therefore not a hyperparameter of the reported models.

Visual token budgets. Text-CoT pages and scene graphs are rendered on a fixed 768×768 canvas (729 visual tokens); the other forms keep the resolution of their source (for crops, the cropped region), so under the Stage-1 budget a single helper occupies between about 4 and 1,250 visual tokens, with a median of about 250. Stage 1 caps the whole sequence at 1500 visual tokens with at most 1280 tokens for any single image; Stage 2, which no longer carries helper images, caps the sequence at 2000 visual tokens.

Teacher targets. The two stages align to different targets, both precomputed offline and kept frozen. The Stage-1 targets Z of Eq. (3) use adaptive average pooling along the token axis, so each helper yields 8 vectors independently of its resolution; at layer 0 they are pooled visual input embeddings. They are computed under a budget of 2000 visual tokens per sequence (1280 per image) rather than the 1500 per sequence used in Stage-1 training; the pooling makes the shapes agree. The Stage-2 targets $H^\star$ stack the hidden states of all layers at the slots generated by the frozen Stage-1 model with the helper images in context (shape $[\,N{+}1,\ N_h{\cdot}8,\ 3584\,]$ for N_h helper images), each slot attending to all preceding tokens, under a budget of 1000 visual tokens per sequence (500 per image).

Attention masking. In the forward pass that computes the losses, Stage 1 applies a 4D attention mask that hides the visual tokens of each helper from every position outside that helper's block of slots; all other visibility is causal. Answer tokens, thinking markers and the slots of other blocks therefore cannot attend to a helper, so helper information can reach the answer only through the latent slots. In the first pass, which generates the slots, each slot attends to all preceding tokens, so block k attends to every helper image that precedes it, not only to $I_h^{(k)}$. Stage 2 uses no such mask, since the helper images are removed from the input altogether. Our implementation also supports an optional masking curriculum; it is disabled for all Qwen2.5-VL-7B models reported here.

Evaluation. All open models are evaluated closed-book through a single VLMEvalKit (Duan et al., 2024) pipeline with identical prompts, one grader, and temperature 0.

A.3 Evaluation suite

Table 7 lists the nine benchmarks of Sec. 3 with the capability each probes and the metric it is scored with.

Table 7: **Evaluation suite.** Nine benchmarks, their capability, and the metric each is scored with. Most are multiple-choice and scored by accuracy; POPE is scored by F1 over its binary yes/no probes, and QSpatial++ requires an open-ended metric distance and is scored by a tolerance-based success rate.

Benchmark	Capability	Metric
V* (Wu & Xie, 2024)	fine-grained visual search	multiple-choice accuracy
HR-Bench 4K (Wang et al., 2025a)	high-resolution perception	multiple-choice accuracy
HR-Bench 8K (Wang et al., 2025a)	high-resolution perception	multiple-choice accuracy
MME-RealWorld-Lite (Zhang et al., 2025b)	real-world perception	multiple-choice accuracy
RealWorldQA (AI, 2024)	real-world QA	accuracy
BLINK (Fu et al., 2024a)	multi-skill perception	multiple-choice accuracy (val)
POPE (Li et al., 2023)	object hallucination	F1 over yes/no probes
CV-Bench (Tong et al., 2024)	2D/3D understanding	accuracy, mean of the 2D and 3D subsets
QSpatial++ (Liao et al., 2024)	quantitative spatial reasoning	success rate, $\delta \leq 2$ tolerance

A.4 Sources for frontier-model numbers

The frontier rows of Table 2 are *reported* numbers; none were produced by our pipeline, and protocols differ across sources (image-resolution caps, prompting, and answer matching vary). We list each cell's source; where published numbers conflict we cite the one from the most authoritative source (benchmark authors $>$ large-lab technical reports $>$ third-party papers) and note alternatives. "—" in the table means no published number was found after an extensive search. For the quantitative-spatial column we compare only on the QSpatial++ split: each frontier cell below is the source's QSpatial++ number, never a full-benchmark average. Because the nine cells of a frontier row come from nine different evaluations (different model snapshots, prompts, image-resolution caps and answer matching), we do not report an average score for these rows: averaging them would yield a number not tied to any single protocol. Frontier rows are reference points for individual benchmarks only, and should not be read as a like-for-like ranking against the rows we ran ourselves.

GPT-5. V* 74.87: S1-VL (Li et al., 2026b, Table 6); an independent evaluation reports 71.3 (Yang et al., 2026a, Avg@8). HR-4K 74.6: Yang et al. (2026a, Table 1) (Avg@8, direct answer; 74.2 pass@1 in Yang et al., 2025). HR-8K 71.62: TikArt (Ding et al., 2026, Table 1) (unified API protocol). MME-RW-Lite 57.0: SpatialThinker (Batra et al., 2025, Table 3) (Lite subset, GPT-5-0807). BLINK 71.0: Qwen3-VL Technical Report (Bai et al., 2025a, Table 2) (reasoning = high; 62.8 at minimal reasoning). RWQA 82.8: Qwen3-VL Technical Report (Bai et al., 2025a, Table 2) (reasoning =

Table 8: **Per-tier means** over the 25 ablations. The average score rises across all tiers, and eight of the nine benchmark means also rise monotonically (HR8K decreases by 0.13 from $K{=}3$ to $K{=}5$). The within-tier std decreases $1.09 \rightarrow 0.42$, and all 25 models outperform base.

Tier (n)	V*	HR4K	HR8K	MME-RW	BLINK	POPE	RWQA	CV-B	QSp++	**Avg.**	Δbase	std↓
$K{=}1$ (8)	79.2	71.1	68.0	50.2	53.9	85.8	68.3	74.2	49.1	66.65	+1.99	1.09
$K{=}3$ (8)	79.3	71.3	68.8	50.5	55.0	87.3	70.2	76.8	49.8	67.67	+3.01	0.45
$K{=}5$ (8)	80.2	71.6	68.7	51.3	56.1	88.1	71.1	77.7	52.9	68.61	+3.95	0.42
$K{=}8$ (1)	81.7	71.8	70.9	54.3	56.9	89.2	72.0	78.5	54.5	**69.97**	+5.31	—
base	75.9	68.9	65.2	45.5	55.6	86.6	69.2	73.4	41.6	64.66	0	—

high). QSpatial++ 63.37: HiSpatial (Liang et al., 2026, Table 1) (its ScanNet split is 73.53; the two-split average 68.45 is not comparable to our column). POPE, CV-Bench: no published GPT-5 numbers found.

GPT-4o. V* 62.8: Pixel Reasoner (Su et al., 2026, Table 1); other reports give 65.2 (Lai et al., 2026) and 67.5 (Zhang et al., 2025a). HR-4K 68.0 and HR-8K 63.9: the benchmark authors' leaderboard (Wang et al., 2025b) (gpt-4o-2024-05-13); the HR-Bench paper's cyclic-permutation protocol yields lower values (59.0/55.5) (Wang et al., 2025a). MME-RW-Lite 51.6: SpatialThinker (Batra et al., 2025, Table 3) (Lite subset, GPT-4o-0513); the benchmark authors report 45.2 on the full English set (Zhang et al., 2025b). BLINK 60.0: the benchmark authors' leaderboard (Fu et al., 2024b) (val split; OpenCompass detail-high gives 68.0 OpenCompass Contributors, 2024). POPE 86.9: InternVL3 report (Zhu et al., 2025a) (average of the three POPE splits). RWQA 75.4: OpenVLM Leaderboard (OpenCompass Contributors, 2024) (VLMEvalKit, GPT-4o-0513 detail-high, verified entry). CV-Bench 79.4: SpatialThinker (Batra et al., 2025, Table 1) (2D 75.8 / 3D 83.0, zero-shot greedy). QSpatial++ 61.06: the benchmark authors (Liao et al., 2024, Table 2) (gpt-4o-2024-05-13, standard prompt; ScanNet split 69.41).

Gemini-2.5-Pro. V* 79.2: Pixel Reasoner (Su et al., 2026, Table 1); corroborated at 79.10 by S1-VL (Li et al., 2026b). HR-4K 87.3: Qwen3-VL Technical Report (Bai et al., 2025a, Table 2) (thinking); a lower 67.3 is reported under a different API configuration (Yang et al., 2026a). HR-8K 83.62: TikArt (Ding et al., 2026, Table 1). BLINK 70.6: Seed1.5-VL report (Guo et al., 2025, Table 6) (thinking; reprinted by Bai et al., 2025a). RWQA 78.0: Seed1.5-VL report (Guo et al., 2025, Table 6) (thinking). CV-Bench 86.3: GCA (Chen et al., 2026, Table 1) (2D 81.2 / 3D 92.5; 84.36 in Huang et al., 2026, Table 3). QSpatial++ 40.59: HiSpatial (Liang et al., 2026, Table 1) (ScanNet split 55.46). MME-RW-Lite 58.30: S1-VL (Li et al., 2026b, Table 6); a substantially lower 49.2 is reported by Yang et al. (2026a, Table 1) under an Avg@1 protocol, so this cell is the least settled of the three frontier rows. POPE: no published number found; the Gemini 2.5 report (Comanici et al., 2025) does not cover it.

A.5 Full modality-count ablation

Table 10 lists all 25 ablation models behind Sec. 4.1: eight single-form models ($K{=}1$), eight random triples ($K{=}3$), eight random quintuples ($K{=}5$), and the all-eight model ($K{=}8$), each trained on a 40K budget ($40\text{K}/K$ examples per form) under the configurations of Sec. 3. The exceptions are four single-form models whose forms have fewer usable examples and which train on all of them: depth (18,370), edge (18,328), scene graph (19,088) and segmentation (30,477). Form ids: `1` bbox-crop, `2` bbox-highlight, `3` depth, `4` edge, `5` interleaved helper, `6` scene-graph, `7` segmentation, `8` text-CoT. The last column is a per-benchmark min–max normalized composite (each benchmark scaled to $[0, 100]$ across the 26 rows, then averaged); it agrees with the average-score ranking, placing ours first (90.7) and base last (29.3).

Table 8 reports the per-tier means and Table 9 the statistical tests summarized in Sec. 4.1.

A.6 Full latent-versus-text results

Table 11 gives every score behind Sec. 4.2. Three variants are trained from the 40K eight-form corpus, each re-expressing one or both of the text-serializable forms as explicit decoded text while all remaining forms stay latent, and each is then evaluated under three inference modes: the model

Table 9: **The trend holds under every test** (all on the same 25 models). Left: the question each test addresses; right: the test and its result. The Jonckheere–Terpstra and Mann–Whitney tests are one-sided; the regression and rank-correlation p-values are two-sided.

Question	**Test and result**	p
Does the average score rise with K?	OLS slope $+0.483 \pm 0.079$ (SE) per form (R^2=62.2%)	2.8×10^{-6}
Is the rise monotone (no linearity)?	Spearman ρ=0.879	7.1×10^{-9}
—same question, a second rank estimator	Kendall τ=0.765	1.6×10^{-6}
Are the tiers ordered?	Jonckheere–Terpstra: 205/216 pairs concordant, z=4.79	8.2×10^{-7}
Are the tiers separated?	Mann–Whitney $K5>K3$ and $K5>K1$: win rate 100% (64/64)	7.8×10^{-5}
Do *which* forms matter, beyond *how many*?	nested $F(7,16)$=0.85: no detectable identity effect	0.57

Table 10: **Full modality-count ablation**: 25 models × 9 benchmarks, plus base.

K	**forms**	V*	HR4K	HR8K	MME-RW	BLINK	POPE	RWQA	CV-B	QSp++	**Avg.**	**norm**
1	bbox_crop {1}	79.58	72.05	69.75	50.86	54.32	86.10	70.70	72.08	54.40	67.76	65.1
1	bbox_highlight {2}	80.13	71.68	68.20	51.01	52.81	85.10	67.80	71.59	60.40	67.64	57.2
1	helper {5}	79.08	70.25	67.62	50.17	50.18	87.80	69.90	74.66	52.50	66.91	51.4
1	text_cot {8}	81.15	72.00	69.38	48.93	57.07	85.30	71.10	78.44	37.60	66.77	65.9
1	depth {3}	79.58	69.88	66.62	49.09	55.81	85.14	67.71	76.81	57.41	67.56	53.1
1	edge {4}	74.65	70.38	65.88	49.09	54.92	88.83	63.66	76.21	40.92	64.95	36.0
1	scene_graph {6}	79.58	71.38	68.25	51.75	50.13	81.89	64.97	69.58	45.66	64.80	35.1
1	segmentation {7}	79.90	71.25	68.02	50.70	56.17	86.45	70.59	74.62	43.56	66.81	59.8
3	{1,3,6}	78.44	71.25	68.88	50.65	54.81	88.20	67.70	77.22	46.50	67.07	60.5
3	{2,4,5}	80.63	71.88	69.50	51.26	51.92	88.60	70.60	72.94	56.40	68.19	68.6
3	{1,7,8}	79.06	72.38	69.50	47.99	57.23	87.40	72.40	78.97	43.70	67.63	71.4
3	{3,4,6}	78.53	70.80	68.75	51.07	55.18	85.80	69.80	75.77	51.50	67.47	59.9
3	{3,4,5}	77.49	70.75	68.00	53.31	52.45	87.16	69.80	76.69	56.44	68.01	61.2
3	{6,7,8}	79.58	71.50	67.50	48.04	56.86	89.30	71.63	78.45	47.50	67.82	68.2
3	{3,6,7}	79.76	70.70	69.00	50.85	54.87	88.13	69.02	76.65	43.68	66.96	60.9
3	{4,5,8}	80.63	71.50	69.12	50.86	56.55	84.19	70.59	77.85	52.46	68.19	69.0
5	{1,3,4,5,6}	80.63	71.50	68.12	53.52	54.92	87.10	70.30	74.59	53.50	68.24	69.2
5	{2,4,6,7,8}	80.10	72.00	68.88	48.78	56.65	88.20	70.60	77.67	53.50	68.49	73.1
5	{1,2,3,7,8}	79.49	71.62	68.25	49.72	56.34	87.20	71.50	78.53	52.50	68.35	70.5
5	{3,4,5,6,7}	80.10	71.88	68.75	53.99	53.82	88.30	68.80	74.06	56.50	68.47	70.8
5	{4,5,6,7,8}	81.15	72.00	69.38	53.31	53.02	89.13	71.63	77.94	58.42	69.55	82.0
5	{3,5,6,7,8}	79.58	70.75	68.25	51.43	57.92	88.75	72.03	79.99	52.48	69.02	76.7
5	{3,4,6,7,8}	80.10	71.50	68.12	48.57	57.34	88.99	71.24	78.80	49.50	68.24	71.9
5	{3,4,5,7,8}	80.10	71.88	69.50	50.91	58.50	86.88	72.42	79.91	46.59	68.52	78.4
8	**{1–8} (ours)**	**81.70**	71.75	**70.88**	**54.30**	56.85	**89.19**	72.03	78.53	54.46	**69.97**	**90.7**
—	base (Qwen2.5-VL-7B)	75.92	68.88	65.25	45.49	55.65	86.58	69.15	73.40	41.58	64.66	29.3

routes freely between the two forms (*free*), or is required to emit the latent form (*forced latent*), or is required to emit text (*forced text*). The scene-graph variant routes to the latent for every question, so its free and forced-latent runs produce identical predictions and are reported once. Table 12 gives the routing and token statistics for the same runs. For CV-Bench, the 2D/3D split is $73.85/83.42$ for the scene-graph variant under free and forced latent, $73.75/75.25$ for both-forced-latent, and $68.71/78.03$ for text-CoT-forced-latent.

Table 11: **Latent versus decoded text**: three training variants × three inference modes, against the all-latent model. The average score is over all nine benchmarks.

decoded form	**inference**	V*	HR4K	HR8K	MME-RW	BLINK	POPE	RWQA	CV-B	QSp++	**Avg.**
— (all latent, ours)	latent	**81.70**	71.75	**70.88**	**54.30**	**56.85**	**89.19**	72.03	78.53	54.46	**69.97**
scene graph	free = forced latent	79.58	72.88	69.25	52.63	54.55	88.71	**72.42**	**78.63**	54.34	69.22
	forced text	79.06	72.50	69.25	53.00	54.50	88.88	71.90	77.58	**55.44**	69.12
text CoT	free (routes to text)	76.44	72.12	66.50	52.32	54.39	84.63	66.54	74.84	54.34	66.90
	forced latent	81.15	71.12	66.25	52.94	52.29	88.96	70.59	73.37	54.34	67.89
	forced text	75.92	72.62	66.25	52.32	53.02	84.63	66.67	74.84	54.34	66.73
both	free (routes to text)	78.01	72.62	65.88	51.17	52.81	84.04	66.27	74.78	**55.44**	66.78
	forced latent	79.06	71.63	67.88	53.10	52.87	88.21	70.33	74.50	**55.44**	68.11
	forced text	78.01	**73.00**	66.62	52.27	53.39	84.10	66.27	74.74	53.47	66.87

A.7 INFERENCE-COST MEASUREMENT

Table 13 lists every number behind Fig. 5. All eight methods were run from their official checkpoints over the same 2,657 questions—V* (191), HR-Bench 4K (800), HR-Bench 8K (800), RealWorldQA

Table 12: **Routing and inference cost**, per mode, for the runs of Table 11. *Think steps* counts the decoding steps spent between the thinking markers and *span* the sequence positions they occupy. All columns are means over the evaluation suite; the latent block is a fixed $L{=}8$ steps over 9 positions on every question, independent of difficulty. Input length is identical across runs because the question and image are unchanged.

run	routing	input	think steps	span	output	total
scene graph, free	100% latent (unprompted)	3288	8.0	9.0	15.0	3303
scene graph, forced latent	100% latent (forced)	3288	8.0	9.0	15.0	3303
scene graph, forced text	100% text	3288	26.8	26.8	32.8	3320
text CoT, free	100% text (collapsed)	3288	177.1	177.1	183.1	3471
text CoT, forced latent	100% latent (forced)	3288	8.0	9.0	15.8	3303
text CoT, forced text	100% text (forced)	3288	177.1	177.1	183.2	3471
both, free	99.2% text (collapsed)	3288	177.6	177.6	183.7	3471
both, forced latent	100% latent (forced)	3288	8.0	9.0	15.1	3303
both, forced text	100% text (forced)	3288	179.1	179.1	185.2	3473

(765) and QSpatial++ (101)—with identical prompts. The accuracy column is the unweighted mean of the corresponding five columns of Table 2, reported here so that cost and capability can be read together.

Token counts. These are recorded per question rather than estimated. The input length is the length of the tokenized prompt after multimodal expansion, and the number of visual tokens is counted separately from the image placeholders so that the vision-tower FLOPs can be computed exactly. *Think tokens* for a latent method are the positions of the latent span. The block is $L{=}8$ slots: in training the span is written as an opening delimiter, eight slot positions and a closing delimiter, and only the eight slot positions are treated as latent. At inference the opening delimiter's own embedding is overwritten by the first injected state and so is never consumed, which makes the measured span a fixed 9 positions per question. The number of injection steps is 8 in both cases. For a text method the think budget is the generated non-latent tokens preceding the final `\boxed{`, obtained by re-tokenizing the decoded text; this is the only approximate quantity in the table, to within a few tokens. Tool-using methods additionally log the number of rounds and tool calls, and the cumulative prefill across rounds.

FLOPs. We count analytically rather than reading hardware counters. Since all eight methods share the Qwen2.5-VL-7B backbone, the architecture constants are identical and the entire difference between rows arises from the number of tokens each method processes. Writing $d{=}3584$, $N{=}28$ layers, $d_{kv}{=}512$, $d_{ff}{=}18944$ and $V{=}152064$ for the language model,

$$\text{lin/tok} = 2\left[N\left(2d^2 + 2d\,d_{kv} + 3d\,d_{ff}\right) + dV\right],$$
$$\text{prefill}(T) = \text{lin}\cdot T + 2dN\,T^2, \qquad \text{decode}(T,n) = \text{lin}\cdot n + 4dN\left(Tn + \tfrac{n(n-1)}{2}\right),$$

and for the vision tower, with $P{=}4n_{\text{vision}}$ patches, the linear term plus windowed attention ($64P$ over 28 layers), global attention (P^2 over 4 layers) and the merger. Norms, RoPE, softmax and activations are omitted; together they are under 2%. A latent step and an ordinary decoding step run the same forward pass (hidden-state injection only overwrites the input embedding), so decode FLOPs are counted uniformly over the generated tokens. For the tool-using methods the extra prefill is the cumulative per-round total minus the first input, scored under the assumption that the prefix cache hits, which is the reading least favourable to us.

Latency. We measure end-to-end wall-clock time, one question at a time. Timing covers only the generate call; image preprocessing is recorded separately and excluded, and `torch.cuda.synchronize()` brackets the HuggingFace path. Each benchmark is preceded by ten untimed warm-up questions and the first five rows of every output file are discarded at aggregation. Prefix caching is disabled, since the warm-up and the shared prompt prefix would otherwise make the timings optimistic. Each model is sharded across independent processes, each of which starts a fresh engine and repeats the warm-up, which also limits drift over a long run.

Table 13: **Measured inference cost**, 2,657 questions per model on one H100 at batch size one. *vs. ours* is the latency ratio. Accuracy is the unweighted mean of the same five benchmarks.

Method	latency (s)	vs. ours	decode TFLOPs	think tok.	output tok.	5-bench avg.
Ours	**0.681**	1.00×	**0.294**	9.0	**15.9**	**70.16**
LVR-7B (Li et al., 2026a)	0.821	1.21×	0.359	**8.0**	22.6	62.18
Mirage-VSP (Yang et al., 2026b)	0.934	1.37×	0.706	33.5	35.1	65.71
base (direct answer)	1.212	1.78×	1.001	47.3	49.6	64.16
Monet-7B (Wang et al., 2026)	2.298	3.37×	2.589	129.1	134.0	62.57
DeepEyes (Zheng et al., 2026)	2.564	3.77×	4.728	179.6	184.2	68.13
PixelReasoner (Su et al., 2026)	3.072	4.51×	8.645	159.5	172.4	65.77
SkiLa-7B (Tong et al., 2025)	3.687	5.42×	2.456	29.3	154.2	65.77

B Extended Related Work

Multimodal Chain-of-Thought. Multimodal reasoning employs diverse intermediate representations to expose different forms of visual evidence. Textual rationales support multi-step reasoning (Zhang et al., 2023; Xu et al., 2025; Yao et al., 2026), while regions and bounding boxes localize relevant objects (Shao et al., 2024; Yang et al., 2023). Scene graphs capture compositional relations (Mitra et al., 2024; Batra et al., 2025), segmentation masks provide fine-grained object extent (Lai et al., 2024), and depth representations encode 3D spatial geometry (Chen et al., 2024; Liao et al., 2024). Sketches and edges emphasize structural cues (Hu et al., 2024), whereas interleaved traces combine textual and visual reasoning (Gu et al., 2026; Li et al., 2025a).

Thinking with images. A parallel line of work incorporates visual operations directly into reasoning to preserve information that may be lost through textualization. Models may crop or search relevant regions (Shao et al., 2024; Wu & Xie, 2024), learn zooming and multi-turn visual interaction (Su et al., 2026; Zheng et al., 2026; Lai et al., 2026; Qiao et al., 2025), or manipulate images through executable tools (Zhang et al., 2025a). Others externalize reasoning as generated images or sketches (Li et al., 2025b; Hu et al., 2024), while Set-of-Mark prompting augments inputs with visual marks to facilitate grounding (Yang et al., 2023). These methods demonstrate the value of diverse visual reasoning forms. We further study whether different visual reasoning forms and operations can be supervised into a shared latent reasoning space in which they benefit one another.

Latent reasoning. Latent reasoning replaces explicit chain-of-thought with continuous hidden-state computation, avoiding token-by-token verbalization (Zhu et al., 2025b). Prior work internalizes textual reasoning through hidden-state distillation (Deng et al., 2024; 2023), recurrent latent steps (Hao et al., 2024; Shen et al., 2025; Cheng & Van Durme, 2024; Geiping et al., 2026), soft tokens (Zhang et al., 2026), or learned latent codes (Su et al., 2025). Multimodal variants similarly internalize a predefined form, such as visual regions (Li et al., 2026a), latent visual thoughts (Yang et al., 2026b), interleaved text and visual latents (Wang et al., 2026), expert features (Qin et al., 2025), or sketches (Tong et al., 2025). In contrast, we study whether heterogeneous visual reasoning forms can be unified in a single shared latent reasoning space.

C Limitations

Our study has several limitations. The diversity trend is measured at a single supervision budget of 40K examples and for up to eight forms; how it behaves at larger budgets or with more forms remains to be tested. And the latent block has a fixed size of L slots, so the amount of latent computation does not adapt to the difficulty of a question. Finally, the latent slots are not mapped to the vocabulary, which makes the intermediate reasoning harder to inspect than an explicit rationale.