INTERLUDE: REASONING BETWEEN TOKENS IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal reasoning mixes two kinds of intermediate steps: symbolic relations that language expresses precisely and continuous visual judgments that text can only approximate. Approaches using a single medium face a trade-off: explicit chain-of-thought can discard fine visual details, while purely latent reasoning can lose the readable scaffold. We propose **Interlude**, retaining language as the backbone while punctuating it with latent segments: short, undecoded spans of recurrent hidden-state computation used by subsequent text. We seek states that remain useful to subsequent text while retaining visual evidence that is difficult to verbalize. Motivated by information under-saturation and the modality gap, we combine semantic guidance on the language-head readout with a geometric margin favoring an intermediate reference over either modal pole. Interlude improves over Qwen3-VL-8B-Instruct on all six evaluated benchmarks, with an average gain of 4.72%, and achieves the highest Overall score among the compared explicit and latent reasoning baselines. Its Overall score also exceeds that of Qwen3-VL-8B-Thinking while using just 7.6% of its decoded tokens in our efficiency evaluation. Geometric analysis shows a pronounced increase in visual alignment for the component orthogonal to the word-embedding principal subspace, with little change within that subspace.

1 INTRODUCTION

Chain-of-thought (CoT) prompting improves reasoning by making intermediate steps explicit (Wei et al., 2022), and vision-language models (VLMs) now scale inference-time computation similarly, generating extended textual traces that organize visual evidence before producing an answer (Dong et al., 2025). However, these reasoning traces consist of steps that differ in nature. Some steps involve discrete symbolic relations, such as spatial comparisons, which language can express with minimal information loss. Conversely, other steps are continuous visual judgments, such as metric estimation or occlusion ordering, which language can only approximate. These two types of steps often interleave within a single problem: in Figure 1(a), deciding which circles lie above four feet requires both a relation that words can state exactly and an estimate they can only approximate.

Approaches using a single reasoning medium emphasize different kinds of reasoning. Explicit CoT expresses intermediate steps as words, which suit symbolic relations but can omit fine visual details. Textual rationales can be unfaithful (Turpin et al., 2023; Jacovi & Goldberg, 2020; Chen et al., 2024b), while language priors can outweigh visual evidence (Long et al., 2026). Purely latent reasoning avoids committing every step to words, but provides less explicit textual structure. Existing methods train latent states through text compression (Shen et al., 2025), semantic alignment (Wang et al., 2026), visual-token reconstruction (Li et al., 2025), or visual-latent distillation and policy optimization (Wang et al., 2025b). These choices motivate two questions: which reasoning steps should bypass language, and how should latent states be trained so that subsequent text can use them?

Addressing the first question, we retain language as the backbone and assign difficult visual judgments to brief, undecoded segments of hidden-state computation. Subsequent text attends to these states to continue reasoning. For the second question, we use language-head supervision to encourage semantic compatibility with textual generation. Strongly favoring text may limit the visual detail a state retains, while favoring vision alone may weaken its compatibility with the textual path. This motivates jointly supervising semantic content and cross-modal geometry.

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.

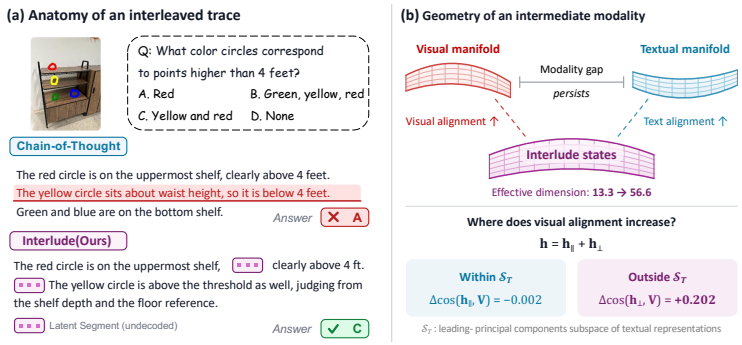


Figure 1: **(a)** CoT approximates the yellow circle's height in words; Interlude keeps that step latent. **(b)** Interlude states approach both modalities, with a larger increase in visual alignment outside the word-embedding principal subspace.

Subsequent text accesses latent states through self-attention. The language-head readout provides a separate way to supervise their semantic content. Prior work describes a bottleneck in this read-out (Yang et al., 2018), anisotropy in language representations (Ethayarajh, 2019), and a geometric gap between visual and textual embeddings (Liang et al., 2022). These observations motivate our hypothesis of *information under-saturation*: latent states may have capacity for visual evidence beyond the dominant directions associated with text. We therefore combine semantic supervision with a geometric objective that allows states to depart from those directions. Section 3 characterizes the backbone's representation geometry, finding weak visual alignment and low effective dimensionality. We use *intermediate modality* to describe states trained for both semantic compatibility and visual grounding, without requiring them to coincide with either modal pole. Figure 1(b) summarizes the geometric changes observed after training.

We implement this approach as **Interlude**. Each latent segment spans at most three hidden states recursively fed back as inputs; it emits no content words, and subsequent text attends to it through the base architecture. Semantic guidance supervises the language-head readout over a stage-dependent concept window while retaining normalization over the full vocabulary. This supplies a soft semantic target without imposing a one-hot token label. The intermediate-modality constraint uses a margin to favor a reference between the textual and visual centroids over either centroid. It does not require the state to lie on the arc joining them or specify an orientation outside their plane. Both objectives use ordered concept sets extracted from teacher traces with correct answers. Interlude thus punctuates the verbal trace with latent computation and trains the two reasoning media together.

Our contributions are threefold:

• ***Paradigm Formulation***: We use the hypothesis of information under-saturation to formulate latent segments as an intermediate modality between discrete text and continuous visual evidence.

• ***Methodological Innovation***: We introduce **Interlude**, a reasoning paradigm interleaving textual scaffolds with latent segments. It is optimized via semantic guidance and a novel intermediate-modality constraint.

• ***Empirical Performance***: Interlude improves over Qwen3-VL-8B-Instruct across all six evaluated benchmarks, with an average gain of 4.72 percentage points, and achieves the highest Overall score among the compared explicit and latent reasoning baselines.

2 Related Work

2.1 Chain-of-Thought Reasoning in VLMs

CoT prompting elicits intermediate natural-language steps (Wei et al., 2022; Xu et al., 2025) and has been extended through zero-shot prompting, self-consistency, decomposition and rationale supervision (Kojima et al., 2022; Wang et al., 2023; Zhou et al., 2023; Zelikman et al., 2022). In VLMs it has been instantiated with multimodal rationale datasets, rationale-conditioned inference and grounded procedures that tie rationales to image regions, visual marks or structured stages (Lu et al., 2022; Zhang et al., 2024; Shao et al., 2024; Yang et al., 2023; Zhang et al., 2025; Dong et al., 2025). Search and reinforcement learning develop multi-step reasoning (Yang et al., 2025b; Yao et al., 2025; Meng et al., 2025), but longer rationales do not imply faithful ones (Turpin et al., 2023;

Jacovi & Goldberg, 2020; Chen et al., 2024b). Interlude keeps text as the interpretable scaffold but moves what it expresses poorly into latent segments.

2.2 Latent Space Reasoning

Latent reasoning asks whether intermediate computation can be carried by continuous hidden states instead of decoded tokens. In language-only settings, pause tokens, implicit CoT, Coconut and compressed reasoning representations show that computation can proceed without verbalizing every step (Goyal et al., 2024; Deng et al., 2023; Hao et al., 2025; Cheng & Van Durme, 2024; Tan et al., 2025). VLM-oriented methods extend this: Heima compresses multimodal CoT into hidden thinking tokens, LVR trains latent states to reconstruct answer-critical visual tokens, Monet optimizes continuous visual thoughts, and Laser aligns latent states with dynamic semantic windows (Shen et al., 2025; Li et al., 2025; Wang et al., 2025b; 2026). Hybrid methods also interleave textual and latent reasoning: SkiLa generates continuous visual sketches, SwimBird switches among textual, visual, and interleaved modes, and HyLaR trains hybrid trajectories with decoupled policy optimization (Tong et al., 2025; 2026; Cheng et al., 2026). Our method differs in jointly constraining the latent state's textual readout and its geometry relative to both modality poles (Eq. 7).

3 Where Should a Visual Thought Live?

The language head maps a hidden state $\mathbf{h}$ to logits through $W_U\mathbf{h}$. A perturbation Δ changes the logits by $W_U\Delta$; a small change leaves the next-token distribution nearly unchanged. We call the hypothesis that hidden states have capacity beyond their immediate textual readout ***information under-saturation***. The linear-softmax bottleneck (Yang et al., 2018) and anisotropy in language representations (Gao et al., 2019; Ethayarajh, 2019) motivate this hypothesis, but do not identify usable directions. The visual–textual embedding gap (Liang et al., 2022) further motivates examining visual alignment outside the dominant word-embedding directions.

Table 1: Backbone measurements, before any Interlude training, on ERQA and VSI-Bench items. $\mathbf{H}$ denotes reasoning hidden states, $\mathbf{V}$ projected visual tokens, $\mathbf{T}$ input word embeddings, and $\mathcal{S}_T$ their leading-PC subspace. Angles are between centroids; cosines are per-example means.

Quantity	Backbone
$\angle(\mathbf{H}, \mathbf{V})$	89.4°
$\cos(\mathbf{H}, \mathbf{V})$	0.008
$\cos\big(\mathrm{Proj}_{\mathcal{S}_T}\mathbf{H},\ \mathbf{V}\big)$	0.029
$\cos\big(\mathrm{Proj}_{\mathcal{S}_T^\perp}\mathbf{H},\ \mathbf{V}\big)$	0.007
Effective dim. of $\mathbf{H}$	13.3

On ERQA and VSI-Bench, we compare reasoning states, projected visual tokens and input word embeddings using the protocol in Section 6. Table 1 shows weak visual alignment (89.4° between centroids; 0.008 mean per-example cosine). Similarity is low both outside (0.007) and inside (0.029) the word-embedding principal subspace; effective dimensionality is 13.3. These measurements show weak alignment and concentrated variation; they do not establish that unused directions can preserve visual evidence without disrupting textual reasoning.

The resulting design question is whether latent computation can strengthen visual grounding while preserving semantic compatibility with subsequent text. This requires two complementary targets: concept-level supervision alone does not directly constrain visual alignment, while visual similarity alone does not explicitly preserve relevant semantics. We therefore pair readout supervision with a geometric preference for an intermediate reference, giving both textual and visual centroids a role. Subsequent text accesses states through self-attention; the readout is a proxy for semantic compatibility, not a guarantee of downstream use. Section 4.3 instantiates these requirements with semantic guidance and an intermediate-modality constraint.

We test whether training improves visual alignment and task performance, with benefits varying by the visual judgments required. We hypothesize a larger visual-similarity increase outside the word-embedding principal subspace; the loss does not enforce this pattern or fix the in-subspace component. Effective dimensionality tests whether variation spreads across more directions. Sections 5 and 6 examine these hypotheses; Appendix B develops the geometric motivation.

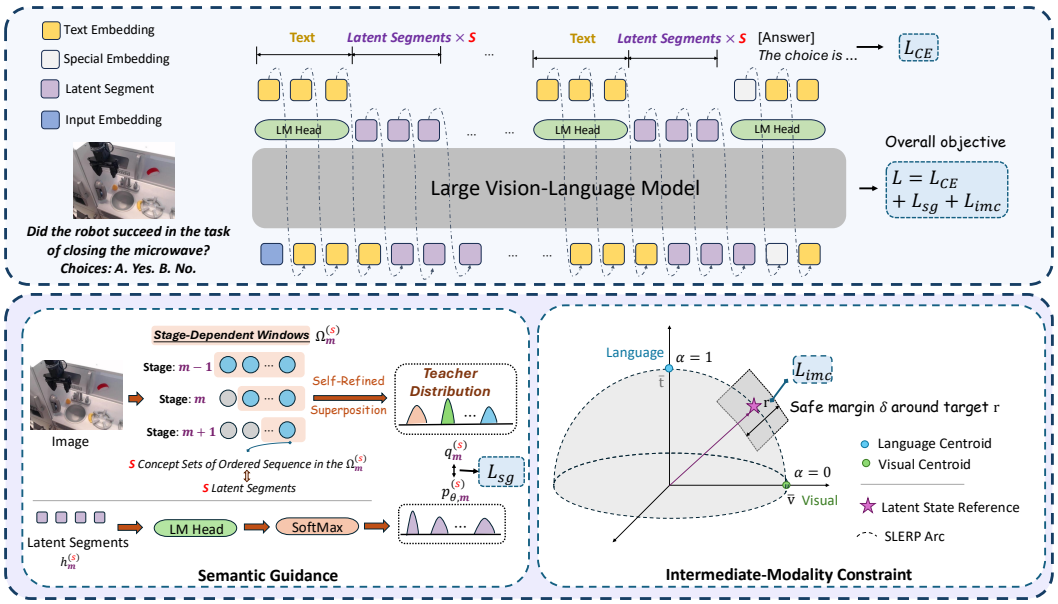


Figure 2: Architecture and training objective of **Interlude**. The model interleaves text with latent segments during reasoning, and optimizes latent segments using semantic guidance and an intermediate-modality constraint between language and visual centroids.

4 METHOD

Following the joint requirements in Section 3, Interlude combines semantic supervision with a geometric objective. This raises three design questions: how a latent segment operates (§ 4.1), what each of its steps should represent (§ 4.2), and how its states are trained for semantic compatibility and visual grounding (§ 4.3).

4.1 AN INTERLEAVED REASONING PARADIGM

Language provides the reasoning scaffold: text states the plan, records observations and delivers the answer. Latent segments add hidden-state computation at steps where visual judgments may be difficult to express precisely in words. Given an image I and a question Q, the output is a single sequence in two media,

$$y = u_1\,[\mathbf{H}_1]\,u_2\,[\mathbf{H}_2]\cdots u_M\,[\mathbf{H}_M]\,u_{M+1}\,\hat{y}, \tag{1}$$

where $\hat{y}$ is the answer, each u_m is an ordinary segment of decoded text and each $\mathbf{H}_m$ is a *latent segment*: a short span of hidden states that is never decoded.

At standard text positions, the model applies the language head and samples tokens; within a latent segment, it feeds each hidden state back as the next input embedding without emitting content words. Its states occupy sequence positions that the subsequent text segment u_{m+1} can access through standard self-attention. The model f_θ acts as its own recurrent operator, requiring no architectural changes. During training, the language head provides semantic supervision for these states (§ 4.3); at inference, it is used only to predict the closing boundary within a latent segment, without decoding the latent content into words. For the m-th latent segment,

$$\mathbf{h}_m^{(s)} = f_\theta\big(\mathbf{h}_m^{(s-1)} + \mathbf{e}_{\text{step}}(s)\,\big|\,\mathcal{C}_m\big), \quad s = 1,\ldots,k_m, \;\; k_m \in \{1,2,3\}, \tag{2}$$

where $\mathbf{h}_m^{(0)}$ is the hidden state at the opening boundary token <|latent|>, $\mathcal{C}_m$ is the preceding KV cache, and $\mathbf{e}_{\text{step}}(s)$ distinguishes steps within the span.

4.2 SYNTHESIZING THE SUPERVISION

Training **Interlude** requires traces in which the boundaries between the two media are marked and the content of each latent segment is specified. A teacher VLM writes a verbal reasoning trace for

each image-question pair; traces whose final answer is incorrect are discarded; and within each retained trace the positions carrying a continuous visual judgment are marked as latent segments. Each is decomposed into an ordered sequence of concept sets $\{\mathcal{S}_m^{(1)}, \ldots, \mathcal{S}_m^{(k_m)}\}$ running from global anchors toward local cues. The direction follows global precedence in human vision (Navon, 1977): early states capture scene-level context, while later states focus on evidence that determines the answer. Filtering for answer correctness before concept extraction helps control supervision quality. The sets describe how the state should evolve rather than specify a unique correct path, and the objectives below are correspondingly soft. The corpus contains 50K examples (Appendix C).

4.3 Shaping the latent state into an intermediate modality

Semantic guidance: stage-windowed semantic supervision. Semantic guidance supervises the state's language-head readout over a stage-dependent window of current and future concepts, encouraging semantic compatibility with subsequent text. The window narrows as the segment progresses, providing broader supervision at early steps and more specific targets at later ones. The soft target uses the model's detached prediction restricted to this window. The student distribution remains normalized over the full vocabulary, so the loss encourages probability mass on the relevant concepts without imposing a one-hot token target. The concept sets provide supervision and never appear in the output sequence.

As illustrated in Figure 2, at step s of latent segment m, the readout distribution is

$$\ell_m^{(s)} = W_U \mathbf{h}_m^{(s)}, \qquad p_{\theta,m}^{(s)} = \mathrm{softmax}(\ell_m^{(s)}), \tag{3}$$

and the window is $\Omega_m^{(s)} = (\mathcal{S}_m^{(s)} \cup \cdots \cup \mathcal{S}_m^{(k_m)}) \setminus \mathcal{X}$, where $\mathcal{X}$ removes stop words and uninformative high-frequency fragments; by construction $\Omega_m^{(1)} \supseteq \cdots \supseteq \Omega_m^{(k_m)}$. With the teacher distribution $q_m^{(s)}$ defined in Appendix A, the loss is

$$\mathcal{L}_{\mathrm{sg}} = -\sum_{m=1}^{M} \frac{1}{k_m} \sum_{s=1}^{k_m} \sum_{k \in \Omega_m^{(s)}} q_m^{(s)}(k) \log p_{\theta,m}^{(s)}(k). \tag{4}$$

Intermediate-modality constraint. Semantic guidance provides concept-level targets for the readout, but does not directly constrain similarity to visual tokens. Direct visual alignment, as used in prior latent reasoning methods (Li et al., 2025), provides one way to encourage visual grounding. We instead use a stage-dependent reference between the textual and visual centroids, and train the state to have greater cosine similarity to this reference than to either centroid. This gives both modalities a role in the geometric objective.

Let $\mathbf{e}(c_j)$ denote the token embedding for each concept token $c_j \in \mathcal{S}_m^{(s)}$, and let $\{\mathbf{z}_{m,i}\}_{i=1}^{N_v}$ represent the projected visual tokens of the current example. Using the textual centroid as a query, we retrieve relevant visual-token indices, denoted by $\mathcal{I}_m^{(s)}$, and define

$$\bar{\mathbf{t}}_m^{(s)} = \frac{1}{|\mathcal{S}_m^{(s)}|} \sum_{c_j \in \mathcal{S}_m^{(s)}} \mathbf{e}(c_j), \qquad \bar{\mathbf{v}}_m^{(s)} = \frac{1}{|\mathcal{I}_m^{(s)}|} \sum_{i \in \mathcal{I}_m^{(s)}} \mathbf{z}_{m,i}. \tag{5}$$

Let $\mathbf{t}_m^{(s)}$ and $\mathbf{v}_m^{(s)}$ denote the ℓ_2-normalized centroids, defined by $\mathbf{t}_m^{(s)} = \bar{\mathbf{t}}_m^{(s)}/\|\bar{\mathbf{t}}_m^{(s)}\|_2$ and $\mathbf{v}_m^{(s)} = \bar{\mathbf{v}}_m^{(s)}/\|\bar{\mathbf{v}}_m^{(s)}\|_2$. With $\theta_m^{(s)} = \arccos({\mathbf{v}_m^{(s)}}^{\top} \mathbf{t}_m^{(s)})$, the reference is defined by their spherical linear interpolation,

$$\mathbf{r}_m^{(s)} = \tfrac{\sin((1-\tau_m^{(s)})\theta_m^{(s)})}{\sin\theta_m^{(s)}} \mathbf{v}_m^{(s)} + \tfrac{\sin(\tau_m^{(s)}\theta_m^{(s)})}{\sin\theta_m^{(s)}} \mathbf{t}_m^{(s)}, \qquad 1 \geq \tau_m^{(1)} \geq \cdots \geq \tau_m^{(k_m)} \geq 0, \tag{6}$$

with a schedule that starts near the textual centroid, where compatibility with the high-level reference matters most, and moves toward the visual centroid as the evidence becomes more specific. Omitting the indices m, s, we define a two-sided margin loss,

$$\mathcal{L}_{\mathrm{imc}} = \sum_{m=1}^{M} \frac{1}{k_m} \sum_{s=1}^{k_m} \Big(\big[d(\mathbf{h}, \mathbf{r}) - d(\mathbf{h}, \mathbf{v}) + \delta \big]_+ + \big[d(\mathbf{h}, \mathbf{r}) - d(\mathbf{h}, \mathbf{t}) + \delta \big]_+ \Big). \tag{7}$$

Here, $d(\mathbf{a}, \mathbf{b}) = 1 - \cos(\mathbf{a}, \mathbf{b})$, $[\cdot]_+ = \max(0, \cdot)$ and δ is a margin. Gradients vanish once both margin conditions are satisfied, so the constraint does not pull states onto a single point. The constraint encourages greater cosine similarity to the intermediate reference than to either modal centroid, without requiring $\mathbf{h}_m^{(s)}$ to coincide with that reference.

Overall objective. The three terms are combined as

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{sg}} + \mathcal{L}_{\mathrm{imc}}, \tag{8}$$

where $\mathcal{L}_{\mathrm{CE}}$ is the cross-entropy on the text segments and the answer, including boundary-token supervision. Both latent objectives average over stages within each segment and sum over segments. Text weighting and boundary supervision are detailed in Appendix D. Together, the three terms train the textual path, provide semantic supervision for latent segments, and encourage visual grounding during hidden-state computation. A more detailed discussion of the geometric motivation is provided in Appendix B. At inference, generating <|latent|> opens a latent segment. The segment ends when the model predicts <|/latent|>, successive-state cosine similarity exceeds 0.99, or three steps are reached. Stage concept sets are required only during training.

5 Experiments

We evaluate Interlude on six benchmarks covering embodied reasoning (ERQA (Gemini Robotics Team et al., 2025)), visual-spatial reasoning from video (VSI-Bench (Yang et al., 2025a)), real-world grounding (RealWorldQA (xAI, 2024)), sensitivity to language priors (HallusionBench (Guan et al., 2024)), general multimodal capability (MMStar (Chen et al., 2024a)), and fine-grained visual understanding (V*Bench (Wu & Xie, 2024)), with Qwen3-VL-8B-Instruct as the backbone. Training and evaluation details are given in Appendix D, the baselines are described in Appendix E, and the benchmarks in Appendix F.

5.1 Main Experiments

Table 2: Main results comparing **Interlude** with baselines and ablations. Across model and ablation rows, best scores are in **bold** and second-best scores are underlined, including ties; Δ is the absolute gain over Qwen3-VL-8B-Instruct and is excluded from ranking. † marks additional baselines with public code and weights. Overall is the unweighted mean of the six benchmark scores. Scores are rounded to one decimal place; the Overall gain is computed before rounding.

Model	ERQA	VSI-Bench	RealWorldQA	HallusionBench	MMStar	V*Bench	Overall
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	35.3	53.2	68.5	52.9	63.9	79.1	58.8
Qwen3-VL-8B-Instruct (Bai et al., 2025a)	42.8	57.1	70.2	61.1	68.6	83.8	63.9
Qwen3-VL-8B-Thinking (Bai et al., 2025a)	43.5	54.8	73.5	65.4	**73.9**	77.5	64.8
InternVL3.5-8B (Wang et al., 2025c)	37.0	53.9	67.5	54.5	69.3	73.8	59.3
VL-Rethinker (Wang et al., 2025a)	37.5	52.7	68.2	71.1	63.2	64.6	59.6
Vision-R1 (Huang et al., 2025)	32.5	48.2	64.7	63.8	62.7	78.2	58.4
Monet (Wang et al., 2025b)	34.5	51.8	68.9	56.4	60.3	83.3	59.2
LVR (Li et al., 2025)	36.8	54.5	69.3	65.2	57.9	81.7	60.9
Laser (Wang et al., 2026)	36.0	52.9	69.5	67.7	60.3	83.2	61.6
SwimBird† (Tong et al., 2026)	43.8	57.3	73.1	68.2	71.2	85.5	66.5
SkiLa† (Tong et al., 2025)	35.0	52.6	69.3	61.9	64.8	84.3	61.3
HyLaR-7B† (Cheng et al., 2026)	36.0	53.7	70.5	63.7	62.0	83.8	61.6
Interlude (Ours)							
Interlude (Ours)	**48.3**	**59.2**	**73.6**	**71.8**	72.1	**86.9**	**68.7**
$\Delta\uparrow$ vs. Qwen3-VL-8B-Instruct	+5.5	+2.1	+3.4	+10.7	+3.5	+3.1	+4.7
Ablation Study							
Interlude (w/o Latent Segment)	44.5	58.3	71.9	69.5	70.8	83.2	66.4
Interlude (w/o Intermediate-Modality Constraint)	44.0	57.5	72.5	71.2	72.1	83.8	66.9
Interlude (w/o Semantic Guidance)	43.5	59.0	70.6	68.8	69.7	85.3	66.2
Qwen3-VL-8B-Instruct (SFT on the synthetic traces)	42.3	57.8	71.5	66.5	68.8	83.9	65.1

We evaluate Interlude across complementary visual reasoning tasks to examine whether interleaving textual reasoning with latent computation improves performance over the same backbone (Table 2). Interlude outperforms Qwen3-VL-8B-Instruct on all six benchmarks and achieves the highest Overall score among the compared explicit and latent reasoning baselines.

The largest gains occur on HallusionBench and ERQA, where answering correctly requires following visual evidence despite misleading language priors or resolving spatial and physical relationships. Improvements on V*Bench further extend to fine-grained visual details. This pattern is consistent with our motivation: retaining intermediate visual judgments in hidden states may help preserve information that is difficult to express faithfully in text. The gains on RealWorldQA and MMStar suggest that this benefit also carries over to broader visual understanding tasks.

We also observe improvement on VSI-Bench despite using no video examples in training, suggesting that image-based supervision can transfer to video-based spatial reasoning. Its more modest gain indicates that the benefits vary with task demands. The following ablations examine how synthetic-data fine-tuning, latent computation, and the two supervision objectives contribute to these improvements.

5.2 Where the gain comes from

The ablation study in Table 2 uses the same six-benchmark Overall score as the main results. All gains below are computed before rounding. Ordinary verbal CoT fine-tuning on the same data achieves 65.1, a gain of 1.2 over the backbone, compared with the full model's 4.7-point gain. Interlude adds 3.5 points over this CoT baseline. The trained model with latent segments suppressed gains 1.2 points over CoT, and enabling the segments adds another 2.3 points. The suppressed configuration is trained with Eqs. 4 and 7 but never opens a latent segment at test time. Its improvement suggests that the benefits of Interlude training extend to evaluation using only the textual path.

Removing semantic guidance reduces Overall to 66.2, slightly below the 66.4 obtained with latent segments suppressed. This comparison suggests that latent computation benefits from semantic supervision in this setting. Removing the intermediate-modality constraint yields 66.9, with the largest drops on ERQA and V*Bench.

Specifically, performance drops by 4.3 points on ERQA, 3.1 points on V*Bench, and 1.7 points on VSI-Bench, while remaining unaffected on MMStar. These ablations support the complementary contributions of the two objectives, with particularly large benefits from the intermediate-modality constraint on ERQA and V*Bench.

5.3 Which subtasks benefit

We analyze Interlude on ERQA subtasks to examine its strengths and limitations. ERQA groups its items into subtasks that differ in how readily the decisive step can be expressed in words. This breakdown provides a finer test of the hypothesis in § 3 than aggregate benchmark scores. Figure 3 shows the gains concentrated on Pointing, Spatial Reasoning, State Estimation, Task Reasoning, Trajectory Reasoning and Action Reasoning. They are largest when the decisive step involves estimating a continuous scene property: where a target is, how a scene is laid out, what state an object is in. Gains vary with a task's emphasis on interpreting the current scene versus simulating changes over time or across viewpoints. This is consistent with the hypothesis that a latent segment can retain judgments that language would express only approximately, even with at most three steps of continuous state. Multi-view Reasoning is the only subtask that declines, and two factors plausibly contribute. One is depth: reconstructing 3D structure across views may exceed what three steps of continuous state can compute. The other is direction: the concept sets of § 4.2 run from global anchors toward local cues, whereas cross-view correspondence is usually established the other way round, matching local land-

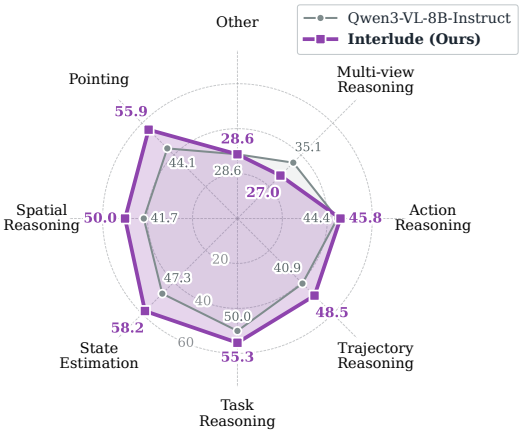


Figure 3: Performance of Interlude across eight subtasks of ERQA, improving over the baseline on six subtasks, matching it on Other, and declining on Multi-view Reasoning.

marks before recovering global geometry. That ordering is a property of how the supervision is built rather than of latent computation.

5.4 What the extra computation costs

Under identical settings, we measure FLOPs, wall-clock latency, and decoded tokens per answer (Figure 4). Relative to Qwen3-VL-8B-Instruct, Interlude requires 29.8% more FLOPs, increases latency by 67.3%, and generates 65.5% more tokens. Relative to Qwen3-VL-8B-Thinking, it uses 41.8% of the FLOPs, 7.9% of the wall-clock time, and 7.6% of the decoded tokens, while achieving a 3.9-point higher Overall score across the six benchmarks in Table 2, computed before rounding. The higher FLOP ratio compared with the decoded-token ratio is consistent with computation beyond token emission, including multimodal prefill and latent steps. Latent steps avoid emitting intermediate words but still require sequential model computation.

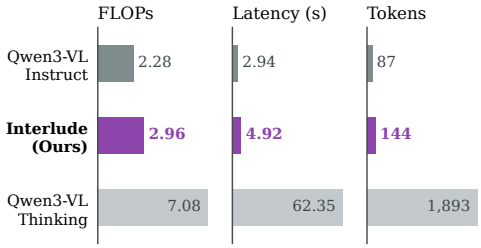


Figure 4: Inference cost. Interlude adds a modest overhead to Qwen3-VL-Instruct and costs far less than Qwen3-VL-Thinking.

6 Is the latent segment an intermediate modality?

Section 3 characterized the backbone and proposed empirical tests of visual alignment and representational spread. Here we repeat those measurements on the trained model to examine these hypotheses.

How the quantities are measured. For each analysis item, we compare three vector groups in the language model's shared representation space: $\mathbf{V}$, the projected visual-token embeddings; $\mathbf{T}$, the word embeddings obtained from the model's input embedding matrix; and $\mathbf{H}$, the hidden states at reasoning positions before training and within latent segments after training. We compute angles between ℓ_2-normalized group centroids, rather than individual tokens, using the arccosine of their inner products. This follows the formulation $\theta_m^{(s)} = \arccos(\mathbf{v}_m^{(s)\top}\mathbf{t}_m^{(s)})$ associated with Eq. 5, but the aggregation now covers the entire item rather than a single stage's concept set and retrieved visual tokens. The cosine distributions in Figure 6 are computed per-example rather than pooled globally. We use the same ERQA and VSI-Bench items before and after training to control for differences in the input sample set. Appendix B provides further geometric analysis.

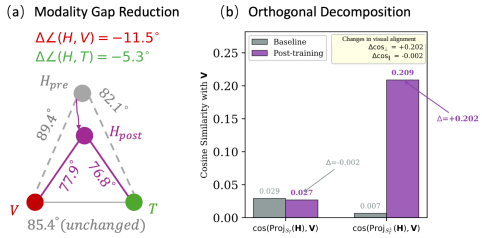


Figure 5: (a) Angles among the $\mathbf{V}$, $\mathbf{T}$ and $\mathbf{H}$ centroids before and after training. $\mathbf{H}$ moves toward both poles while the $\mathbf{T}$–$\mathbf{V}$ angle is unchanged. (b) Decomposition of $\mathbf{H}$ relative to the word-embedding principal subspace: visual similarity rises sharply for the orthogonal component and changes little within the subspace.

The states move toward both poles. Training pulls $\mathbf{H}$ toward both $\mathbf{V}$ and $\mathbf{T}$ and leaves the $\mathbf{T}$–$\mathbf{V}$ angle unchanged (Figure 5(a)). The per-example cosines show the same pattern (Figures 6): $\cos(\mathbf{H}, \mathbf{V})$ rises by 0.180 and $\cos(\mathbf{H}, \mathbf{T})$ by 0.087, while $\cos(\mathbf{T}, \mathbf{V})$ stays at 0.070. The states are therefore reshaped without disturbing the representations against which they are measured.

The unchanged $\cos(\mathbf{T}, \mathbf{V})$ also serves as a reference point. Before training, $\cos(\mathbf{H}, \mathbf{V}) = 0.008$ lies below $\cos(\mathbf{T}, \mathbf{V}) = 0.070$: the reasoning states were farther from the visual tokens than the word embeddings themselves. After training, both $\cos(\mathbf{H}, \mathbf{V}) = 0.188$ and $\cos(\mathbf{H}, \mathbf{T}) = 0.216$ exceed $\cos(\mathbf{T}, \mathbf{V})$, which is the pattern expected of a state positioned between the two poles rather than alongside either one. Relative to the backbone measurements in Table 1, the centroid angle

$\angle(\mathbf{H}, \mathbf{V})$ falls from $89.4°$ to $77.9°$ and the per-example cosine rises from 0.008 to 0.188; the near-orthogonality identified in Section 3 is thus reduced by training rather than being a fixed property of the backbone.

Visual alignment increases more in the orthogonal component. We next examine how visual alignment changes within and outside the dominant directions of word embeddings. Let $\mathcal{S}_T$ be the subspace spanned by the leading principal components of the input word embeddings $\mathbf{T}$ and $\mathcal{S}_T^{\perp}$ its orthogonal complement. This is an empirical word-embedding subspace, not a characterization of the language head's null space. We decompose

$$\mathbf{h} = \underbrace{\mathrm{Proj}_{\mathcal{S}_T}(\mathbf{h})}_{\text{in-subspace}} + \underbrace{\mathrm{Proj}_{\mathcal{S}_T^{\perp}}(\mathbf{h})}_{\text{orthogonal}}. \tag{9}$$

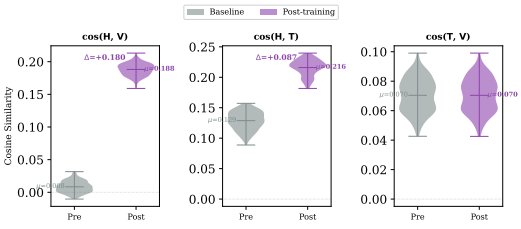


Figure 6: Per-example cosine similarities before and after training, with their means.

Before training, both components have low cosine similarity with $\mathbf{V}$. After training, the orthogonal component's similarity increases by $+0.202$, while the in-subspace component changes by -0.002 (Figure 5(b)). Thus, visual alignment increases much more strongly in the orthogonal component. These are separately normalized cosine measurements, so they do not quantify additive contributions to the full state's similarity. The contrast is consistent with our motivation for allowing latent states to depart from the dominant directions of word embeddings.

The states use more directions, with stronger visual alignment. The effective dimensionality of $\mathbf{H}$ rises from 13.3 to 56.6, more than quadrupling after training. The analyzed post-training directions align more strongly with the visual subspace $\mathcal{S}_V$ than with the word-embedding principal subspace $\mathcal{S}_T$ (Figure 7). This pattern is consistent with increased visual alignment across a broader set of hidden-state directions.

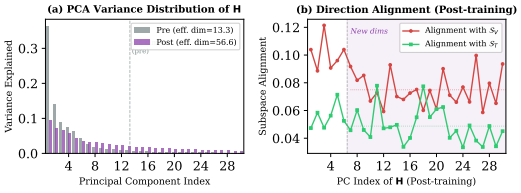


Figure 7: (a) PCA variance of $\mathbf{H}$ before and after training: the effective dimensionality rises from 13.3 to 56.6. (b) The analyzed post-training directions align more strongly with the visual subspace $\mathcal{S}_V$ than with the textual subspace $\mathcal{S}_T$.

Summary. The measurements show greater alignment with both modal references, a larger increase in visual similarity for the orthogonal component, and higher effective dimensionality after training. These patterns support the geometric motivation for an intermediate modality. The performance gains and ablations establish its usefulness on the evaluated tasks, while the geometric analysis describes how its representations differ from the backbone's. Isolating the contribution of individual subspaces would require direct interventions on those components.

7 Conclusion

We propose a multimodal reasoning paradigm that takes natural language as the explicit reasoning backbone while adaptively invoking latent reasoning at language bottlenecks. The explicit and latent reasoning processes complement each other, forming an interleaved reasoning flow. Based on this paradigm, we introduce **Interlude**, which uses carefully constructed training data and tailored supervision to shape latent segments into an intermediate modality between vision and language. This design gives the model continuous states for additional computation, with semantic and geometric supervision guiding how those states are used. Across the six evaluated benchmarks, Interlude consistently improves over Qwen3-VL-8B-Instruct and achieves the highest Overall score among the compared explicit and latent reasoning baselines. The ablations further support the contributions of latent computation and both supervision objectives to these gains.

Reproducibility Statement

The model, training objective and data-construction procedure are described in Section 4 and Appendix C; the teacher distribution is given in Appendix A; training and inference details are given in Appendix D; baselines and benchmarks are described in Appendices E and F. The evaluation protocol is stated in Appendix D. We will release the data synthesis pipeline, the training and inference code, the model implementation and the trained weights. An anonymized version of the implementation code is available at `https://anonymous.4open.science/r/Interlude-DEA4/`, and we commit to make the model weights publicly accessed upon publication.

The Use of Large Language Models

A large language model was used in three roles. First, it served as an object of study and a component of the method, synthesizing the 50K training traces, including the placement of latent segments and the stage concept sets, as described in Section 4.2 and Appendix C. Second, it supported the experiments as a coding assistant, helping write parts of the code, and every program was verified and further refined by the authors before use. Third, it assisted with language editing of the paper. The authors take full responsibility for the content of this paper.

References

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. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f8ee6a3d766b426d2618e555b5aeb39-Abstract-Conference.html.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Measuring and improving chain-of-thought reasoning in vision-language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 192–210, 2024b. URL https://aclanthology.org/2024.naacl-long.11/.

Jeffrey Cheng and Benjamin Van Durme. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*, 2024. URL https://arxiv.org/abs/2412.13171.

Tao Cheng, Shi-Zhe Chen, Hao Zhang, Yixin Qin, Jinwen Luo, and Zheng Wei. HyLaR: Hybrid latent reasoning with decoupled policy optimization. In *Computer Vision – ECCV 2026*, volume 17023 of *Lecture Notes in Computer Science*, pp. 613–630. Springer Nature Switzerland, September 2026. doi: 10.1007/978-3-032-37592-6_35. URL https://link.springer.com/chapter/10.1007/978-3-032-37592-6_35.

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. URL https://arxiv.org/abs/2311.01460.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-V: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9062–9072, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Dong_Insight-V_Exploring_Long-Chain_Visual_Reasoning_with_Multimodal_Large_Language_Models_CVPR_2025_paper.html.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 55–65, 2019. URL https://aclanthology.org/D19-1006/.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1907.12009.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini Robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*, 2025. URL https://arxiv.org/abs/2503.20020.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *International Conference on Learning Representations*, volume 2024, pp. 27896–27923, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/76917808731dae9e6d62c2a7a6afb542-Abstract-Conference.html.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.html.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Itxz7S4Ip3.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-R1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. URL https://arxiv.org/abs/2503.06749.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4198–4205, 2020. URL https://aclanthology.org/2020.acl-main.386/.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

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*, 2025. URL https://arxiv.org/abs/2509.24251.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/702f4db7543a7432431df588d57bc7c9-Abstract-Conference.html.

Lin Long, Changdae Oh, Seongheon Park, and Sharon Li. Understanding language prior of LVLMs by contrasting chain-of-embedding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=wnTnwWKu72.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. MM-Eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. URL https://arxiv.org/abs/2503.07365.

David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9(3):353–383, 1977. doi: 10.1016/0010-0285(77)90012-3. URL https://doi.org/10.1016/0010-0285(77)90012-3.

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. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0ff38d72a2e0aa6dbe42de83a17b2223-Abstract-Datasets_and_Benchmarks_Track.html.

Xuan Shen, Yizhou Wang, Yufa Zhou, Xiangxi Shi, Pu Zhao, Yanzhi Wang, and Jiuxiang Gu. Efficient reasoning with hidden thinking. *arXiv preprint arXiv:2501.19201*, 2025. URL https://arxiv.org/abs/2501.19201.

Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Ruihua Song, and Jian Luan. Think silently, think fast: Dynamic latent compression of LLM reasoning chains. In *Advances in Neural Information Processing Systems*, volume 38, pp. 4646–4668, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0706261aedab63814a2b73c32564b4c4-Abstract-Conference.html.

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*, December 2025. URL https://arxiv.org/abs/2512.16584.

Jintao Tong, Shilin Yan, Hongwei Xue, Xiaojun Tang, Kunyu Shi, Guannan Zhang, Ruixuan Li, and Yixiong Zou. SwimBird: Eliciting switchable reasoning mode in hybrid autoregressive MLLMs. *arXiv preprint arXiv:2602.06040*, February 2026. URL https://arxiv.org/abs/2602.06040.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. VL-Rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 30865–30891, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2c84844a559e4f962752570bff456ae4-Abstract-Conference.html.

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*, 2025b. URL https://arxiv.org/abs/2511.21395.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025c. URL https://arxiv.org/abs/2508.18265.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yubo Wang, Juntian Zhang, Yichen Wu, Yankai Lin, Nils Lukas, and Yuhan Liu. Forest before trees: Latent superposition for efficient visual reasoning. *arXiv preprint arXiv:2601.06803*, 2026. URL https://arxiv.org/abs/2601.06803.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

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 (CVPR)*, pp. 13084–13094, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.html.

xAI. Grok-1.5 Vision Preview. https://x.ai/news/grok-1.5v, April 2024. Published April 12, 2024. Introduces the RealWorldQA benchmark. Accessed: 2026-09-26.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2087–2098, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Xu_LLaVA-CoT_Let_Vision_Language_Models_Reason_Step-by-Step_ICCV_2025_paper.html.

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. URL https://arxiv.org/abs/2310.11441.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Thinking_in_Space_How_Multimodal_Large_Language_Models_See_Remember_CVPR_2025_paper.html.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-Onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2376–2385. IEEE, 2025b. URL https://openaccess.thecvf.com/content/ICCV2025/html/Yang_R1-Onevision_Advancing_Generalized_Multimodal_Reasoning_through_Cross-Modal_Formalization_ICCV_2025_paper.html.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkwZSG-CZ.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, YuXin Song, Haocheng Feng, Li Shen, and Dacheng Tao. Mulberry: Empowering MLLM with o1-like reasoning and reflection via collective Monte Carlo tree search. In *Advances in Neural Information Processing Systems*, volume 38, pp. 29918–29952,

2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2b0c4fc1bd52329279c51ad2ddec9ace-Abstract-Conference.html.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.

Juntian Zhang, Chuanqi Cheng, Yuhan Liu, Wei Liu, Jian Luan, and Rui Yan. Weaving context across images: Improving vision-language models through focus-centric visual chains. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27782–27798, 2025. URL https://aclanthology.org/2025.acl-long.1347/.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=y1pPWFVfvR.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WZH7099tgfM.

A Teacher Distribution Construction

For completeness, we describe the teacher distribution used in Eq. 4. For each latent state $\mathbf{h}_m^{(s)}$, we first gather the logits within the stage window $\Omega_m^{(s)}$ and construct a detached soft target:

$$\tilde{q}_m^{(s)} = \mathrm{sg}\left[\mathrm{softmax}\left(\ell_m^{(s)}|_{\Omega_m^{(s)}}/T\right)\right]. \tag{10}$$

As in Laser (Wang et al., 2026), we apply an entropy-triggered intervention when the restricted soft target is overly flat. Let $\bar{\mathcal{S}}_m^{(s)} = \mathcal{S}_m^{(s)} \setminus \mathcal{X}$. The prototype $y_m^{(s)}$ is uniform over $\bar{\mathcal{S}}_m^{(s)}$ and zero on $\Omega_m^{(s)} \setminus \bar{\mathcal{S}}_m^{(s)}$. The final teacher distribution is

$$q_m^{(s)} = \beta_m^{(s)} y_m^{(s)} + \left(1 - \beta_m^{(s)}\right) \tilde{q}_m^{(s)}, \tag{11}$$

where

$$\begin{aligned} \beta_m^{(s)} &= \alpha_{\mathrm{mix}} \mathbb{1}\left[\bar{H}(\tilde{q}_m^{(s)}) > \eta\right], \\ \bar{H}(q) &= -\frac{\sum_{k\in\Omega_m^{(s)}} q(k)\log q(k)}{\log|\Omega_m^{(s)}|}. \end{aligned} \tag{12}$$

If $\bar{\mathcal{S}}_m^{(s)}$ is empty, we set $\beta_m^{(s)} = 0$.

B Detailed Geometric Motivation

The information under-saturation hypothesis. For a latent state $\mathbf{h} = \mathbf{h}_m^{(s)}$, the language head produces next-token logits through a linear readout

$$\ell(\mathbf{h}) = W_U \mathbf{h}, \qquad W_U \in \mathbb{R}^{|\mathcal{V}|\times d}. \tag{13}$$

Thus, the language distribution observes $\mathbf{h}$ only through the linear readout defined by W_U. A perturbation Δ changes the logits by

$$\ell(\mathbf{h}+\Delta) - \ell(\mathbf{h}) = W_U \Delta. \tag{14}$$

Consequently, directions with small language-head sensitivity, i.e., small $\|W_U\Delta\|$ relative to $\|\Delta\|$, can carry additional information while only weakly perturbing the next-token distribution.

This view is consistent with two known geometric properties of language models. First, the linear-softmax readout imposes a low-rank bottleneck on the family of output distributions (Yang et al., 2018). Second, contextual language representations are often anisotropic and concentrate in narrow conic regions (Gao et al., 2019; Ethayarajh, 2019). Together, these observations suggest that the hidden state has representational capacity that is not fully saturated by the language-generation readout alone. We call this hypothesized capacity ***information under-saturation*** (IUS).

Geometric separation between visual and textual representations. A second ingredient is the geometric separation between visual and textual representations. Prior work on multimodal representation learning observes that even semantically aligned visual and textual embeddings may occupy separated regions or cones in the shared space (Liang et al., 2022). In a VLM, visual tokens are projected into the LM hidden space before being processed by the language model. After this projection, visual evidence need not coincide with the dominant directions used by textual representations. Instead, it can be represented along directions that are compatible with, but not identical to, the language manifold.

This separation is important for latent reasoning. If visual evidence were forced to collapse directly into discrete language tokens, fine-grained spatial or relational information could be lost. If it were represented as purely visual tokens, it could become difficult for subsequent language generation to use. This motivates seeking a transition region in which visual evidence remains available while the state is still accessible to the language pathway.

How the objectives complement each other. IUS and the modality gap motivate combining semantic and geometric supervision. A latent segment $\mathbf{H}_m = \{\mathbf{h}_m^{(1)}, \ldots, \mathbf{h}_m^{(k_m)}\}$ is trained to support subsequent textual reasoning while retaining useful visual cues. The two losses guide different aspects of this representation.

$\mathcal{L}_{\text{sg}}$ supervises the language-head readout of $\mathbf{h}_m^{(s)}$, encouraging probability mass on the stage's relevant concepts. Subsequent text accesses the latent states through self-attention, so this readout supervision serves as an auxiliary semantic objective. $\mathcal{L}_{\text{imc}}$ instead encourages the state to have a smaller cosine distance to an intermediate reference than to either modal centroid. It does not require the state to coincide with that reference or fix its direction outside the modal plane, although normalization couples the size of that component to the loss.

This design gives the model a semantic target and a geometric preference while allowing latent states to depart from either modal anchor. The ablations support using both objectives, and the geometric measurements show a larger increase in visual alignment outside the word-embedding principal subspace. These observations are consistent with the proposed design; they do not establish that the losses act on independent subspaces or that the added visual alignment is invisible to the language head.

C Data Details

C.1 Constructing interleaved reasoning traces

We construct 50K synthetic training examples from VCR and VIRL-39K images using Claude-4.7-opus as the teacher model. For each image-question pair, the teacher generates a visually grounded reasoning trace and a final answer. We discard traces with incorrect final answers before constructing latent supervision.

Within each retained trace, we identify steps that require continuous visual judgments, such as reconstructing spatial layout, combining gaze directions, resolving occlusion, or anticipating physical changes. These steps become latent segments interleaved with the verbal reasoning. The preceding text specifies the visual subproblem, while the following text uses its result to advance toward the answer. This organization preserves an explicit reasoning scaffold while allocating difficult visual judgments to hidden-state computation.

C.2 Stage-wise concept supervision

Each example consists of an image, a question, a final answer, and a reasoning trace whose latent boundaries are paired with stage annotations. The m-th boundary corresponds to the m-th ordered stage list. Each latent segment is annotated with an ordered sequence of concept sets $\{\mathcal{S}_m^{(1)}, \ldots, \mathcal{S}_m^{(k_m)}\}$. Here, $k_m \in \{1, 2, 3\}$ is the number of stages in the segment. Early stages identify the scene and relevant entities; later stages focus on local attributes and relations that determine the answer. An intermediate stage can bridge these levels, while a single-stage segment can combine context and local evidence. The progression therefore specifies how a visual judgment should be refined without prescribing a unique verbal explanation.

Each concept set supervises one recurrent step. The concepts define the semantic windows and textual anchors used by the two latent objectives; they are neither appended to the input prompt nor emitted as an additional reasoning trace. During training, the number of annotated stages determines the length of each segment. At inference time, the model receives neither the concept annotations nor their stage counts.

Training traces contain 2–5 latent triggers, with an average of 2.95 per example (Figure 8). Each trigger begins a visual subproblem, whereas its stages describe the successive refinements within that subproblem.

C.3 Illustrative example

Figure 9 illustrates this structure with a question about the focus of attention of a seated audience. The first latent segment organizes head orientations and gaze directions. Its concept sets move from

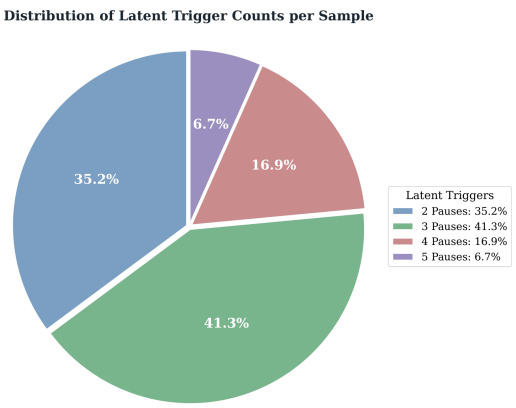


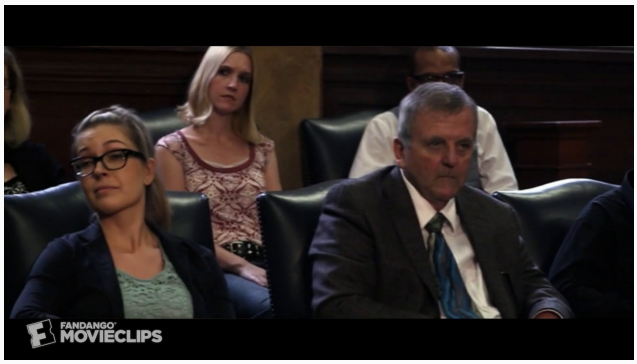
Figure 8: Distribution of latent triggers per training example. A trigger opens a latent segment whose internal stages refine a visual judgment from scene-level context toward answer-relevant evidence.

the audience and individual heads to their directional alignment. The second segment combines these sightlines to estimate the location and height of their intersection, progressing from individual gaze lines to an off-camera target at approximately seated eye level. Each segment has two stages, giving four supervised recurrent steps across two triggers. The intervening text links the two visual judgments, and the final text expresses the resulting spatial conclusion.

Stage-wise Supervision of Interleaved Latent Reasoning

Task Type: physical

Question: Based on the heads and gazes of the people seated, where is the focal point of their attention located relative to the camera?



Reasoning with Latent Thought Triggers:

To work out where the audience is looking I first need the orientation of each visible head; let each one cast a forward gaze ray r_i. `<|latent|>`

Latent Thought 1: Head Orientation and Gaze Direction

Abstract	courtroom gallery; seated audience rows; front woman head; front man head; back blonde woman head.
Concrete	back bespectacled man head; head yaw left; gaze direction left; off-axis alignment.

Those four head-rays all swing off the camera axis toward the same side of the frame, none of them facing the lens and none drifting to the opposite wall. With that bundle of rays in mind, I now ask where they intersect and at what vertical level, since a slight downward or level pitch tells me whether the target is elevated like a bench or low like a seated witness. `<|latent|>`

Latent Thought 2: Sightline Convergence and Target Location

Abstract	courtroom gallery; seated audience rows; front man gaze line; front woman gaze line.
Concrete	back woman gaze line; converging sightlines; off-frame right direction; seated eye-level height.

Following those rays outward, they meet just past the right edge of the frame at roughly seated eye height, consistent with a speaker or witness positioned off-camera to that side rather than an elevated judge.

Answer: Off the camera's right side, roughly level with the seated viewers.

Figure 9: An illustrative example of stage-wise concept supervision. The natural-language reasoning trace contains two adaptive latent thought triggers. Each trigger is paired with abstract and concrete latent key tokens that supervise the model to preserve intermediate visual cues and refine them into task-relevant spatial evidence.

D Training and Inference Details

D.1 Text and boundary supervision

The training sequence interleaves verbal reasoning with latent-segment boundaries and ends with the final answer. Textual cross-entropy supervises the assistant response; the question, system prompt, image positions, and padding do not contribute to this loss. The opening boundary `<|latent|>` is supervised within the text sequence, and the final state of each latent segment is trained to predict the closing boundary `<|/latent|>`. These targets teach the model when to enter hidden-state computation and when to resume verbal reasoning.

To balance a short answer against a longer reasoning trace, we assign unit weight to reasoning tokens and weight

$$w_{\text{ans}} = \min\left(10, \max\left(1, \frac{L_{\text{reason}}}{L_{\text{answer}}}\right)\right) \tag{15}$$

to answer tokens. Here, L_{reason} and L_{answer} denote the token counts of the reasoning and answer portions; the latter includes the answer separator and end-of-turn token. The text loss is normalized by the sum of the supervised token weights. A unit-weight closing-boundary loss is added for each latent segment. Together, these terms form $\mathcal{L}_{\text{CE}}$.

D.2 Latent computation and supervision

Each latent step reuses the complete language-model stack of Qwen3-VL-8B-Instruct. A learned step embedding distinguishes successive iterations. The resulting states are normalized with RMSNorm and supplied as a prefix to the next text segment, allowing subsequent reasoning to attend to the full latent trajectory. Training unrolls each segment for $k_m \in \{1, 2, 3\}$ steps, as specified by its concept annotations. The initial recurrent state is the hidden state at the opening boundary, and the preceding context supplies the attention cache.

The preceding context is detached between text segments, while gradients from the subsequent text propagate through the injected latent states. Visual encoding is kept fixed. The language-model stack, language head, step embeddings, and latent-output normalization are optimized, coupling latent supervision with the ability to use the resulting states in later reasoning.

For semantic guidance, the student distribution is normalized over the full vocabulary, while the detached soft target is restricted to the current and future concept sets. The target construction is given in Appendix A. For the intermediate-modality constraint, the textual concept centroid retrieves the six visual tokens with highest cosine similarity; their mean forms the visual anchor. Textual and visual anchors are detached so that the geometric objective shapes the latent state relative to these references.

Table 3 summarizes the latent-supervision parameters. The SLERP coefficient decreases from abstract to concrete stages, moving the intermediate reference from the textual side toward the visual

side. Both latent losses are averaged over supervised stages within each segment and summed over segments. They enter the total objective with unit coefficients.

Table 3: Parameters of the latent-supervision objectives. The SLERP coefficient τ is defined in Eq. 6.

Parameter	Value
Soft-target temperature T	1.0
Normalized-entropy threshold η	0.6
Entropy-triggered mixture α_{mix}	0.7
Retrieved visual tokens	6
Geometric margin δ	0.05
SLERP τ: abstract / bridge	0.75 / 0.50
SLERP τ: unified / concrete	0.50 / 0.25

D.3 Inference and evaluation

Generating `<|latent|>` initiates recurrent latent computation. A segment ends when the language head predicts `<|/latent|>`, the cosine similarity of successive states exceeds 0.99, or three recurrent steps have been completed. The latent states occupy context positions without being decoded into words, and the following text attends to them through ordinary self-attention. Concept annotations are used only for training.

Training uses 16 H20 GPUs. Benchmark evaluation uses greedy decoding. Generation ends at the end-of-sequence token or the generation budget, and the final answer is read from the answer segment. Representation-analysis examples are drawn from ERQA and VSI-Bench.

E Baseline Details

We compare Interlude with general-purpose VLMs, reasoning-enhanced VLMs, and latent reasoning methods.

(1) General-purpose VLM baselines.

- **Qwen2.5-VL-7B-Instruct** (Bai et al., 2025b) is a strong open-source vision-language model with broad multimodal understanding capabilities, including visual recognition, object localization, document understanding, chart reasoning, and video understanding.
- **Qwen3-VL-8B-Instruct** (Bai et al., 2025a) is used as the main reference model in our experiments. It provides a stronger multimodal backbone and supports interleaved image, text, and video contexts.
- **Qwen3-VL-8B-Thinking** is the reasoning-oriented variant of Qwen3-VL (Bai et al., 2025a). It improves complex reasoning mainly by extending explicit reasoning traces, making it a natural comparison for Interlude, which instead introduces compact latent segments.
- **InternVL3.5-8B** (Wang et al., 2025c) is included as another competitive open-source VLM baseline with strong general multimodal ability, reasoning capability, and inference efficiency.

(2) Reasoning-enhanced VLM baselines.

We further compare with VLMs explicitly optimized for reasoning.

- **VL-Rethinker** (Wang et al., 2025a) enhances the slow-thinking ability of VLMs through reinforcement learning and explicit self-reflection, using mechanisms such as selective sample replay and forced rethinking.
- **Vision-R1** (Huang et al., 2025) is a reasoning-oriented multimodal model trained with cold-start multimodal chain-of-thought data and reinforcement learning to encourage stronger reasoning behaviors.

These comparisons test whether explicit textual reflection and RL-induced reasoning are sufficient for the visual reasoning tasks considered in this work.

(3) Latent reasoning baselines.

The most direct competitors to Interlude are latent reasoning methods, which move part of the reasoning process from natural-language tokens to hidden or visual-latent representations.

- **Monet** (Wang et al., 2025b) enables MLLMs to reason in latent visual space by generating continuous embeddings that function as intermediate visual thoughts.
- **LVR** (Li et al., 2025) performs latent visual reasoning by training the model to generate latent states that reconstruct answer-critical visual tokens in the visual embedding space.
- **Laser** (Wang et al., 2026) aligns latent states with dynamic semantic windows, allowing the model to preserve a superposition of plausible future semantics before committing to more specific visual judgments.
- **SwimBird** (Tong et al., 2026), based on Qwen3-VL-8B, switches among text-only, latent-visual, and interleaved reasoning modes.
- **SkiLa** (Tong et al., 2025), based on Qwen2.5-VL-7B, interleaves textual reasoning with continuous visual sketches trained through semantic reconstruction.
- **HyLaR-7B** (Cheng et al., 2026), based on Qwen2.5-VL-7B, combines textual and continuous visual reasoning using Canvas-in-Latents and decoupled policy optimization.

These baselines cover explicit, latent, and hybrid reasoning. Interlude combines semantic guidance with a geometric constraint that positions latent states relative to textual and visual anchors without requiring them to lie on the arc joining the two.

F Benchmark Details

We evaluate Interlude on six complementary multimodal benchmarks. Together, these benchmarks cover embodied reasoning, visual-spatial understanding, real-world visual grounding, hallucination-sensitive reasoning, general multimodal capability, and fine-grained visual understanding. Unless otherwise specified, we follow the official evaluation protocol of each benchmark and report the corresponding percentage score. The Overall score in Table 2 is the unweighted average across all six benchmarks.

ERQA. ERQA, short for Embodied Reasoning Question Answering, is a multimodal benchmark for evaluating embodied reasoning capabilities in realistic physical scenarios (Gemini Robotics Team et al., 2025). It contains interleaved image-text multiple-choice questions that require models to reason about spatial relations, object states, world knowledge, and robot-relevant task progress. We include ERQA because it directly tests whether a VLM can ground its reasoning in visual evidence when answering questions about physical scenes. In our fine-grained analysis, ERQA is further divided into subtasks such as Pointing, Spatial Reasoning, State Estimation, Task Reasoning, Trajectory Reasoning, Action Reasoning, Multi-view Reasoning, and Other, allowing us to examine where latent reasoning is most beneficial.

VSI-Bench. VSI-Bench is a video-based visual-spatial intelligence benchmark designed to test whether multimodal models can see, remember, and recall spaces from sequential visual observations (Yang et al., 2025a). It contains more than 5,000 question-answer pairs about real indoor scenes and covers configurational, measurement-estimation, and spatiotemporal reasoning tasks. VSI-Bench is particularly relevant to Interlude because many of its questions require continuous spatial reasoning, scene memory, and viewpoint-aware understanding, which may be difficult to express faithfully using only discrete textual chain-of-thought.

RealWorldQA. RealWorldQA evaluates practical real-world visual understanding and spatial grounding (xAI, 2024). The benchmark consists of natural images from vehicles and other everyday scenarios, each paired with a question and an easily verifiable answer. Its questions often involve

object counting, direction, relative position, scene layout, and commonsense physical interpretation. We use RealWorldQA to assess whether Interlude improves grounding in ordinary real-world images rather than only in synthetic or highly curated settings.

HallusionBench. HallusionBench is a diagnostic benchmark for studying entangled language hallucination and visual illusion in large vision-language models (Guan et al., 2024). It contains human-crafted image-context reasoning questions with control structures designed to reveal whether a model is answering according to the visual evidence or relying on misleading language priors. This benchmark is central to our evaluation because Interlude is designed to reduce visually ungrounded reasoning by allowing difficult visual computations to be performed in hidden space before returning to natural language.

MMStar. MMStar is a multimodal benchmark that requires visual information consisting of 1,500 human-selected samples that are designed to reduce text-only shortcuts and data leakage (Chen et al., 2024a). It evaluates six core multimodal capabilities and eighteen fine-grained dimensions, covering a broad range of perception and reasoning skills. We include MMStar as a general multimodal benchmark to test whether Interlude improves overall VLM capability beyond task-specific spatial or hallucination-sensitive settings.

V*Bench. V*Bench evaluates fine-grained visual understanding in high-resolution, visually complex images (Wu & Xie, 2024). It contains 191 images from SA-1B and covers attribute recognition and spatial relationship reasoning. The questions require locating small, task-relevant details within a larger scene, complementing the other benchmarks' tests of visual grounding. Table 2 reports overall V*Bench accuracy.

G The Interlude Reasoning Process: A Case Study

Figure 10 illustrates how Interlude handles a height-based visual reasoning problem. The question cannot be answered by color recognition alone: the model must first associate each colored circle with a shelf level, estimate its vertical position relative to the floor, and then compare it with the 4-foot threshold. Interlude allocates these operations to different reasoning media. The textual trace first states the high-level plan, while latent segments are inserted at the key visual bottlenecks corresponding to the top, middle, and bottom shelf regions. These latent segments preserve continuous visual evidence such as shelf height, perspective, and floor reference, which would be difficult to express faithfully as discrete tokens. The following textual segments then use the latent segments to make interpretable judgments: the red circle on the upper shelf and the yellow circle on the middle shelf are above the threshold, whereas the green and blue circles on the lower shelf are not. This leads to the correct answer, C.

This example highlights the central advantage of Interlude. Rather than producing a longer textual chain of thought, the model separates readable problem decomposition from visually grounded hidden-state computation. The language segments provide an interpretable scaffold, while the latent segments carry the continuous spatial reasoning needed for the final decision.

The Interlude Interleaved Reasoning Process: A Case Study

Question: What color circles correspond to points higher than 4 feet?

Choices:
A. Red.
B. Green, yellow, and red.
C. Yellow and red.
D. None of the above.

Please answer directly with only the letter of the correct option and nothing else.



Reasoning:
I need to identify the vertical position of each colored circle on the shelving unit and estimate its height relative to the floor. **<|latent|>** shelf top **<|/latent|>** The red circle sits on the uppermost shelf, which is clearly above the 4-foot threshold. **<|latent|>**shelf middle **<|/latent|>** The yellow circle rests on the middle shelf, and judging from the visible shelf depth and the floor reference, this shelf is also well above 4 feet. **<|latent|>** shelf bottom **<|/latent|>** The green and blue circles are on the bottom shelf area, which sits well below 4 feet, so they do not qualify. **<|latent|><|/latent|>** Since only the red and yellow circles lie higher than 4 feet, the answer is C.

Final Answer: C
Correct Answer: C

Figure 10: Case study of Interlude on height-based visual reasoning. The task requires the model to identify the shelf level of each colored circle and compare its height with a 4-foot threshold. Interlude uses natural-language text to organize the reasoning process, while inserting latent segments at visual bottlenecks such as the top, middle, and bottom shelf regions. The subsequent textual segments use these latent segments to conclude that the red and yellow circles are above the threshold, producing the correct answer C. The **red latent markers** are visualization labels for latent segments rather than decoded natural-language tokens.