ER-LVR: RESIDUAL-ADAPTIVE DISPERSION FOR LATENT VISUAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent visual reasoning lets a vision–language model reason over visual evidence without first translating it into words: at designated positions it outputs a continuous hidden state in place of a token and feeds it back as the next input embedding. Existing methods train these visual thoughts by regressing each such state onto a target embedding from the model's own frozen vision encoder. In practice the alignment is imperfect, and the error that remains reflects how well a particular visual thought was formed, yet a single vector carries no trace of it: a visual thought that missed its target re-enters the model's reasoning looking exactly like one that hit. We show that this is a property of the objective. Squared-error alignment is the degenerate boundary of the energy-score family, at which the score constrains only the mean of the predicted distribution; in the strictly proper interior the scale is constrained as well, and at the optimum it equals the distance from the distribution's mean to the target, so the spread of a visual thought tracks its own alignment error. We instantiate this as ER-LVR, which replaces the deterministic alignment head with a reparameterized stochastic head trained under an energy objective, and verify that the learned geometry converges to the predicted landmark while two control objectives sharing the same head do not. On spatial planning ER-LVR exceeds the strongest prior latent method by 2.3 points, and by 3.2 with policy optimization of the text policy; it matches or exceeds every latent baseline on out-of-domain visual understanding, and improves joint accuracy on paired semantic queries over a same-backbone baseline.

1 INTRODUCTION

Reasoning in vision–language models has largely been organized around two strategies. The first reasons in natural language, describing the image and then chaining textual steps, and has been scaled through reinforcement learning with verifiable rewards (Xiao et al., 2026). The second lets the model act on the image while it reasons, cropping, zooming, or executing code to bring evidence back into context. Reinforcement learning has been used to induce zoom-in behaviour without supervision (Zheng et al., 2026), to prevent models from bypassing a nascent visual tool (Wang et al., 2025), to learn adaptive visual search (Zhang et al., 2025b), and to coordinate detectors and drawing operations under a single controller (Su et al., 2025; Wu et al., 2026); related work substitutes executable code for a fixed tool interface (Zhao et al., 2025) or interleaves region recognition with refinement (Jiang et al., 2025). These methods are effective, but every visual operation is a round trip through an external interface, which costs latency and constrains what the model can look at.

A third strategy keeps the reasoning inside the model. In latent visual reasoning the model outputs a continuous hidden state at designated positions in place of a token and feeds it back as the next input embedding, forming a chain of visual thoughts in the shared vision–text embedding space (Li et al., 2026a; Yang et al., 2026). The idea has developed quickly, and variants differ mainly in what the target encodes and how the span is scheduled: reasoning over latent visual states beyond both images and language (Wang et al., 2026a), contrastive rather than hard alignment (Ding et al., 2026), region-level attribute supervision (Xu et al., 2026a), auxiliary expert heads (Qin et al., 2025), dynamically interleaved latent and textual steps (Liu et al., 2025; Ma et al., 2025), parallel rather than autoregressive generation (Lu et al., 2026), and compression or superposition for efficiency (Wang et al., 2026b). Others plan directly through image sequences (Xu et al., 2026b) or internalize manipulations that would otherwise require a tool (Zhang et al., 2025a; Li et al., 2025).

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.

Across this family the training signal is close to uniform: each hidden state produced in the span is regressed onto a target region embedding from the model's own frozen vision encoder, under a mean-squared error. In practice the regression does not close, and in our runs the error plateaus at a strictly positive value. The gap that remains is informative, because it reflects how faithfully a particular visual thought was formed, yet a single vector cannot express it. Each visual thought is therefore passed back into the model's own reasoning with no indication of whether it landed near its target or far from it, and one that missed is indistinguishable from one that hit. The latent may still encode its region accurately; what it cannot do is report how accurately. This is a narrower failure than the latents being uninformative, and it is consistent with analyses finding that they can be removed or corrupted with little change in accuracy (Viveiros et al., 2026; Li et al., 2026b; Guo et al., 2026).

We propose ER-LVR (Figure 1), which trains each visual thought as a distribution of embeddings rather than a single vector. Each visual thought is drawn from a reparameterized stochastic head, and training scores the samples against the target with an energy objective. Squared-error alignment is the degenerate boundary of the energy-score family, the exponent at which the score constrains only the mean of the predicted distribution and is indifferent to its spread (Proposition 1); in the strictly proper interior the scale is constrained as well, and the minimizing scale equals the distance from the distribution's mean to the target (Proposition 2), so the spread of a visual thought tracks its own alignment error.

Whether a verifiable reward can improve models of this kind is unsettled. Policy optimization applied to the latent generator has been reported not to improve, and sometimes to degrade, latent quality (Wang et al., 2026a; Lu et al., 2026; Xu et al., 2026a); under standard reinforcement learning the fraction of responses that use the latent path has been observed to fall toward zero (Zhu et al., 2026); and the estimator is unstable when the action is continuous and high-dimensional (Deng et al., 2026). We find that the picture separates by which parameters the reward is allowed to touch. Optimizing the text policy improves a supervised ER-LVR model by 1.9 points on maze planning, consistently across seeds, once the reward discriminates partial progress and gradients are accumulated across problems. Continuing to train the noise tilt under the energy objective during this stage adds 0.17 points

Our contributions are as follows.

- **ER-LVR.** Squared-error alignment is the non-strictly-proper boundary of the energy score, constraining only the mean of the latent distribution and leaving the alignment residual unexpressed (Proposition 1), while a strictly proper member of the same family fixes that scale to the residual (Propositions 2 and 3). We instantiate this as ER-LVR, which draws each visual thought from a reparameterized stochastic head trained under the energy objective and uses that head on the decode path, so the mechanism trained is the mechanism deployed (§2).
- **Policy optimization.** A verifiable reward improves a supervised ER-LVR model when applied to the text policy; continuing to train the noise tilt alongside it adds nothing stable, and the reward gradient into the head carries no consistent direction across rollouts (§2.6).
- **Evaluation and analysis.** ER-LVR exceeds the strongest prior latent method on spatial planning, matches or exceeds every latent baseline out of domain, and improves joint accuracy on paired semantic queries over a same-backbone baseline. Geometric diagnostics confirm that the learned spread converges to the predicted landmark while two control objectives sharing the same head do not (§4).

2 METHOD

2.1 BACKGROUND: LATENT VISUAL REASONING

A vision–language model π_θ encodes an image $\boldsymbol{x}$ with a frozen vision tower V and a frozen multimodal projector Π into visual tokens sharing a space with the text embeddings, then decodes a response to a question q. Latent visual reasoning extends this loop (Li et al., 2026a; Yang et al., 2026): the response carries a *latent span* delimited by two special tokens, so that

$$\boldsymbol{y} = \big(y_{1:s},\ \langle\text{LVR}\rangle,\ \boldsymbol{z}_1, \ldots, \boldsymbol{z}_B,\ \langle/\text{LVR}\rangle,\ y_{s+1:T}\big), \tag{1}$$

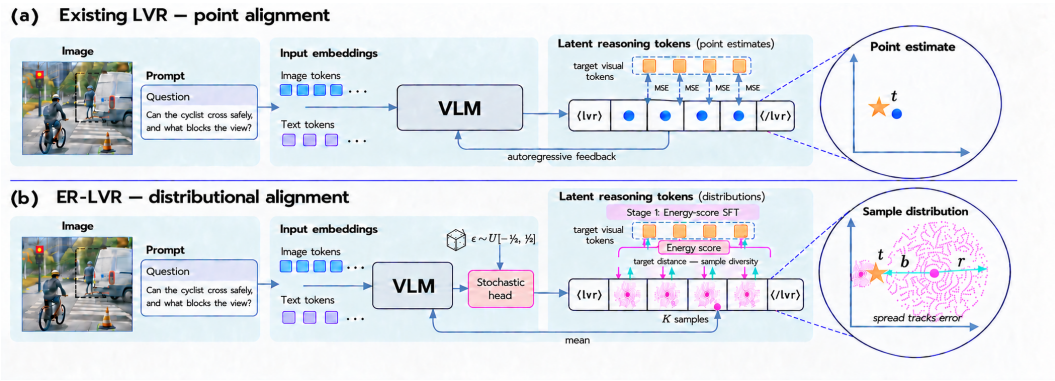


Figure 1: **ER-LVR.** (a) Existing latent visual reasoning aligns each latent to a pooled target visual token under a squared error and feeds back a single vector, so a latent that missed its target is indistinguishable downstream from one that hit it. (b) ER-LVR draws each visual thought from a stochastic head and trains the K samples under an energy score; the spread of the learned distribution tracks its distance to the target ($r^\star = b$), and the mean of the draws is fed back.

where the model itself generates $\langle$LVR$\rangle$ after reading q and decoding then switches mode for a fixed budget of B steps. In latent mode no token is sampled: at step i the backbone hidden state $\boldsymbol{h}_i \in \mathbb{R}^d$ is mapped to a vector $\boldsymbol{z}_i$ which is written into the input embedding of the next position, $\boldsymbol{e}_{i+1} = \boldsymbol{z}_i$, in place of a token embedding. After B steps $\langle$/LVR$\rangle$ is generated and ordinary decoding resumes conditioned on the span. The budget is the same at training and at inference.

The vectors inside the span are never observed, so training requires a target. Given a dataset annotated with a region of interest $\mathrm{ROI}(\boldsymbol{x}, q)$, the frozen encoders supply

$$\boldsymbol{t}_i = \mathrm{pool}_i\big(\Pi(V(\boldsymbol{x}))\big[\mathrm{ROI}\big]\big) \in \mathbb{R}^m, \qquad \mathcal{L}_{\text{align}} = \sum_{i=1}^{B} \|\boldsymbol{z}_i - \boldsymbol{t}_i\|^2, \tag{2}$$

with m the target dimension, used in the normalization of equation 6. The head outputs directly into the target space, so no separate projection is required. Since V and Π are frozen, $\boldsymbol{t}_i$ is fixed. The whole image is embedded once; the token positions falling inside the ROI box are selected and average-pooled to the step budget. Unless noted we pool to a single target shared across all B steps, so $\boldsymbol{t}_1 = \cdots = \boldsymbol{t}_B$; a per-step variant pooling to B distinct targets is implemented and ablated, matching the shared target on V* (82.20) and scoring within a point of it on MMVP. Region annotations enter only through equation 2: at inference no box is available, no target is formed, and the span is generated from the model's own hidden states.

2.2 Setup

Stochastic latent head. In place of a deterministic map from $\boldsymbol{h}_i$ to the fed-back vector (Figure 1a) we use a reparameterized head g_ψ (Figure 1b),

$$\boldsymbol{z}_i = g_\psi(\boldsymbol{h}_i, \boldsymbol{\varepsilon}), \qquad \boldsymbol{\varepsilon} \sim \mathcal{U}[-\tfrac{1}{2}, \tfrac{1}{2}]^{d_{\boldsymbol{\varepsilon}}}, \tag{3}$$

a stack of residual gated-MLP blocks that refine exogenous noise conditioned on $\boldsymbol{h}_i$ in a single forward pass (Shao et al., 2025). The noise carries its own dimension $d_{\boldsymbol{\varepsilon}}$, independent of the hidden size d and the target size m, and enters through a learned projection; a multiplier σ scales it, held fixed at σ=1 during training and used as the only inference-time control (§2.7). Drawing K noise vectors yields samples $\{\boldsymbol{z}_i^{(k)}\}_{k=1}^{K}$ from an implicit conditional distribution $P_\psi(\cdot \mid \boldsymbol{h}_i)$. Reparameterization makes expectations under P_ψ differentiable in ψ by the pathwise estimator, so training requires no score-function term.

2.3 THE ENERGY SCORE AND ITS MEAN-MATCHING BOUNDARY

Definition 1 (Energy score). *For a predictive distribution P on $\mathbb{R}^m$, an observation $\boldsymbol{y}$, and $\alpha \in (0, 2)$,*

$$S_\alpha(P, \boldsymbol{y}) = \mathbb{E}_{\boldsymbol{u},\boldsymbol{u}'\sim P} \|\boldsymbol{u} - \boldsymbol{u}'\|^\alpha - 2\,\mathbb{E}_{\boldsymbol{u}\sim P} \|\boldsymbol{u} - \boldsymbol{y}\|^\alpha . \tag{4}$$

The first term rewards intra-sample spread and the second draws samples toward $\boldsymbol{y}$. S_α is strictly proper for $\alpha \in (0, 2)$ (Gneiting & Raftery, 2007), and degenerates at the upper endpoint.

Proposition 1 (Mean-matching squared error is the $\alpha = 2$ boundary). *At $\alpha = 2$, $-S_2(P, \boldsymbol{t}) = 2\,\|\mathbb{E}_P[\boldsymbol{z}] - \boldsymbol{t}\|^2$, which depends on P only through its mean and is therefore indifferent to the spread of P. The per-sample form $\frac{1}{K}\sum_k \left\|\boldsymbol{z}^{(k)} - \boldsymbol{t}\right\|^2$ is a distinct objective, strictly increasing in the spread. Neither selects a finite non-zero scale.*

Proof sketch. The variance terms cancel on substitution into equation 4; Appendix A. □

Empirical behaviour of the boundary objective. Trained at the boundary, the alignment error settles within roughly one hundred optimizer steps at a strictly positive floor, around 9% of the target norm in our runs, and converges to a single vector per position.

2.4 SCALE SELECTION IN THE STRICTLY PROPER INTERIOR

Away from the boundary the objective constrains the scale of P as well as its location. Writing $b = \|\boldsymbol{\mu} - \boldsymbol{t}\|$ for the distance from the centre of P to the target, the following holds.

Proposition 2 (Fixed-centre scale optimum). *Let P be isotropic with centre $\boldsymbol{\mu}$ and radius r. Holding $\boldsymbol{\mu}$ fixed, and in the high-dimensional concentration regime, the objective for $\alpha \in (0, 2)$ is*

$$\mathcal{E}_\alpha(r) = 2\big(b^2 + r^2\big)^{\alpha/2} - 2^{\alpha/2} r^\alpha, \tag{5}$$

with unique interior minimum $r^\star = b$, independently of α.

Proof sketch. Setting $\mathcal{E}'_\alpha(r) = 0$ and writing $u = r/b$ gives a scale-free condition whose left side is strictly increasing for $\alpha < 2$, so its unique root is $u = 1$; at $\alpha = 2$ it degenerates. Appendix A. □

Proposition 3 (Deterministic target). *For $\alpha \in (0, 2)$ and a deterministic target, $-S_\alpha(P, \boldsymbol{t}) \geq 0$ over all P, with equality if and only if $P = \delta_{\boldsymbol{t}}$.*

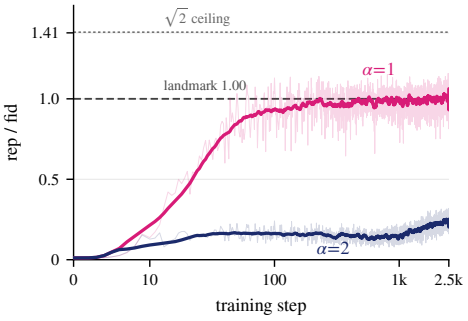


Figure 2: Measured spread ratio during training, VSP lineage, 8 latent steps, seed 42; both arms zero-initialised at 0.011. Faint traces are per-batch values, bold a 31-step centred mean. The strictly proper arm rises through the fixed-centre landmark by step 100 and fluctuates about it thereafter (final 1.06); the α=2 boundary arm never acquires an attractor, drifting in a 0.15–0.32 band, as Proposition 1 predicts. The trajectory, not either endpoint, is the evidence.

Together the two propositions describe the objective's behaviour at two levels. Over all distributions the optimum for a deterministic target is the point mass, so the selected scale is not a fixed quantity; conditional on a centre at distance b from the target, the selected scale is b itself. The scale therefore tracks the alignment error, sharpening as the centre improves. We refer to this as residual-adaptive dispersion.

Trajectory. We track the spread ratio rep/fid, where $\text{fid} = \mathbb{E}\,\|\boldsymbol{z} - \boldsymbol{t}\|$ and $\text{rep} = \mathbb{E}\,\|\boldsymbol{z} - \boldsymbol{z}'\|$, which equals 1 at $r = b$ (§4). Two predictions follow. The spread should track b over training rather than decay to zero or grow without bound, and, starting from a head with negligible output spread, should rise before following b downward. Figure 2 shows both. Since $r^\star$ does not depend on α, varying the exponent probes optimization rather than the location of the optimum.

Per-sample squared error is monotone in r and drives the spread to zero, recovering a deterministic head, while mean-matching squared error is flat in r by Proposition 1; only the strictly proper interior has an interior optimum in the scale.

2.5 OBJECTIVE

For latent step i with K samples we use the unbiased K-sample estimator of $-S_\alpha$, normalized per dimension:

$$\mathcal{L}_{\text{en}}(i) = \frac{2}{K}\sum_k \frac{\left\|\boldsymbol{z}_i^{(k)} - \boldsymbol{t}_i\right\|^\alpha}{m^{\alpha/2}} - \frac{1}{K(K-1)}\sum_{k\neq j}\frac{\left\|\boldsymbol{z}_i^{(k)} - \boldsymbol{z}_i^{(j)}\right\|^\alpha}{m^{\alpha/2}}, \tag{6}$$

with $\alpha = 1$. The divisor $m^{\alpha/2}$ converts Euclidean distances into per-dimension root-mean-square distances, so the magnitude of the objective does not scale with hidden size and a single weight is comparable across backbones and against a mean-reduced squared-error baseline. Training minimizes $\mathcal{L} = \mathcal{L}_{\text{NTP}} + \beta\sum_i \mathcal{L}_{\text{en}}(i)$, with $\mathcal{L}_{\text{NTP}}$ the next-token loss on the answer and gradients reaching $\{\theta, \psi\}$ pathwise through equation 3; Algorithm 1 gives the step. Exponents below one induce gradient explosion in practice (Shao et al., 2025).

2.6 POLICY OPTIMIZATION: ENERGY-GRPO

Starting from a supervised checkpoint we optimize with GRPO (Shao et al., 2024) on a planning task with checkable answers. In this stage the model has two trainable components, the text policy π_θ and a tilt ϕ on the head's noise law, with the head g_ψ itself held fixed; the arms differ in whether ϕ is updated. The K samples entering equation 6 are then drawn from the tilted law rather than from the untilted noise of equation 3.

For a problem with G sampled responses $\{\boldsymbol{y}_g\}$ we use a shaped reward

$$R(\boldsymbol{y}) = w_{\text{s}}\mathbb{1}[\text{solved}] + w_{\text{v}}\mathbb{1}[\text{valid}] + w_{\text{p}}\,\rho(\boldsymbol{y}) + w_{\text{b}}\mathbb{1}[\text{formatted}], \qquad \hat{A}_g = \frac{R_g - \bar{R}}{\text{std}(R)}, \tag{7}$$

where $\rho \in [0,1]$ measures shortest-path progress and the weights satisfy $w_{\text{s}} > w_{\text{v}} + w_{\text{p}} + w_{\text{b}}$, so any solution outranks any failure by a fixed margin. A binary reward leaves 65–71% of groups with $\text{std}(R) = 0$, where $\hat{A}_g \equiv 0$; shaping reduces this to 43%, and the remaining degenerate groups are dropped from the text update rather than rescued by a near-zero denominator. Text gradients have cosine similarity 0.93 within a problem and 0.02 across problems, so we accumulate $M = 12$ problems per step, raising across-problem agreement to 0.871. The text objective is the usual clipped surrogate with a KL term to the supervised reference.

The reward does not reach the latent path. Measuring the cosine similarity of the latent reward gradient across independent rollouts of one problem gives 0.003 to 0.013 at $K \in \{8, 16\}$, so it carries no stable direction, and the tilt is trained by equation 6 alone, on every problem including those with $\text{std}(R) = 0$, since the energy target is independent of the reward. The total Stage-2 update is therefore

$$\max_\theta \; \mathcal{J}_{\text{GRPO}}(\theta) \quad \text{and} \quad \min_\phi \; \textstyle\sum_i \mathcal{L}_{\text{en}}(i), \tag{8}$$

optimized on the same rollouts, which we call *Energy-GRPO*; the arm with ϕ frozen is plain GRPO.

Latent updates are accepted under an analytic KL trust region on the tilted noise law, and energy samples are keyed so that each visit to a problem draws fresh noise; Algorithm 2 and Appendix B give the details.

2.7 INFERENCE

The trained head is used on the decode path: the embedding fed back at each latent step is $\bar{\boldsymbol{z}}_i = \frac{1}{K}\sum_k g_\psi(\boldsymbol{h}_i, \sigma\boldsymbol{\varepsilon}^{(k)})$, the mean of K draws. Setting $\sigma = 0$ gives a deterministic point estimate; with $\sigma > 0$ the learned scale is available across repeated samples, and several reasoning chains may be drawn and aggregated by self-consistency. Because g_ψ is nonlinear, scaling the input noise does not scale the output spread proportionally, so σ must be set explicitly rather than inferred from the training spread. We report $\sigma = 0$ throughout except where self-consistency is stated, which uses the training value $\sigma = 1$.

Dependence on the latent path. Figure 3 compares accuracy with and without the latent span. Removing it costs the strictly proper model 45.7 ± 1.4 points, while the squared-error model is unchanged within run-to-run variation ($+1.2 \pm 3.9$), matching reports that such latents can be ablated or replaced with little effect (Viveiros et al., 2026; Guo et al., 2026). The energy-trained model instead carries a large part of its competence in the latent channel.

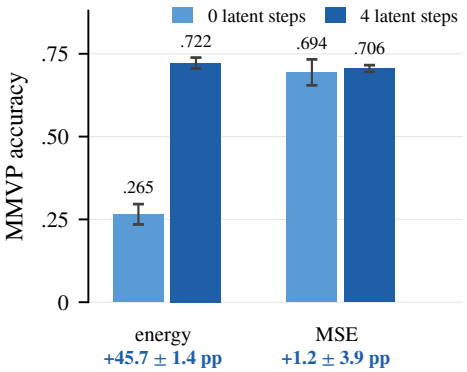


Figure 3: Dependence on the latent path at decode time. Both arms use the identical stochastic head and differ only in the objective. MMVP accuracy with and without the latent span, mean ± s.d. over three runs. Removing the span costs the strictly proper model 45.7 ± 1.4 points, while the squared-error model is unchanged within run-to-run variation ($+1.2 \pm 3.9$), so the energy-trained model has relocated a large part of its competence into the latent channel.

3 Experiments

3.1 Setup

We build on Qwen2.5-VL (Bai et al., 2025) with the vision tower and multimodal projector frozen, so the target space of §2.1 is fixed. Three training lineages are used, each named in the caption of the table it supports. The *VSP lineage* trains on the VSP maze data of Yang et al. (2026) and supports planning (Table 1), geometry (Table 4) and policy optimization (Table 5). The *Zebra-CoT Jigsaw lineage* trains on the Jigsaw split of Zebra-CoT following Ding et al. (2026) and supports the out-of-domain and semantic-consistency results (Tables 2 and 3), so every benchmark there is off the training task. The *Visual-CoT lineage* trains on Visual-CoT, whose ROI boxes supply the alignment target, and supports the objective ablation (Table 6). Unless a caption states otherwise: $\alpha = 1$, $K = 8$, a single pooled target shared across the latent span, per-dimension normalized distances equation 6, decoding with $\sigma = 0$, and the mean of the K draws fed back at each latent step (§2.7; Stage 2 feeds a single draw). Latent budgets and seeds are stated per caption: the main planning rows use one seed and 8 latent steps, the geometry and policy-optimization tables three seeds. All evaluations run through one harness with identical prompts. Hyperparameters are in Appendix C.

Supervised and reinforced arms. Rows are supervised unless labelled Energy-GRPO, which isolates the alignment objective from reward-driven confounds. Where a reinforced arm is reported we state which parameters the reward reaches (§2.6).

3.2 Benchmarks

Group 1: spatial planning (Table 1). VSP (Yang et al., 2026) on seen levels 3–6 and unseen levels 7–8. This is the primary group: the tasks require committing to one of several admissible continuations from a single visual state, and it is the suite on which Mirage and CoLVR report their largest gains, so the comparison is made on their own ground.

Group 2: out-of-domain visual understanding (Table 2). The five benchmarks reported by Ding et al. (2026): V* (Wu & Xie, 2024), MMVP (Tong et al., 2024), MMStar, VisPuzzle and CV-Bench. V* and MMVP are the fine-grained perception benchmarks used by LVR, so this group also carries the perception comparison; the same-backbone, same-data version of that comparison is the objective ablation of §4.

Group 3: semantic consistency (Table 3). SV-QA (Xu et al., 2026a) queries one region from two semantic aspects and scores them jointly, so it penalises a latent that encodes a region under one aspect only. The split released with that benchmark is not public, so we evaluate on our own construction at a larger sample size and report only same-protocol rows, LVR and ER-LVR run through an identical pipeline with a single extraction rule. Published numbers are measured on different items and are not reproduced alongside.

Table 1: **Main results on VSP**, comparing ER-LVR with other VLMs and latent reasoning baselines. Results are reported in accuracy (%), and best results are highlighted in **bold**. Rows marked † are reported by Ding et al. (2026), Tab. 1, as printed, including their reported averages; we did not recompute them. Our rows use the same VSP lineage, 8 latent steps, greedy decoding with σ=0 and the mean of the K draws fed back (§2.7).

Model	Seen					Unseen			Total
	Level 3	Level 4	Level 5	Level 6	Avg.	Level 7	Level 8	Avg.	
Proprietary Model									
GPT-4o† (OpenAI, 2024)	67.00	46.00	24.00	17.00	38.50	11.00	6.00	8.50	28.50
Gemini-2.5-flash† (Gemini Team, 2025)	95.00	86.00	62.00	30.00	62.00	32.00	18.00	25.00	53.83
Open source model									
Qwen2.5-VL-7B† (Bai et al., 2025)	11.00	7.00	7.00	1.00	6.50	1.00	0.00	0.50	4.50
w/ vanilla SFT†	91.00	85.00	71.00	50.00	74.30	36.00	29.00	29.00	59.17
w/ vanilla SFT + GRPO†	89.00	80.00	66.00	49.00	71.00	34.00	14.00	24.00	55.33
Latent model									
Mirage† (Yang et al., 2026)	90.00	84.00	67.00	55.00	74.00	39.00	25.00	32.00	60.00
CoLVR† (Ding et al., 2026)	94.00	**92.00**	80.00	57.00	80.70	43.00	**29.00**	36.00	65.83
ER-LVR (SFT)	**97.00**	**92.00**	82.00	64.00	83.75	**48.00**	26.00	37.00	68.17
ER-LVR + Energy-GRPO	**97.00**	**92.00**	**85.00**	**65.00**	**84.75**	47.00	28.00	**37.50**	**69.00**
Δ vs CoLVR	+3.00	0.00	+5.00	+8.00	+4.00	+4.00	−1.00	+1.50	+3.17

Table 2: **Out-of-domain visual understanding** (accuracy %) against latent-reasoning baselines on the benchmarks of Ding et al. (2026), Tab. 2. Rows marked † are reported from that table; Δ is ER-LVR minus that baseline. Zebra-CoT Jigsaw lineage, so every benchmark here is out-of-domain.

Model	V*	MMVP	MMStar	VisPuzzle	CV-Bench
ER-LVR (σ=0)	**80.63**	**72.67**	**62.60**	**38.50**	**79.04**
CoLVR† (Ding et al., 2026)	80.63	72.00	61.47	37.00	76.95
Δ	0.00	+0.67	+1.13	+1.50	+2.09
LVR † (Li et al., 2026a)	80.10	69.33	58.07	37.00	75.51
Δ	+0.53	+3.34	+4.53	+1.50	+3.53
Mirage† (Yang et al., 2026)	75.92	68.33	55.27	36.75	72.40
Δ	+4.71	+4.34	+7.33	+1.75	+6.64

Table 3: **Semantic consistency on the SV-QA-style paired protocol** (accuracy %). Q_1/Q_2 are the two semantic queries on the same region; *Both* is the fraction of pairs where both are correct. LVR and ER-LVR are both trained by us on the Zebra-CoT Jigsaw lineage, so the comparison is same-backbone, same-target and same-data. Qwen2.5-VL is the base model evaluated on the same items.

Method	V* (n=191)			HR-Bench-4K (n=800)			HR-Bench-8K (n=800)		
	Q_1	Q_2	Both	Q_1	Q_2	Both	Q_1	Q_2	Both
Qwen2.5-VL (Bai et al., 2025)	79.1	79.1	64.9	**70.6**	**83.3**	**59.6**	65.5	**76.3**	51.1
LVR (Li et al., 2026a)	77.0	79.6	65.4	67.4	79.0	55.4	62.0	74.5	48.8
ER-LVR	**80.1**	**82.2**	**69.1**	70.3	79.5	58.0	**67.0**	75.4	**51.3**
gain over LVR	+3.1	+2.6	+3.7	+2.9	+0.5	+2.6	+5.0	+0.9	+2.5

3.3 Baselines and what counts as the primary comparison

The latent baselines for VSP and out-of-domain understanding use the same training-task lineages and test sets as ER-LVR: VSP for the former (the public 600-item test set) and Zebra-CoT Jigsaw for the latter. We use the Mirage (Yang et al., 2026), CoLVR, and other baseline scores reported by Ding et al. (2026). For semantic consistency, we retrain LVR (Li et al., 2026a) and ER-LVR with the same backbone, target, and training data, and evaluate both and base Qwen on identical items ($n = 800$ per HR-Bench setting; Table 3). The objective-only comparison is among the stochastic-head arms in Table 6.

3.4 RESULTS

Planning. On VSP, ER-LVR reaches 68.17 total with supervised training and 69.00 with Energy-GRPO, against 65.83 for CoLVR and 60.00 for Mirage: gains of $+2.34$ and $+3.17$ over the strongest prior latent method (Table 1). Both arms also exceed a proprietary reasoning model on the same protocol. For the reinforced arm the margin over CoLVR is larger on the seen levels ($+4.00$) than on the unseen ones ($+1.50$), and the unseen-level margin is within seed variation at this sample size.

Out-of-domain visual understanding. ER-LVR matches or exceeds every latent baseline on every benchmark in this group (Table 2). Against CoLVR it ties on V^* and leads by $+0.67$ to $+2.09$ on the other four; against LVR the margins are $+0.53$ to $+4.53$, and against Mirage $+1.75$ to $+7.33$. The largest margins over LVR fall on MMStar ($+4.53$) and CV-Bench ($+3.53$), the two benchmarks furthest from the Jigsaw training task.

Semantic consistency. On the joint SV-QA score ER-LVR improves over LVR under an identical protocol by $+3.7$ on V^*, $+2.6$ on HR-Bench-4K and $+2.5$ on HR-Bench-8K (Table 3), with the marginals reported alongside so that a joint-score change can be separated from a change in either query.

4 ABLATIONS

The ablations separate the objective from the architecture and test the mechanism of Proposition 2 independently of accuracy; all arms use the Qwen2.5-VL 7B backbone, matching the baselines reproduced from Ding et al. (2026).

Objective, holding the head fixed. The energy arm introduces both a new loss and a new head, so the head must be controlled for. We train three arms that share the *identical* stochastic head and target and differ only in the objective: per-sample MSE, $\alpha = 2$ (equivalently mean-matching MSE, Prop. 1), and $\alpha = 1$. Table 6 gives the decomposition. Replacing the deterministic head with the stochastic one, while keeping per-sample squared error, recovers 2.09 of the 3.14-point V^* improvement over LVR and 2.00 of the 3.33 points on MMVP; the objective contributes the remaining 1.05 and 1.33. Most of the benchmark gain over LVR is therefore attributable to the head. What the objective adds is consistent across both benchmarks and is accompanied by the geometric signature below, which the per-sample arm does not produce. Self-consistency decoding does not help, lowering V^* from 82.20 to 81.68 and raising MMVP only from 73.00 to 73.33.

Latent geometry. We measure the geometry of the learned latent distribution directly, which tests the mechanism independently of benchmark accuracy. Let

$$\mathrm{fid} = \mathbb{E}\,\|\boldsymbol{z} - \boldsymbol{t}\|\,, \qquad \mathrm{rep} = \mathbb{E}\,\|\boldsymbol{z} - \boldsymbol{z}'\|\,, \qquad \frac{\mathrm{rep}}{\mathrm{fid}} = \frac{\sqrt{2}\,r}{\sqrt{b^2 + r^2}}, \tag{9}$$

per-dimension normalized, the last equality holding under the isotropic concentration model of Proposition 2. These are the bare expectations; equation 6 weights the first by two, and folding that factor in would move the landmark below from 1 to $1/2$. By equation 9 the ratio is increasing in r/b, equals 1 at the fixed-centre optimum $r = b$, and is bounded above by $\sqrt{2}$, attained as $b \to 0$, so a reading above $\sqrt{2}$ indicates a normalization fault rather than a property of the model. The effective rank, the exponential of the eigen-entropy of the $K \times K$ Gram matrix of centred samples, has ceiling $K - 1$ and is invariant to uniform scaling, so it measures the shape of the sample cloud and not its size.

The three arms of the objective ablation share a head, a target and data, so their geometry isolates the objective. Per-sample squared error drives the spread out, so $\mathrm{rep}/\mathrm{fid} \to 0$; the mean-matching boundary constrains only the mean and has no attractor; the strictly proper interior converges to 1. Table 4 confirms all three, with the energy arm at 1.009 ± 0.003. The effective rank is near its ceiling and nearly identical across arms, so the arms differ in the size of the sample cloud rather than its shape, consistent with the isotropy Proposition 2 assumes, though not establishing it, since a cloud with unequal axis lengths can still read near the ceiling.

Table 4: **Latent geometry.** Predicted vs. observed $\mathrm{rep}/\mathrm{fid}$ (Propositions 1 and 2), with the two distance terms and the effective rank. VSP lineage, 8 latent steps; mean $\pm$ half-range over seeds 42–44. The same ordering holds on Visual-CoT (4 steps): 0.074 / 0.177 / 0.946.

Arm	predicted $\mathrm{rep}/\mathrm{fid}$	fid	rep	observed $\mathrm{rep}/\mathrm{fid}$	eff. rank
MSE per-sample	$\to 0$	0.0565	0.0079	0.140 ± 0.007	6.67
α=2 $\equiv$ mean-matching	no attractor	0.0538	0.0145	0.270 ± 0.057	6.64
α=1 (ours)	$\to 1.00$	0.0594	0.0599	1.009 ± 0.003	6.62

Table 5: **Energy-GRPO on VSP.** Official 600-item test set. Arms share a supervised start per seed and use the same shaped text reward (M=12); only Energy-GRPO trains the noise tilt. Δ is the gain over the supervised start in percentage points.

Arm	tilt	seed 42	seed 43	seed 44	mean	Δ
supervised start	—	0.6650	0.6783	0.6700	0.6711	—
GRPO	frozen	0.6833	0.6983	0.6883	0.6900	+1.89
Energy-GRPO	energy-trained	0.6850	0.6950	0.6950	0.6917	+2.06

Table 6: **Objective ablation** (Visual-CoT lineage, 4 latent steps; σ=0 except the last row). All arms share the same target and data and differ only in the objective: the first row is our re-trained LVR with its deterministic head, the rest an identical stochastic head. rep/fid is measured on the same checkpoint.

Arm	V^*	MMVP	rep/fid
deterministic head + MSE (LVR)	79.06	69.67	n/a
stochastic head + MSE per-sample	81.15	71.67	0.074
stochastic head + α=2	80.63	70.67	0.177
stochastic head + α=1 (ours)	**82.20**	73.00	0.946
+ self-consistency decoding	81.68	**73.33**	0.946

Is it the objective or the stochasticity? The per-sample MSE arm shares the head, the noise input and the sample count with α=1 and differs only in the objective, yet it drives the spread out ($\mathrm{rep}/\mathrm{fid} = 0.074$ against 0.946) and sits below α=1 on both benchmarks. Stochastic capacity alone is therefore not sufficient; what separates the arms is the spread the objective selects.

Gradient matching between arms. Equal β does not equalise the auxiliary signal. At identical β the ratio of auxiliary to next-token gradient norm is 2.72 for the energy arm and 1.05 for the squared-error arm, so the arms in Table 6 are matched on β but not on effective auxiliary weight.

Energy-GRPO. Starting from the supervised checkpoint, two Stage-2 arms take the same text updates under the shaped reward and differ only in whether the latent tilt continues to be trained under equation 6. Both improve the supervised model, by 1.89 and 2.06 points, and both are positive in all three seeds (Table 5); the additional 0.17 changes sign across seeds and is within seed variation. The reward reaches only the text policy in both arms: its gradient into the head has cosine similarity 0.003 to 0.013 across independent rollouts of one problem, so it carries no stable direction to accumulate and we do not apply it there.

5 Conclusion

Latent visual reasoning trains its visual thoughts by regressing each latent onto a fixed target embedding. We showed that this is the $\alpha = 2$ boundary of the energy score, where the score depends on the predictive distribution only through its mean. In the strictly proper interior the scale becomes part of what is optimized: holding the centre fixed, the selected scale equals the distance from the centre to the target, for every exponent in $(0, 2)$. We instantiated this as a replacement for the alignment loss, drew each visual thought from a reparameterized stochastic head used on the decode path, and measured the resulting latent geometry against the landmarks the analysis predicts.

AI USAGE CLARIFICATION

We used large language models (LLMs) as a general-purpose writing and coding assistant. Specifically, LLMs were used to help organise the sectioning and structure of the manuscript, to improve clarity and readability, and to assist with routine software engineering tasks such as writing boilerplate code, refactoring, and debugging. All research contributions reported in this paper, namely the analysis of squared-error alignment as the boundary of the energy-score family, the fixed-centre scale result, the stochastic latent head and energy objective, the Energy-GRPO procedure, and the experimental design, originate from the authors. LLMs were not used to generate research ideas, to design the method, or to interpret results. Every LLM-assisted component was reviewed, verified, and edited by the authors: generated text was checked against our own claims and results, and generated code was read, tested, and validated against the numbers reported in the paper. The authors take full responsibility for all content of this submission.

ETHICS STATEMENT

All experiments use publicly released research data: the Visual-CoT, VSP and Zebra-CoT training corpora and the V*, MMVP, MMStar, VisPuzzle, CV-Bench and HR-Bench evaluation benchmarks. Our semantic-consistency evaluation is constructed from images already contained in V* and HR-Bench. We collect no new data, and no human subjects were involved. Our models are fine-tuned from a publicly available vision–language model and inherit its capabilities and biases; the method changes how visual evidence is represented during reasoning rather than extending what the model can perceive or generate, so we regard its additional misuse potential as low relative to the base model.

REPRODUCIBILITY STATEMENT

We provide our implementation, including training and evaluation scripts for the supervised and Energy-GRPO stages and the scripts that reproduce the tables and figures in this paper, as supplementary material. Proofs of all propositions are given in Appendix A, and pseudocode for both training stages in Appendix B. The architecture, training and decoding setup are described in §3 and Appendix C, and each table caption states the training lineage, latent budget and decoding setting of its rows. Rows reproduced from prior work are marked † and taken verbatim from the cited source; all other rows are produced by us under a single evaluation harness.

REFERENCES

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*, 2025. URL https://arxiv.org/abs/2502.13923.

Jingcheng Deng, Zihao Wei, Liang Pang, Junhong Wu, Shicheng Xu, Zenghao Duan, and Huawei Shen. Latent-GRPO: Group relative policy optimization for latent reasoning. *arXiv preprint arXiv:2604.27998*, 2026. URL https://arxiv.org/abs/2604.27998.

Ziyang Ding, Linjian Meng, Yiming Wu, Yuhan Li, Yuhao Liu, and Zhen Zhao. CoLVR: Enhancing exploratory latent visual reasoning via contrastive optimization. *arXiv preprint arXiv:2605.08802*, 2026. URL https://arxiv.org/abs/2605.08802.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Jiawei Guo, Yu Chen, Xiang Wang, Shuai Li, Xinpei Zhao, Huaxing Liu, Shuai Dong, Feifei Zhai1, and Yu Zhou. Beyond visual memory: Mechanistic diagnostics of latent visual reasoning. *arXiv preprint arXiv:2606.01287*, 2026. URL https://arxiv.org/abs/2606.01287.

Chaoya Jiang, Yongrui Heng, Wei Ye, Han Yang, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. VLM-R^3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. In *NeurIPS*, 2025. URL https://openreview.net/forum?id=ERGIrG9GBR.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *ICLR*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/efa0939caa8a906fd0a61a9c60e46e8b-Abstract-Conference.html.

Kelvin Li, Chuyi Shang, Leonid Karlinsky, Rogerio Feris, Trevor Darrell, and Roei Herzig. Latent implicit visual reasoning. *arXiv preprint arXiv:2512.21218*, 2025. URL https://arxiv.org/abs/2512.21218.

You Li, Chi Chen, Yanghao Li, Fanhu Zeng, Kaiyu Huang, Jinan Xu, and Maosong Sun. Imagination helps visual reasoning, but not yet in latent space. In *ICML*, 2026b. URL https://arxiv.org/abs/2602.22766.

Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyue Wei, Sheng Liu, and Xin Eric Wang. Reasoning within the mind: Dynamic multimodal interleaving in latent space. *arXiv preprint arXiv:2512.12623*, 2025. URL https://arxiv.org/abs/2512.12623.

Dongchen Lu, Zhimo Li, Mao Shu, and Huo Cao. DeepLatent: Think with images via parallel latent visual reasoning. *arXiv preprint arXiv:2606.00562*, 2026. URL https://arxiv.org/abs/2606.00562.

Jizheng Ma, Xiaofei Zhou, Geyuan Zhang, Yanlong Song, and Han Yan. LaRe: Latent refocusing for multimodal reasoning. *arXiv preprint arXiv:2511.02360*, 2025. URL https://arxiv.org/abs/2511.02360.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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

Chenze Shao, Darren Li, Fandong Meng, and Jie Zhou. Continuous autoregressive language models. *arXiv preprint arXiv:2510.27688*, 2025. URL https://arxiv.org/abs/2510.27688.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, and Yu Cheng. OpenThinkIMG: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025. URL https://arxiv.org/abs/2505.08617.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *CVPR*, pp. 9568–9578, 2024. doi: 10.1109/CVPR52733.2024.00914. URL https://openaccess.thecvf.com/content/CVPR2024/html/Tong_Eyes_Wide_Shut_Exploring_the_Visual_Shortcomings_of_Multimodal_LLMs_CVPR_2024_paper.html.

André G. Viveiros, Nuno Gonçalves, André F. T. Martins, and Matthias Lindemann. What's holding back latent visual reasoning? *arXiv preprint arXiv:2605.18445*, 2026. URL https://arxiv.org/abs/2605.18445.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. In *NeurIPS*, 2025. URL https://arxiv.org/abs/2505.15966.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond images and language. In *CVPR*, 2026a. URL https://arxiv.org/abs/2511.21395.

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*, 2026b. URL https://arxiv.org/abs/2601.06803.

Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, ChengXiang Zhai, and Klara Nahrstedt. VTool-R1: VLMs learn to think with images via reinforcement learning on multimodal tool use. In *ICLR*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7e6f445a74cdb71931aac64f1e3f49c9-Abstract-Conference.html.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *CVPR*, pp. 13084–13094, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.html.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-R1: Advancing multimodal reasoning capabilities of MLLMs via visual perception reward. In *ICLR*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2cf153951b5e9b39564fc4a0ef6adc1a-Abstract-Conference.html.

Tianrun Xu, Yue Sun, Qixun Wang, Jingyi Lu, Yuan Wang, Tianren Zhang, Longteng Guo, Fengyun Rao, Jing Lyu, Feng Chen, and Jing Liu. Semantic-enriched latent visual reasoning. *arXiv preprint arXiv:2605.19342*, 2026a. URL https://arxiv.org/abs/2605.19342.

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. In *ICLR*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3de60dc2ff7ac8866e805bbf457e2291-Abstract-Conference.html. Oral presentation.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. In *CVPR*, 2026. URL https://arxiv.org/abs/2506.17218.

Chi Zhang, Haibo Qiu, Qiming Zhang, Zhixiong Zeng, Lin Ma, and Jing Zhang. DeepSketcher: Internalizing visual manipulation for multimodal reasoning. *arXiv preprint arXiv:2509.25866*, 2025a. URL https://arxiv.org/abs/2509.25866.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. Adaptive chain-of-focus reasoning via dynamic visual search and zooming for efficient VLMs. *arXiv preprint arXiv:2505.15436*, 2025b. URL https://arxiv.org/abs/2505.15436.

Shitian Zhao, Haoquan Zhang, Shaoheng Lin, Ming Li, Qilong Wu, Kaipeng Zhang, and Chen Wei. PyVision: Agentic vision with dynamic tooling. *arXiv preprint arXiv:2507.07998*, 2025. URL https://arxiv.org/abs/2507.07998.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. In *ICLR*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/cdb347d7516d52a9280ea9d6708911f6-Abstract-Conference.html.

Dongyao Zhu, Zhen Wang, Xi Xiao, Han Jiang, Saeed Vahidian, Wei-Lun Chao, Tanya Berger-Wolf, Yu Su, Raju Vatsavai, and Jianyang Gu. Leveraging latent visual reasoning in silence. *arXiv preprint arXiv:2605.18641*, 2026. URL https://arxiv.org/abs/2605.18641.

A PROOFS

A.1 PROPOSITION 1: THE $\alpha = 2$ BOUNDARY IS MEAN-MATCHING

Let $\boldsymbol{z}, \boldsymbol{z}' \overset{iid}{\sim} P$ with mean $\boldsymbol{\mu}$ and let $\boldsymbol{t}$ be the target. Using $\mathbb{E}\,\|\boldsymbol{z} - \boldsymbol{z}'\|^2 = 2\,\mathbb{E}\,\|\boldsymbol{z} - \boldsymbol{\mu}\|^2$ and the bias–variance split $\mathbb{E}\,\|\boldsymbol{z} - \boldsymbol{t}\|^2 = \mathbb{E}\,\|\boldsymbol{z} - \boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu} - \boldsymbol{t}\|^2$,

$$-S_2(P, \boldsymbol{t}) = 2\,\mathbb{E}\,\|\boldsymbol{z} - \boldsymbol{t}\|^2 - \mathbb{E}\,\|\boldsymbol{z} - \boldsymbol{z}'\|^2 \tag{10}$$

$$= 2\big(\mathbb{E}\,\|\boldsymbol{z} - \boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu} - \boldsymbol{t}\|^2\big) - 2\,\mathbb{E}\,\|\boldsymbol{z} - \boldsymbol{\mu}\|^2 \;=\; 2\,\|\boldsymbol{\mu} - \boldsymbol{t}\|^2 . \tag{11}$$

The variance terms cancel identically, so the objective depends on P only through $\boldsymbol{\mu}$ and is flat in the spread. Consequently S_2 is proper but not *strictly* proper: every distribution with the correct mean attains the optimum. □

A.2 PROPOSITION 2: $r^\star = b$ FOR EVERY $\alpha \in (0, 2)$

Let P be isotropic about $\boldsymbol{\mu}$ with radius r and let $b = \|\boldsymbol{\mu} - \boldsymbol{t}\|$. Under high-dimensional concentration, $\mathbb{E}\,\|\boldsymbol{z} - \boldsymbol{t}\| \to \sqrt{b^2 + r^2}$ and $\mathbb{E}\,\|\boldsymbol{z} - \boldsymbol{z}'\| \to \sqrt{2}\,r$, giving equation 5. Differentiating,

$$\mathcal{E}'_\alpha(r) = 2\alpha r\big(b^2 + r^2\big)^{\alpha/2-1} - \alpha 2^{\alpha/2} r^{\alpha-1}. \tag{12}$$

Setting this to zero, dividing by $\alpha r^{\alpha-1}$ and writing $u = r/b$ gives the scale-free condition

$$f(u) \;\equiv\; u^{2-\alpha}\big(1 + u^2\big)^{\alpha/2-1} \;=\; 2^{\alpha/2-1}. \tag{13}$$

Taking logarithms, $\log f(u) = (2 - \alpha)\log u + (\frac{\alpha}{2} - 1)\log(1 + u^2)$, so

$$\frac{\mathrm{d}\log f}{\mathrm{d}u} = \frac{2-\alpha}{u} + \frac{(\alpha-2)u}{1+u^2} = \frac{2-\alpha}{u\,(1+u^2)} \;>\; 0 \qquad \text{for } \alpha < 2. \tag{14}$$

Hence f is strictly increasing on $(0, \infty)$ and equation 13 has at most one root. Direct substitution shows $u = 1$ is a root for every α, since $f(1) = 1^{2-\alpha} \cdot 2^{\alpha/2-1} = 2^{\alpha/2-1}$. With $\mathcal{E}'_\alpha(0^+) < 0$ and $\mathcal{E}_\alpha(r) \to \infty$ as $r \to \infty$, the root is the unique interior minimum, so $r^\star = b$ independently of α. At $\alpha = 2$ both exponents in equation 13 vanish, the condition degenerates to $1 = 1$, and every scale is stationary, recovering Proposition 1. Substituting $r^\star = b$ gives $\mathcal{E}_\alpha(b) = (2 - 2^{\alpha/2})\,(2b^2)^{\alpha/2}/2^{\alpha/2} \cdot 2^{\alpha/2}$, and for $\alpha = 1$ in particular $\mathcal{E}_1(b) = \sqrt{2}\,b$ against $\mathcal{E}_1(0) = 2b$, so collapse is worse by a factor $\sqrt{2}$. At the optimum $\mathrm{rep}/\mathrm{fid} = \sqrt{2}r^\star/\sqrt{2b^2} = 1$, the landmark used in §4. □

A.3 PROPOSITION 3: GLOBAL OPTIMUM FOR A DETERMINISTIC TARGET

This does not require the isotropic model of equation 5; it follows directly from strict propriety. A deterministic target is the degenerate observation distribution $Q = \delta_{\boldsymbol{t}}$, for which the expected score reduces to a single evaluation, $\mathbb{E}_{Y\sim Q}[S_\alpha(P, Y)] = S_\alpha(P, \boldsymbol{t})$. Strict propriety of S_α for $\alpha \in (0, 2)$ states that this expectation is uniquely maximized at $P = Q$. Evaluating the maximum, $S_\alpha(\delta_{\boldsymbol{t}}, \boldsymbol{t}) = \mathbb{E}\,\|\boldsymbol{u} - \boldsymbol{u}'\|^\alpha - 2\mathbb{E}\,\|\boldsymbol{u} - \boldsymbol{t}\|^\alpha = 0 - 0 = 0$. Hence $S_\alpha(P, \boldsymbol{t}) \le 0$ for every P, with equality if and only if $P = \delta_{\boldsymbol{t}}$; equivalently, in the minimization form used in equation 6, $-S_\alpha(P, \boldsymbol{t}) \ge 0$ with equality only at the point mass. The statement therefore holds over all predictive distributions, isotropic or not. Combining with Proposition 2, the fixed-centre optimum $r^\star = b$ is a constrained solution whose value $\sqrt{2}\,b$ decreases with b and vanishes only in the unconstrained optimum. □

A.4 NON-NEGATIVITY OF THE OBJECTIVE, AND OF ITS ESTIMATOR

For a deterministic target the objective coincides with the *energy distance* between the predictive distribution and a point mass, $-S_\alpha(P, \boldsymbol{t}) = D^2_\alpha(P, \delta_{\boldsymbol{t}})$, which is positive definite for $\alpha \in (0, 2)$; the population objective is therefore non-negative on the whole interval, vanishing only at $P = \delta_{\boldsymbol{t}}$. We note that the elementary argument via the triangle inequality, $\big\|\boldsymbol{z}^{(k)} - \boldsymbol{z}^{(j)}\big\|^\alpha \le \big\|\boldsymbol{z}^{(k)} - \boldsymbol{t}\big\|^\alpha + \big\|\boldsymbol{z}^{(j)} - \boldsymbol{t}\big\|^\alpha$, establishes this only for $\alpha \le 1$, since $(a + b)^\alpha \le a^\alpha + b^\alpha$ requires subadditivity; the positive-definiteness argument is the one that covers the full interval.

The finite-sample estimator in equation 6 is a different matter. It is unbiased but not non-negative, and its variance grows as $\alpha \to 2$, where the population value itself degenerates to $2\,\|\mathbb{E}[\boldsymbol{z}] - \boldsymbol{t}\|^2$ and vanishes for a centred cloud. Negative logged values near the boundary are therefore an estimator artifact rather than evidence of a divergent objective, and should be reported as such. Repulsion weights greater than one remove the guarantee entirely and are not used.

B ALGORITHMS

Algorithm 1 Stage 1: supervised step

Require: batch $\{(\boldsymbol{x}, q, \text{ans}, \text{ROI})\}$; budget B; samples K; exponent α; weight β

$\boldsymbol{t} \leftarrow \text{pool}\big(\Pi(V(\boldsymbol{x}))[\text{ROI}]\big)$ ▷ *frozen encoders; shared across the B steps, equation 2*

for $i = 1$ to B **do**

  obtain $\boldsymbol{h}_i$; draw $\{\boldsymbol{z}_i^{(k)} = g_\psi(\boldsymbol{h}_i, \sigma\boldsymbol{\varepsilon}^{(k)})\}_{k=1}^K$, $\boldsymbol{\varepsilon}^{(k)} \sim \mathcal{U}[-\frac{1}{2}, \frac{1}{2}]^{d_{\boldsymbol{\varepsilon}}}$

  $\mathcal{L}_{\text{en}}(i) \leftarrow$ equation 6

  feed back $\bar{\boldsymbol{z}}_i = \frac{1}{K}\sum_k \boldsymbol{z}_i^{(k)}$ ▷ *sample mean, as in §2.7*

$\mathcal{L} \leftarrow \mathcal{L}_{\text{NTP}} + \beta \sum_i \mathcal{L}_{\text{en}}(i)$; update $\{\theta, \psi\}$ pathwise

Algorithm 2 Stage 2: GRPO on the text policy, with the noise tilt frozen or energy-trained

Require: supervised checkpoint; group size G; accumulation M; clip ϵ; KL threshold κ; arm $\in \{\text{FROZEN}, \text{ENERGY}\}$

for each accumulation window of M problems **do**

  for each problem **do**

    roll out G responses, feeding back a single draw $\boldsymbol{z}_i$ at each latent step; score each with R_g of equation 7

    if $\text{std}(\{R_g\}) = 0$ **then**

      skip the text update for this problem ▷ $\hat{A}_g \equiv 0$

    else

      $\hat{A}_g \leftarrow (R_g - \bar{R})/\text{std}(R)$; accumulate the clipped token surrogate + KL to the reference

    if arm $=$ ENERGY **then**

      accumulate $\sum_i \mathcal{L}_{\text{en}}(i)$ on this problem, regardless of $\text{std}(R)$ ▷ *energy target is reward-independent*

  step the text optimizer ▷ $\max_\theta \mathcal{J}_{\text{GRPO}}$, *equation 8*

  if arm $=$ ENERGY **then**

    step the tilt optimizer to ϕ'; compute $\text{KL}(q_{\phi'}(\cdot \mid \boldsymbol{h}) \,\|\, q_\phi(\cdot \mid \boldsymbol{h}))$ analytically ▷ *noise tilts at fixed $\boldsymbol{h}$*

    if $\text{KL} > \kappa$ **then**

      $\phi \leftarrow (1-\eta)\phi + \eta\phi'$ and roll back optimizer moments

Trust region and sample keying. Stage 2 holds the head g_ψ fixed and trains a tilt ϕ on the noise law: the policy emits $q_\phi(\varepsilon \mid \boldsymbol{h})$, a tilt of the uniform law over ε, and g_ψ maps a draw deterministically to z_i. Write $P_\phi(\cdot \mid \boldsymbol{h})$ for the latent law induced by pushing q_ϕ through g_ψ. The trust region is evaluated on the noise variable, not on the latent: the density of a tilted uniform and its KL against another tilt are closed-form, so Algorithm 2 computes $\text{KL}(q_{\phi'} \,\|\, q_\phi)$ exactly rather than by sampling, and requires no invertibility of g_ψ. Because the same fixed g_ψ transforms both tilts, the data-processing inequality gives

$$\text{KL}(P_{\phi'}(\cdot \mid \boldsymbol{h}) \,\|\, P_\phi(\cdot \mid \boldsymbol{h})) \;\le\; \text{KL}(q_{\phi'}(\cdot \mid \boldsymbol{h}) \,\|\, q_\phi(\cdot \mid \boldsymbol{h}))\,,$$

so constraining the noise-side KL constrains the latent-side divergence. The bound is conditional on a fixed $\boldsymbol{h}$; it does not bound the divergence between whole rollouts, which also change as θ is updated. On rejection we interpolate parameters, $\phi \leftarrow (1-\eta)\phi + \eta\phi'$, rather than rescale the gradient: an adaptive optimizer whose moments derive from that same gradient is near scale-invariant, so a ladder built by scaling could only accept or reject, never shorten. Optimizer moments are rolled back with the parameters. Energy samples are drawn from a key over (seed, problem, visit, latent index), so they are reproducible on resume, shared across group members as common

random numbers, and fresh on each visit to a problem; without the visit index a problem would draw the same K base samples at every visit, letting the tilt reduce equation 6 by rearranging a fixed sample set instead of by widening the distribution.

C Implementation Details

Head. g_ψ projects $\boldsymbol{h}_i$ and the noise vector into a hidden width, refines through residual gated-MLP blocks, applies LayerNorm, and outputs into the target space. The output layer is zero-initialized. We note the exact consequence, since it is easy to misstate: because the head output is fed back directly rather than added residually to $\boldsymbol{h}_i$, zero initialization produces a *zero latent*, not an identity map. The initial centre error is therefore $b_0 = \|\boldsymbol{t}\|$ and the initial spread is zero, so the spread grows into $r^\star$ from below rather than being tuned to it. A residual parameterization $\boldsymbol{z} = \boldsymbol{h}_i + \Delta_\psi(\boldsymbol{h}_i, \boldsymbol{\varepsilon})$ would instead start from the raw hidden state; we report the direct form and note the alternative.

Normalization. All distances are divided by $m^{\alpha/2}$ equation 6, making the loss magnitude invariant to hidden size so that one weight β transfers across backbones and is directly comparable with a mean-reduced MSE baseline.

Inference. The head is engaged on the decode path; $\sigma = 0$ for all reported rows except self-consistency, which uses the training value $\sigma = 1$ and aggregates several chains by majority vote.

Stability. Following continuous-generation practice (Shao et al., 2025) we regularize the latent manifold (variational regularization with KL clipping, latent dropout) so that fed-back latents remain well-behaved when re-entering the reasoning stream.

Optimization. All runs fine-tune Qwen2.5-VL-7B-Instruct in bf16 under DeepSpeed ZeRO-2 on four NVIDIA H200 GPUs, with the vision tower and multimodal projector frozen and the language model fully trainable; no LoRA or other adapter is used. Per-device batch size is 1 throughout, so the effective batch is set by gradient accumulation.

Stage 1, VSP lineage (Tables 1 and 4): learning rate 5×10^{-5}, linear schedule, 10 warmup steps, weight decay 0.01, 2500 steps, accumulation 1; 8 latent steps, K=8, α=1, σ=1, repulsion weight 1, $\beta = 3.33$.

Stage 1, Visual-CoT lineage (Table 6): learning rate 10^{-5}, cosine schedule, warmup ratio 0.03, weight decay 0.1, 2500 steps, accumulation 16; 4 latent steps, a shared target, K=8, α=1, $\beta = 0.1$; sequence packing at 16,384 tokens.

Stage 1, Zebra-CoT Jigsaw lineage (Tables 2 and 3): 200 steps; both the ER-LVR and LVR arms are trained by us under the same recipe.

Stage 2 (Table 5): G=16 sampled responses per problem, M=12 accumulated problems per optimizer step, 8 latent steps; text learning rate 2.5×10^{-6}, tilt-head learning rate 10^{-4}, KL target 0.01 with cap 0.05. Seeds 42, 43 and 44.

D Per-Example Residual Calibration

Proposition 2 predicts a per-example relation, $r^\star = b$, whereas the diagnostic in §4 is an aggregate: over- and under-dispersion on different examples average to the same ratio as uniform agreement, so a reading near the landmark is consistent with the prediction without establishing it. The per-example test is left to future work.

It would draw $K' > K$ samples for each held-out example i and latent step and compare

$$b_i = \|\bar{\boldsymbol{z}}_i - \boldsymbol{t}_i\|, \qquad r_i = \Big(\tfrac{1}{K'} \textstyle\sum_k \Big\|\boldsymbol{z}_i^{(k)} - \bar{\boldsymbol{z}}_i\Big\|^2\Big)^{1/2}, \tag{15}$$

regressing r_i on b_i and asking whether the spread also predicts answer error.

Two constraints govern where it can run. The target $\boldsymbol{t}_i$ is formed only from a region annotation, and none is available at inference (§2.1), so the measurement is confined to held-out splits of an

annotated corpus. It also requires $\sigma > 0$ at decode: at $\sigma = 0$ the K' draws coincide and r_i is identically zero, whereas the rows we report use $\sigma = 0$ (§2.7).

E LIMITATION

Our analysis is carried out under an isotropic latent distribution and a single frozen target, and our experiments use one backbone family and a fixed latent budget per table. Benchmark results in Tables 1–3 are from a small number of seeds, and the semantic-consistency evaluation uses our own construction of the benchmark rather than a public split. Policy optimization is studied on one planning task.