SOFT SPATIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models (LVLMs) commonly perform spatial reasoning through chain-of-thought (CoT), encoding intermediate reasoning as autoregressive sequences of discrete language tokens. Such *hard thinking* requires committing to a single token at each step, even when the correct spatial interpretation remains uncertain. This early commitment constitutes *premature discretization*: an incorrect token selection can propagate errors through subsequent reasoning. We propose **Soft Spatial Reasoning**, a post-training framework that introduces *soft thinking* for spatial tasks in LVLMs. At each intermediate reasoning step, the LVLM forms a continuous soft state by mixing token embeddings rather than selecting a single token, allowing multiple candidate continuations to influence the next step. The appropriate degree of softness, however, can vary across reasoning steps: retaining multiple candidates may preserve a useful spatial interpretation, but if those candidates imply conflicting spatial relations, mixing them may interfere with subsequent reasoning. At the core of Soft Spatial Reasoning is **AdaptSoft**, a controller that uses the current hidden state and predictive uncertainty to adapt the degree of softness at each reasoning step. To train AdaptSoft, we introduce a gradient-alignment learning objective that provides a step-specific learning signal for softness control without intermediate reasoning supervision. Across diverse spatial benchmarks, Soft Spatial Reasoning outperforms hard and fixed-soft CoT baselines using the same backbone, as well as a range of existing LVLMs. The source code is available in the supplementary material.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) Alayrac et al. (2022); Li et al. (2023); Liu et al. (2023); Dai et al. (2023); Bai et al. (2023) have demonstrated capabilities across a range of tasks involving visual understanding Yin et al. (2024); Li et al. (2024) and spatial reasoning Liu et al. (2026); Stogiannidis et al. (2025). To support complex reasoning, these models commonly use chain-of-thought (CoT) Wei et al. (2022), generating intermediate reasoning steps before producing a final answer. Standard CoT expresses these steps as an autoregressive sequence of discrete language tokens, a process we refer to as *hard thinking*. As an alternative to this discrete formulation, recent work on Large Language Models (LLMs) Hao et al. (2024); Zhang et al. (2026b); Zheng et al. (2025a) has explored *soft thinking*, in which continuous intermediate states carry information from multiple candidate continuations Zhu et al. (2026); Zhang et al. (2026a) while the final answer is expressed in language.

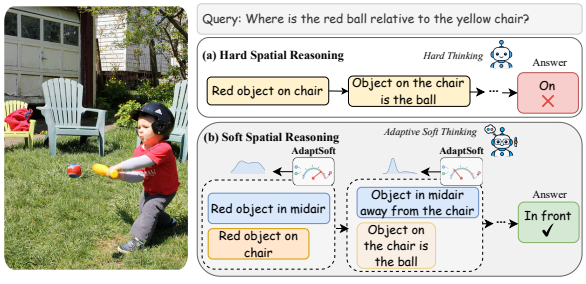


Figure 1: Hard versus adaptive soft thinking for spatial reasoning. (a) Hard thinking commits to one token at each CoT step, allowing early errors to propagate. (b) Our **Soft Spatial Reasoning** forms continuous intermediate states by mixing token embeddings, carrying information from multiple candidate continuations through the CoT. **AdaptSoft** adjusts the mixture's softness at each step based on the current reasoning state and predictive uncertainty.

LVLMs, however, typically use *hard thinking* for spatial reasoning tasks Li et al. (2025); Gholami et al. (2025); Wang & Ling (2025). During the CoT, several spatial interpretations may still appear

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.

plausible to the model Zhang et al. (2025). Consider the question in Figure 1a, which asks where the red ball is relative to the yellow chair. Before resolving which red object is the ball, the model may identify the object on the chair as the ball in its CoT, even though the object in midair remains a plausible candidate. Subsequent reasoning may then treat the object on the chair as the ball, producing the answer `on` rather than `in front`. This illustrates *premature discretization*: committing to a discrete continuation before resolving the spatial interpretation can allow an early error to propagate through the remaining CoT Hao et al. (2024); Hu et al. (2026). Human cognition offers a useful contrast: people can think through a spatial problem before reaching an answer, without putting every thought into words Quiroga et al. (2005); Fedorenko & Varley (2016).

Reasoning without verbalizing every intermediate step is also emerging in LVLMs. Some approaches maintain and refine continuous visual tokens alongside the CoT Yang et al. (2026); Li et al. (2026a;b). In these approaches, softness lies in the visual representations, while the CoT still advances through discrete token selections and may commit prematurely to one interpretation. Other approaches extend soft thinking to the CoT itself Hu et al. (2026); Jeon et al. (2026); Wang et al. (2026b), forming soft states that carry information from multiple candidate continuations and may delay commitment to a single reasoning path. However, the benefit of preserving these alternatives depends on what they represent. In spatial reasoning, alternative interpretations of a scene can imply different spatial relations. During the CoT, a soft state may keep a useful interpretation available at one step; at another, mixing interpretations that imply conflicting relations may interfere with subsequent reasoning. A fixed degree of softness may not suit both situations (Figure 1b).

Research Question

How should LVLMs dynamically control representation softness during spatial reasoning?

To address this question, we propose **Soft Spatial Reasoning** (Figure 1b), a post-training framework that introduces *adaptive soft thinking* for spatial reasoning in LVLMs using soft states at each intermediate CoT step. Central to the framework is **AdaptSoft**, a controller that adjusts the softness of the token-embedding mixtures forming these states. AdaptSoft uses the LVLM's hidden state at each step to account for the ongoing reasoning and predictive uncertainty to gauge the model's confidence in its next continuation. By varying softness, it adjusts the distribution of weights across candidate continuations within each soft state, with the aim of guiding subsequent reasoning toward a correct final answer. We further introduce a **gradient-alignment objective** that measures agreement between each step's contribution to the LVLM policy gradient and a reference gradient. This provides step-level credit alongside task-level rewards for jointly post-training AdaptSoft and the LVLM, without requiring intermediate reasoning annotations.

Our contributions are fourfold:

- We introduce **Soft Spatial Reasoning**, to our knowledge the first framework dedicated to soft thinking for spatial reasoning in LVLMs.
- We design **AdaptSoft** to control softness at each reasoning step using the current reasoning state and predictive uncertainty.
- We develop a **gradient-alignment objective** that provides step-specific credit without external evaluators or intermediate reasoning annotations.
- Extensive experiments covering **19 spatial reasoning categories** demonstrate overall improvements over strong spatial reasoning baselines.

2 Related Works

Hard Spatial Reasoning in LVLMs. Several approaches support spatial reasoning by providing LVLMs with explicit scene geometry and object relationships Cheng et al. (2024); Ma et al. (2025b); Cai et al. (2025b); Liu et al. (2025); Chen et al. (2024b); Hu et al. (2025); Wang et al. (2025c); Cai et al. (2025a); Chen et al.; Sultan et al. (2026); Daxberger et al.; Hong et al. (2023); Wu et al. (2025a); Xu et al. (2026a); Zhao et al. (2025b). Others organize spatial evidence into intermediate representations for grounding and reasoning across perspectives Bigverdi et al. (2025); Wan et al. (2025); Ning et al. (2025); Gholami et al. (2025); Lee et al. (2025); Yang et al. (2025a); Chen

et al. (2026c); Zhou et al. (2026a). Post-training methods optimize grounded reasoning and spatial predictions through supervision or task-specific rewards Kancheti et al. (2026); Xu et al. (2025); Zheng et al. (2025b); Ma et al. (2026); Chen et al. (2025b); Batra et al. (2025); Wang & Ling (2025); Wu et al. (2025b); Sarch et al. (2025); Zhao et al. (2025a); Li et al. (2025); Chen et al. (2026b); Li et al. (2026c), whereas training-free methods guide inference through spatial prompting or interventions on visual processing Liao et al. (2024); Ma et al. (2024); Mitra et al. (2024); Chen et al. (2025a); Yan et al. (2026). These methods strengthen spatial evidence and its use, but premature commitment to a single continuation can remain in discrete CoT.

Continuous and Soft Thinking. Recent work explores continuous intermediate states in place of discrete CoT generation in LLMs Hao et al. (2024); Wei et al. (2026b); Zhou et al. (2026c). Soft thinking retains multiple candidate continuations through mixtures of token embeddings Zhang et al. (2026b); Butt et al. (2026); Zheng et al. (2025a); Wang et al. (2025a). Other work explores discrete CoT paths Dang et al. (2026); Zhou et al. (2026b); Yu et al. (2026); Wei et al. (2026a) or switches between continuous and discrete reasoning Shi et al. (2026); Xu et al. (2026b;c). In LVLMs, continuous representations support multimodal reasoning, including soft thinking within the CoT Hu et al. (2026); Sun et al. (2026); Chen et al. (2026a); Shen et al. (2025); Pham & Ngo (2026); Ma et al. (2025a); Jeon et al. (2026); Wang et al. (2026b); Ray et al. (2026); Huang & Shan (2026), while complementary approaches construct or refine visual representations for subsequent reasoning Yang et al. (2026); Li et al. (2026a;b). These approaches do not learn step-specific softness for token-embedding mixtures. Soft Spatial Reasoning uses the hidden state and predictive uncertainty to adjust softness, with gradient alignment guiding training.

3 METHOD

Soft Spatial Reasoning (Figure 2a) is a post-training framework that enables adaptive soft thinking for spatial tasks in LVLMs through continuous intermediate states and discrete final answers. Its controller, **AdaptSoft** (Figure 2b), determines step-specific softness from the current hidden state and predictive uncertainty. **Gradient-Alignment Learning** (Figure 3) trains this controller by deriving step-specific credit from alignment between each step's policy gradient and a reference gradient.

3.1 PROBLEM SETUP

Given an image I and a spatial reasoning query q, an LVLM policy π_θ generates a rollout $o = (C, a)$ consisting of a CoT C and a discrete final answer a. Hard thinking represents C as a sequence of discrete language tokens, each conditioning subsequent predictions. Soft thinking instead represents C as continuous states (Section 3.2).

The LVLM policy is post-trained using Group Relative Policy Optimization (GRPO) Shao et al. (2024), which uses relative rewards within groups of sampled rollouts without requiring ground-truth CoT supervision. As illustrated in Figure 2a, for each image–query pair (I, q), the rollout policy $\pi_{\theta_{\text{old}}}$ samples a group of G rollouts:

$$o_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid I, q), \qquad i = 1, \ldots, G. \tag{1}$$

Each rollout receives an answer reward $R^{(i)}_{\text{ans}}$ and a format reward $R^{(i)}_{\text{fmt}}$. The answer reward assigns full credit (1) when the final answer is correct and no credit (0) otherwise, while the format reward evaluates compliance with the required reasoning-and-answer structure. Using reward weights w_{ans} and w_{fmt}, and a small constant ϵ for numerical stability, we define the task reward and its group-normalized advantage as

$$r_i = w_{\text{ans}} R^{(i)}_{\text{ans}} + w_{\text{fmt}} R^{(i)}_{\text{fmt}}, \qquad A^{\text{task}}_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^G)}{\text{std}(\{r_j\}_{j=1}^G) + \epsilon}. \tag{2}$$

3.2 SOFT CHAIN-OF-THOUGHT (COT)

Within each rollout o_i, the intermediate CoT is represented by a sequence of T_i soft states $\{s_{i,t}\}_{t=1}^{T_i}$, formed by feeding continuous mixtures of token embeddings back into the language backbone at successive reasoning steps (Figure 2a) Zhang et al. (2026b); Butt et al. (2026). After the reasoning steps, the model switches to discrete token generation for the final answer.

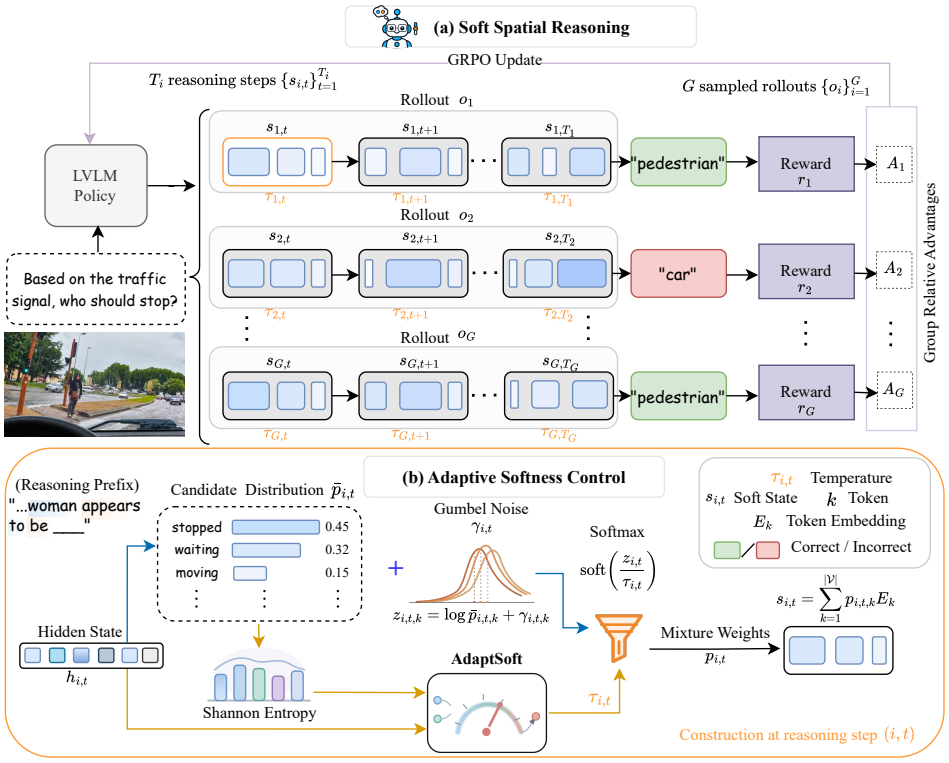


Figure 2: Overview of **Soft Spatial Reasoning**. (a) The LVLM policy is post-trained to carry multiple candidate continuations through soft states during reasoning, then generate the final answer as discrete tokens. Rollout rewards yield group-relative advantages for the GRPO update. (b) At each reasoning step, **AdaptSoft** sets the mixture temperature from the current hidden state and candidate-distribution entropy. The resulting mixture weights combine token embeddings into the next soft state.

Stochastic Soft Rollout. To produce different soft CoTs within the rollout group, independent Gumbel perturbations are applied to the vocabulary log-probabilities at each reasoning step Zheng et al. (2025a); Wu et al. (2026), as illustrated in Figure 2b. Specifically, conditioned on (I, q) and the preceding states $s_{i,<t}$, the rollout policy produces the unperturbed distribution $\bar{p}_{i,t}$. For each token k, an independent Gumbel sample $\gamma_{i,t,k}$ is then added to $\log \bar{p}_{i,t,k}$, yielding the perturbed score $z_{i,t,k}$:

$$\bar{p}_{i,t} = \pi_{\theta_{\text{old}}}(\cdot \mid I, q, s_{i,<t}), \qquad \gamma_{i,t,k} \overset{\text{i.i.d.}}{\sim} \text{Gumbel}(0,1),$$
$$z_{i,t,k} = \log \bar{p}_{i,t,k} + \gamma_{i,t,k}. \tag{3}$$

A temperature-controlled softmax (Figure 2b) converts the perturbed scores into token-mixture weights $p_{i,t}$, whose weighted combination of token embeddings forms the soft state $s_{i,t}$:

$$p_{i,t} = \text{softmax}\left(\frac{z_{i,t}}{\tau_{i,t}}\right), \qquad s_{i,t} = \sum_{k=1}^{|\mathcal{V}|} p_{i,t,k} E_k, \tag{4}$$

where $\mathcal{V}$ is the language vocabulary. At each step, the softmax is applied to its top-$\mathcal{K}$ candidate tokens, with $p_{i,t,k} = 0$ for all other tokens. Here, k indexes vocabulary tokens, and $E_k \in \mathbb{R}^d$ is the embedding of token k. The temperature $\tau_{i,t} > 0$ controls mixture concentration: lower values move $s_{i,t}$ toward the highest-scoring token's embedding, while higher values spread weight across token embeddings.

Gumbel-Reparameterized Likelihood. GRPO requires a likelihood ratio between the current and rollout policies at each reasoning step. Unlike a discrete CoT step, a soft state $s_{i,t}$ combines token embeddings without selecting an individual token, so the standard sampled-token likelihood does not apply. The perturbed score vector $z_{i,t}$ is the sampled variable that, given $\tau_{i,t}$, deterministically specifies $s_{i,t}$ (Equations (3) and (4)). The likelihood ratio is therefore defined over $z_{i,t}$, evaluating the same recorded score vector under both policies.

During the policy update, the sampled scores and corresponding temperatures from the preceding steps reconstruct $s_{i,<t}$, ensuring that both policies are evaluated on the same reasoning prefix. Conditioned on $(I, q, s_{i,<t})$, the current policy produces $\bar{p}^{\theta}_{i,t}$. For each recorded score $z_{i,t,k}$, subtracting the current token log-probability yields the Gumbel noise value implied by the current policy. Evaluating the implied noise values under independent standard Gumbel densities gives the conditional joint log-density of $z_{i,t}$, with conditioning on $(I, q, s_{i,<t})$ omitted for compactness:

$$\begin{aligned} \bar{p}^{\theta}_{i,t} = \pi_\theta(\cdot \mid I, q, s_{i,<t}), \qquad \widetilde{\gamma}^{\theta}_{i,t,k} = z_{i,t,k} - \log \bar{p}^{\theta}_{i,t,k}, \\ \log P_\theta(z_{i,t}) = \sum_{k=1}^{|\mathcal{V}|} \left[-\widetilde{\gamma}^{\theta}_{i,t,k} - \exp\left(-\widetilde{\gamma}^{\theta}_{i,t,k}\right) \right]. \end{aligned} \tag{5}$$

Evaluating the same scores under the rollout policy gives $P_{\theta_{\text{old}}}(z_{i,t})$, yielding the soft-step GRPO ratio:

$$\rho^{\text{soft}}_{i,t}(\theta) = \exp(\log P_\theta(z_{i,t}) - \log P_{\theta_{\text{old}}}(z_{i,t})). \tag{6}$$

As detailed in Section 3.3, the rollout temperatures are reintroduced through AdaptSoft when reconstructing the soft states, allowing gradients from subsequent reasoning steps to propagate to its parameters. A derivation of the density and likelihood ratio is provided in Appendix Section C.1.

3.3 AdaptSoft: Adaptive Softness Control

Step-Specific Softness Controller. At reasoning step t of rollout o_i, let $h_{i,t} \in \mathbb{R}^d$ denote the final-layer hidden state and $\bar{p}_{i,t}$ the unperturbed next-token distribution. The hidden state summarizes the accumulated reasoning context, while the entropy of $\bar{p}_{i,t}$ measures predictive uncertainty among candidate continuations. The entropy is standardized as

$$\widehat{H}_{i,t} = \frac{H(\bar{p}_{i,t}) - \mu_H}{\sigma_H + \epsilon_H}, \tag{7}$$

where $H(\cdot)$ denotes Shannon entropy. The mean μ_H and standard deviation σ_H are estimated once from the next-token entropies of the initial policy and remain fixed throughout post-training. The constant $\epsilon_H > 0$ prevents division by zero.

Before entering the controller, $h_{i,t}$ is layer-normalized and projected to $d_p < d$ dimensions using a fixed random matrix $\mathbf{P} \in \mathbb{R}^{d_p \times d}$. This compact representation reduces the rollout storage required for controller training. The projected state is concatenated with $\widehat{H}_{i,t}$ and passed to an MLP f_ϕ, whose scalar output parameterizes the temperature:

$$u_{i,t} = f_\phi\Big(\Big[\mathbf{P}\ \mathrm{LN}(h_{i,t}) \,\Big\|\, \widehat{H}_{i,t}\Big]\Big), \qquad \tau_{i,t} = \tau_0 + \Delta \tanh(u_{i,t}), \tag{8}$$

where LN denotes layer normalization, $\|$ denotes concatenation, and ϕ contains the learnable controller parameters. Since $\tanh(u_{i,t}) \in (-1, 1)$, τ_0 sets the base temperature and Δ sets its maximum step-specific deviation, yielding $\tau_{i,t} \in (\tau_0 - \Delta, \tau_0 + \Delta)$. We require $0 < \Delta < \tau_0$ to keep all temperatures above zero.

Within this range, AdaptSoft uses the current reasoning context and predictive uncertainty to adjust the softness of each mixture. As described in Section 3.2, the sampled soft states are reconstructed during the policy update from the recorded scores and rollout temperatures. To train AdaptSoft without altering these sampled states, the controller inputs and outputs are retained during rollout. A stop-gradient construction uses the recorded temperature in the forward computation while passing gradients through its recomputed value to ϕ; details are provided in Appendix Section C.2.

3.4 Post-Training Optimization

The LVLM policy is optimized with GRPO, while AdaptSoft is trained through step-specific gradient alignment.

GRPO for the LVLM Policy. For each sampled step, $\rho_{i,t}(\theta)$ compares its likelihood under the current and rollout policies. Soft reasoning steps use the density ratio $\rho^{\text{soft}}_{i,t}(\theta)$ in Equation 6, while

discrete final-answer tokens use the token-probability ratio. Each ratio is weighted by the shared rollout advantage A_i^{task}, encouraging steps from higher-reward rollouts and discouraging those from lower-reward rollouts. The clipped surrogate limits the incentive for large ratio changes, while KL regularization over the LVLM's next-token distributions penalizes deviation from π_{ref}. Averaging over steps and rollouts gives

$$\mathcal{J}(\theta) = \mathbb{E}_{\{o_i\}\sim\pi_{\theta_{\text{old}}}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\Big(\min\big\{\rho_{i,t}(\theta)A_i^{\text{task}}, \\ \text{clip}\left(\rho_{i,t}(\theta), 1-\delta, 1+\delta\right)A_i^{\text{task}}\big\} - \eta D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big)\Big)\Bigg]. \tag{9}$$

Here, $|o_i|$ counts soft reasoning steps and discrete final-answer tokens, δ is the clipping threshold, and η weights the KL penalty against the reference LVLM policy π_{ref}. The KL distributions are conditioned on (I, q) and the preceding steps. Optimization minimizes $\mathcal{L}_{\text{GRPO}}(\theta) = -\mathcal{J}(\theta)$.

Step-Specific Gradient Alignment for AdaptSoft. To train AdaptSoft, we introduce a gradient-alignment objective that provides a step-specific learning signal for each predicted temperature. The shared rollout-level advantage in GRPO does not directly distinguish how softness should vary across reasoning steps. The proposed objective provides this signal by comparing each step's contribution to the LVLM policy gradient with a reference gradient (Figure 3). Each batch of sampled rollouts is divided into disjoint reference and optimization subsets. The reference subset provides a gradient G_{ref}, while each soft reasoning step t of rollout o_i in the optimization subset contributes a gradient $g_{i,t}$. Both gradients are computed with respect to the LVLM output-layer weights W.

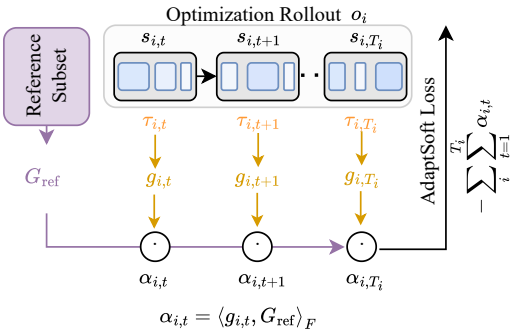


Figure 3: **Gradient-Alignment Learning.** At each soft reasoning step t, the temperature $\tau_{i,t}$ controls the state $s_{i,t}$, whose contribution $g_{i,t}$ to the LVLM output-layer gradient is compared with a reference gradient G_{ref} computed from a disjoint subset of rollouts. The resulting alignment scores $\alpha_{i,t}$ are aggregated to form the AdaptSoft loss.

The alignment score $\alpha_{i,t}$ is defined as the Frobenius inner product of these gradients. The AdaptSoft loss $\mathcal{L}_{\text{AS}}(\phi)$ is the negative sum of the alignment scores across soft reasoning steps in the optimization subset:

$$G_{\text{ref}} = \nabla_W \mathcal{L}_{\text{GRPO}}^{\text{ref}}, \qquad g_{i,t} = \nabla_W \mathcal{L}_{\text{GRPO}}^{(i,t)}, \\ \alpha_{i,t} = \langle g_{i,t}, G_{\text{ref}}\rangle_F, \qquad \mathcal{L}_{\text{AS}}(\phi) = -\sum_i \sum_{t=1}^{T_i} \alpha_{i,t}, \tag{10}$$

where $\mathcal{L}_{\text{GRPO}}^{\text{ref}}$ is the GRPO loss on the reference subset, $\mathcal{L}_{\text{GRPO}}^{(i,t)}$ is the contribution of step t in rollout o_i to the GRPO loss, and i ranges over the optimization subset. For a small update $W' = W - \lambda g_{i,t}$ with step size $\lambda > 0$, the first-order change in the reference loss is $-\lambda\alpha_{i,t}$. Minimizing $\mathcal{L}_{\text{AS}}$ thus favors temperatures whose resulting step-specific gradients align with the reference gradient.

The gradients $g_{i,t}$ are computed through the reconstructed soft states and remain differentiable with respect to the temperatures. Holding G_{ref} and the GRPO logit residuals fixed, differentiation provides a first-order alignment signal to temperatures through subsequent reasoning steps. To emphasize differences across reasoning steps, the mean temperature gradient is subtracted within each rollout:

$$\tilde{g}_{i,t}^{\tau} = \frac{\partial \mathcal{L}_{\text{AS}}}{\partial \tau_{i,t}} - \frac{1}{T_i}\sum_{t'=1}^{T_i}\frac{\partial \mathcal{L}_{\text{AS}}}{\partial \tau_{i,t'}}. \tag{11}$$

The centered gradients satisfy $\sum_{t=1}^{T_i} \tilde{g}_{i,t}^{\tau} = 0$. Alignment scores are computed using the outer-product structure of the output-layer gradients, avoiding a separate $|\mathcal{V}| \times d$ gradient matrix for each soft reasoning step.

Table 1: OmniSpatial Jia et al. (2025) results across 10 spatial reasoning task categories. Soft Spatial Reasoning is compared against proprietary models, general open-source LVLMs, soft thinking models, and specialized spatial reasoning models. Proprietary models are included as reference points; the best open-source or specialized result is shown in **bold**. Average accuracy is weighted by category sample size.

Method	Average	Dynamic Reasoning		Spatial Interaction			Complex Logic		Perspective Taking		
		Mani-pulation	Motion Anal.	Traffic Anal.	Loca-lization	Geospa. Strategy	Pattern Rec.	Geometric Reasoning	Ego Centric	Allo Centric	Hypo-thetical
Reference Baselines											
Random Choice	24.98	24.86	26.30	25.88	23.43	27.27	21.44	24.77	22.55	24.84	25.78
Human Evaluation	92.63	94.62	96.07	91.38	95.11	92.15	89.02	85.90	98.53	94.30	90.26
Proprietary Models											
GPT-4.1-mini Open (2025)	48.87	64.32	56.53	59.06	60.19	56.36	29.28	30.19	72.55	39.57	39.28
Gemini-2.5-flash-preview Team et al. (2023)	52.12	67.57	62.72	68.24	73.33	60.91	38.14	34.19	75.49	35.90	33.73
o4-mini OpenAI (2025)	52.77	72.97	59.83	60.00	73.33	61.82	34.02	36.77	73.53	40.69	40.96
Gemini-2.5-flash Team et al. (2023)	53.16	70.27	64.74	61.18	72.38	58.18	35.05	36.13	74.12	40.96	32.53
Open-weights Models											
LLaVA-1.5-7B Liu et al. (2024)	34.97	54.46	31.23	35.29	36.19	33.94	29.01	24.18	55.60	34.66	36.14
InternVL3-8B Zhu et al. (2025)	41.60	52.43	40.87	48.94	51.05	44.77	24.95	28.63	64.20	38.62	40.96
InternVL3-14B Zhu et al. (2025)	45.94	54.32	60.17	50.35	51.81	51.45	28.04	28.26	68.04	35.37	34.46
Qwen2.5-VL-7B Wang et al. (2024)	39.18	58.38	35.09	50.12	45.33	44.00	31.13	29.42	64.51	33.19	37.35
Gemma-3-12B Gemma Team et al. (2025)	43.71	54.05	54.91	54.12	47.62	45.45	16.49	30.32	63.73	36.70	33.73
Soft Thinking Models											
LVR Li et al. (2026a)	41.86	54.05	35.83	44.70	53.33	48.18	28.86	27.74	72.54	38.29	**50.60**
Laser Wang et al. (2026b)	42.78	55.40	45.95	57.64	54.28	50.00	20.61	24.51	72.54	33.51	44.57
LaCoT Sun et al. (2026)	45.66	59.45	53.75	56.47	50.47	48.18	28.86	33.54	67.64	34.57	44.57
Spatial Reasoning Models											
SpaceMantis-13B Chen et al. (2024a)	36.36	47.03	36.59	40.94	34.86	33.09	22.27	24.39	49.22	38.25	39.28
SpaceQwen2.5-VL-3B Chen et al. (2024a)	40.25	58.11	39.88	41.18	40.95	40.91	29.90	25.81	63.73	**38.83**	39.76
SpaceThinkerQwen2.5VL-3B Chen et al. (2024a)	40.42	47.84	53.06	43.29	35.43	38.73	24.33	28.00	58.04	35.11	31.08
VST-RL-7B Yang et al. (2025b)	41.09	56.75	43.39	44.75	46.66	42.72	25.51	28.38	72.54	32.89	43.37
SoFar-Qwen2.5VL-3B Qi et al. (2025)	45.14	56.49	51.16	54.12	53.14	**52.73**	**31.75**	22.88	71.60	36.56	41.69
SpatialLadder-3B Li et al. (2025)	40.50	59.45	39.01	50.58	48.57	42.72	26.80	23.87	71.56	35.10	39.75
Backbone and Our Variants											
Qwen3-VL-8B-Thinking (Base) Bai et al. (2025)	43.90	57.14	54.28	36.04	54.95	48.67	24.77	23.12	71.56	31.11	38.82
Hard Thinking + GRPO	45.92	61.32	54.30	55.88	58.33	46.18	25.90	25.00	72.92	36.57	42.53
Soft Thinking + GRPO	46.93	61.82	55.85	52.21	62.62	49.77	25.26	25.00	**76.35**	36.23	45.88
Soft Spatial Reasoning (Ours)	**49.68**	**63.51**	**62.14**	**57.65**	**62.86**	50.00	29.90	**34.84**	75.49	34.93	46.34

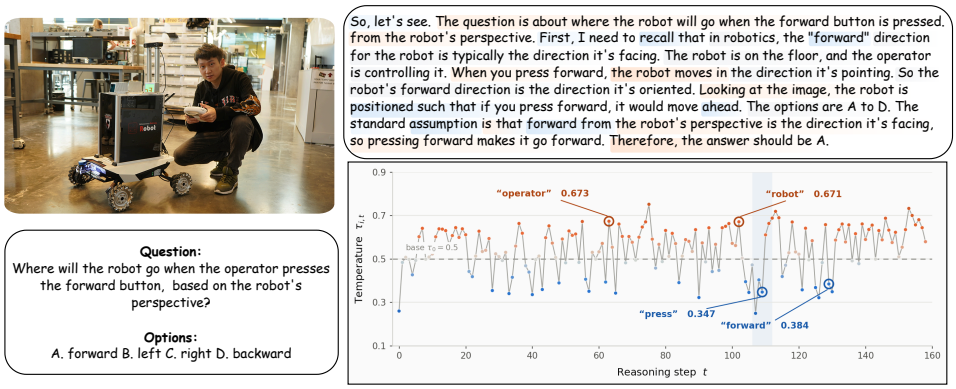


Figure 4: Soft Spatial Reasoning on a randomly selected example from OmniSpatial. AdaptSoft varies softness across the CoT. For readability, greedy decoding is used only to display the soft states as text. Shading shows each span's mean temperature (blue: lower; orange: higher). The plot below shows temperature at every reasoning step and labels several words with their temperatures.

Parameter Updates. Each training step updates the LVLM parameters θ using $\mathcal{L}_{\text{GRPO}}$ from Equation 9. The AdaptSoft parameters ϕ are updated by backpropagating the centered temperature gradients from Equation 11 through the controller in Equation 8 (detailed in Appendix Section C.3).

4 Experiments

4.1 Implementation Details

We instantiate Soft Spatial Reasoning with a Qwen3-VL-8B-Thinking backbone Bai et al. (2025) and train for two epochs on the OmniSpatial training split Jia et al. (2025). GRPO uses $G = 8$ rollouts per prompt, a batch size of 64 prompts, and two optimizer updates per batch. The LVLM policy is optimized with AdamW at a learning rate of 10^{-6} and regularized toward its frozen initialization using a KL coefficient of 10^{-3}. AdaptSoft is implemented as a two-layer MLP with projection dimension $d_p = 8$ and optimized with AdamW at a learning rate of 10^{-3}. Additional optimization, preprocessing, sequence-length, and system details are provided in Appendix Section B.1.

Table 2: SpatiaLab Wasi et al. (2026) results across 6 spatial reasoning task categories in the zero-shot setting. Proprietary models are included as reference points; the best open-source or specialized result is shown in **bold**. Average accuracy is weighted by category sample size.

Method	Average	Question Categories					
		3D Geom.	Dep. & Occu.	Orientation	Relat. Posit.	Size & Scale	Spati. Navig.
Reference Baselines							
Random Choice	25.00	25.00	25.00	25.00	25.00	25.00	25.00
Human Baseline	87.57	93.70	74.13	91.58	91.51	88.89	87.76
Proprietary Models							
GPT-4o-mini Hurst et al. (2024)	46.50	47.06	39.00	47.03	47.17	49.60	49.79
Gemini-2.5-Flash Team et al. (2023)	48.29	44.96	48.26	48.02	56.13	42.46	51.05
Mistral Medium 3.1 Mistral AI (2025)	47.93	46.64	49.81	47.52	61.79	41.67	41.77
Open-weights Models							
Qwen2.5-VL-3B-Instruct Wang et al. (2024)	41.43	41.18	35.52	46.04	40.09	47.22	39.24
InternVL3.5-4B Wang et al. (2025b)	43.29	42.86	42.86	42.08	54.72	36.51	42.19
Gemma-3-4B-it Gemma Team et al. (2025)	40.57	43.70	34.36	46.53	45.75	37.30	37.97
LLaVA-1.5-7B Liu et al. (2024)	38.64	40.33	34.74	31.68	40.56	40.07	43.88
Qwen2.5-VL-7B-Instruct Wang et al. (2024)	41.00	42.86	37.84	42.57	46.23	42.06	35.44
Soft Thinking Models							
LVR Li et al. (2026a)	45.78	**48.31**	43.62	**51.98**	49.52	40.07	43.03
Laser Wang et al. (2026b)	44.78	46.21	43.24	46.03	50.47	40.07	43.88
LaCoT Sun et al. (2026)	45.50	46.63	40.15	47.52	53.30	43.25	43.88
Spatial Reasoning Models							
SpaceOm Chen et al. (2024a)	41.36	42.44	38.61	48.02	37.74	42.86	39.24
SpaceThinker-Qwen2.5VL-3B Chen et al. (2024a)	40.64	40.34	37.84	47.03	38.21	43.25	37.97
SpaceQwen2.5-VL-3B-Instruct Chen et al. (2024a)	40.14	31.51	35.14	37.62	37.74	**50.79**	47.26
SpatialLadder-3B Li et al. (2025)	35.28	39.91	34.61	39.90	38.20	25.39	33.19
Backbone and Our Variants							
Qwen3-VL-8B-Thinking (Base) Bai et al. (2025)	44.78	42.01	47.10	47.02	51.88	40.47	41.35
Hard Thinking + GRPO	46.21	43.70	47.88	47.52	54.25	41.27	43.88
Soft Thinking + GRPO	46.86	44.12	48.26	48.02	55.66	41.67	44.73
Soft Spatial Reasoning (Ours)	**48.71**	46.22	**49.42**	48.51	**58.96**	42.46	**48.10**

4.2 BASELINES

We compare Soft Spatial Reasoning against open-source LVLMs, including models specialized in spatial reasoning and soft thinking, and report proprietary models as reference points. We also train controlled hard- and soft-thinking baselines using the same backbone and post-training setup. Evaluation covers three complementary settings. The held-out OmniSpatial test set Jia et al. (2025) evaluates post-training performance across its spatial-reasoning taxonomy. SpatiaLab Wasi et al. (2026) measures zero-shot transfer to realistic, unconstrained scenes from an unseen benchmark. MindCube Wang et al. (2026a) evaluates zero-shot spatial mental modeling from limited views through cognitive mapping, perspective-taking, and mental simulation. Dataset and evaluation details are provided in Appendix Section D.1.

4.3 RESULTS

What Adaptive Softness Adds to Spatial Reasoning. On OmniSpatial, Soft Spatial Reasoning leads the non-proprietary models in weighted-average accuracy (Table 1), surpassing InternVL3-14B, SoFar, and LaCoT by 3.74, 4.54, and 4.02 percentage points, respectively. With the same backbone and post-training setup, basic soft thinking improves on hard thinking by 1.01 points, while adaptive softness adds another 2.75 points. Geometric Reasoning illustrates this distinction: hard and basic soft thinking both score 25.00, whereas adaptation reaches 34.84. Motion Analysis also rises from 55.85 to 62.14 over basic soft thinking. The absence of a geometric gain from basic soft thinking suggests that preserving alternatives alone may not suffice. Geometry and motion can involve competing spatial configurations whose usefulness changes across reasoning steps. Adaptive softness may help retain useful possibilities while limiting interference from conflicting relations as the reasoning context changes. Relative to the original backbone, the full model improves in all ten categories, including a 21.61-point gain in Traffic Analysis.

Figure 4 illustrates how Soft Spatial Reasoning answers a spatial question about an image. The displayed CoT traces the model's reasoning about the robot's movement, reaching the correct answer as softness varies across steps. The text highlights connect this reasoning to the temperature plot below: "operator" and "robot" correspond to higher temperatures, while "press" and "forward" correspond to lower temperatures. These adjustments let broader candidate continuations contribute at some steps and concentrate their influence at others, showing how adaptive softness operates throughout reasoning. Further analysis of temperature variation appears in Section A.

What Adaptive Softness Adds in Zero-Shot Transfer. On the unseen SpatiaLab benchmark (Table 2), Soft Spatial Reasoning achieves 48.71, exceeding the strongest prior open-weights model, LVR, by 2.93 percentage points. The matched-backbone comparison shows that adaptive softness contributes a further 1.85 points over basic soft thinking, compared with the 0.65-point gain from hard to basic soft thinking. Thus, the larger benefit from adjusting softness across reasoning steps observed on OmniSpatial persists under zero-shot transfer. Unlike on OmniSpatial, adaptive softness improves over basic soft thinking in every category. The gains are nevertheless uneven: Relative Position (+3.30) and Spatial Navigation (+3.37) improve substantially more than Orientation (+0.49) and Size & Scale (+0.79). The consistent direction of these improvements supports the transferability of learned softness control, while their differing magnitudes indicate that its benefit depends on the spatial task.

Table 3: Zero-shot results on MindCube Wang et al. (2026a) across three spatial mental-modeling settings. Proprietary models are included as reference points, and Overall is weighted by setting size.

Method	Overall	Rotation	Among	Around
Reference Baseline				
Random Choice	32.35	36.36	32.29	30.66
Proprietary Models				
Gemini-2.5-Pro Team et al. (2023)	47.05	85.50	25.95	38.40
Claude-4-Sonnet Anthropic (2024)	44.75	48.42	44.21	47.62
Open-weight Models				
InternVL3-8B Wang et al. (2025b)	37.50	26.00	**42.03**	**36.00**
Qwen3-VL-8B-Thinking (Base) Bai et al. (2025)	33.62	45.14	33.26	30.44
Soft-Thinking Models				
LVR Li et al. (2026a)	29.53	36.07	29.44	26.59
Laser Wang et al. (2026b)	27.29	37.65	30.12	21.02
Spatial Reasoning Models				
SpaceMantis Chen et al. (2024a)	22.81	37.65	21.26	29.32
SpaceQwen Chen et al. (2024a)	33.28	38.02	33.71	26.32
Proposed Method				
Soft Spatial Reasoning (Ours)	**38.13**	**47.36**	38.18	32.37

What Adaptive Softness Adds to Spatial Mental Modeling. MindCube Wang et al. (2026a) requires integrating limited views to infer unseen spatial relations and reason about hypothetical movements. In this zero-shot setting, Soft Spatial Reasoning achieves 38.13 weighted-average accuracy, the highest among the compared non-proprietary models, and improves on its backbone by 4.51 points (Table 3). Improvements span all three settings, extending the framework's benefits to reasoning about partially observed scenes. Adaptive softness may support this process by preserving plausible spatial interpretations while combining evidence from different views.

Ablation Study. All four ablations reduce accuracy in every perspective-taking category (Table 4). Replacing AdaptSoft's step-specific gradient alignment with the rollout-level task advantage produces the largest average decline (4.07 points), including a 6.52-point drop in *Hypothetical* reasoning. This suggests that the shared GRPO signal provides less guidance for controlling softness at individual steps. Removing the final-layer hidden state $h_{i,t}$ from AdaptSoft reduces average accuracy by 3.53 points, compared with 2.11 points without predictive uncertainty, supporting the use of the ongoing reasoning state alongside predictive uncertainty to control softness. Omitting gradient centering reduces accuracy in all three categories, with a smaller average decline of 1.57 points.

Table 4: OmniSpatial *Perspective Taking* ablations Jia et al. (2025). All variants share the backbone and policy GRPO setup, with post-training and evaluation on the corresponding training and test splits. Averages are sample-weighted.

Variant	Avg.	Ego centric	Allo centric	Hypo thetical
Soft Spatial Reasoning (Full)	**42.75**	**81.37**	**32.53**	**41.46**
w/o $h_{i,t}$ in AdaptSoft	39.22	77.45	29.52	36.14
w/o predictive uncertainty	40.64	78.43	30.85	38.55
w/o gradient alignment	38.68	76.47	29.26	34.94
w/o gradient centering	41.18	79.41	31.12	39.76

5 Conclusion

We introduced **Soft Spatial Reasoning**, a post-training framework that enables adaptive soft thinking for spatial reasoning in LVLMs. **AdaptSoft** controls softness using the reasoning state and predictive uncertainty, with step-specific guidance from gradient alignment. Across three benchmarks, our framework achieves the highest weighted-average accuracy among the evaluated non-proprietary models, with gains extending to unseen benchmarks.

Limitations. AdaptSoft lacks an explicit measure of uncertainty in visual evidence. Future work could incorporate visual uncertainty at individual CoT steps, allowing softness to reflect ambiguity in both visual evidence and language generation.

Ethics Statement

This work studies spatial reasoning in LVLMs using existing question–answering benchmarks. Improved benchmark accuracy does not establish reliability in applications involving navigation or interaction with the physical world, where incorrect spatial judgments could lead to harmful decisions. Such applications require additional evaluation and safeguards beyond those studied here.

Reproducibility Statement

Section 3 specifies the construction of soft states, the AdaptSoft controller, and the post-training objectives. Section 4 describes the training and evaluation setup, baselines, and metrics. The appendix provides implementation details, hyperparameters, benchmark splits and evaluation protocols, and derivations of the training objectives. We provide the full source code as supplementary material and will release the trained model weights after acceptance.

AI Use Statement

Generative AI tools were used to polish the manuscript's writing for clarity, readability, and consistency. The authors reviewed the suggested edits and take responsibility for the final text, methods, experiments, results, and claims.

References

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

Sonnet Anthropic. Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet. *URL https://api. semanticscholar. org/CorpusID*, 273639283:24, 2024.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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

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

Natasha Butt, Ariel Kwiatkowski, Ismail Labiad, Julia Kempe, and Yann Ollivier. Soft tokens, hard truths. In *International Conference on Learning Representations*, volume 2026, pp. 114650–114675, 2026.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9490–9498. IEEE, 2025a.

Zhipeng Cai, Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. Depthlm: Metric depth from vision language models. *arXiv preprint arXiv:2509.25413*, 2025b.

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

Chao Chen, Zhixin Ma, Yongqi Li, Yupeng Hu, Yinwei Wei, Wenjie Li, and Liqiang Nie. Reasoning in the dark: Interleaved vision-text reasoning in latent space. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 39117–39129, 2026a.

Pingyi Chen, Yujing Lou, Shen Cao, Jinhui Guo, Lubin Fan, Yue Wu, Lin Yang, Lizhuang Ma, and Jieping Ye. Sd-vlm: Spatial measuring and understanding with depth-encoded vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. *arXiv preprint arXiv:2503.01773*, 2025a.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26428–26438, 2024b.

Siyi Chen, Mikaela Angelina Uy, Chan Hee Song, Faisal Ladhak, Adithyavairavan Murali, Qing Qu, Stan Birchfield, Valts Blukis, and Jonathan Tremblay. Spacetools: Tool-augmented spatial reasoning via double interactive rl. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37109–37120, 2026b.

Zhangquan Chen, Ruihui Zhao, Chuwei Luo, Mingze Sun, Xinlei Yu, Yangyang Kang, and Ruqi Huang. Sifthinker: Spatially-aware image focus for visual reasoning. *arXiv preprint arXiv:2508.06259*, 2025b.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, et al. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2613–2624, 2026c.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Haoran Dang, Cuiling Lan, Hai Wan, Xibin Zhao, and Yan Lu. Temperature as a meta-policy: Adaptive temperature in llm reinforcement learning. *arXiv preprint arXiv:2602.11779*, 2026.

Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms, 2025. *URL https://arxiv. org/abs/2503.13111 v2*.

Evelina Fedorenko and Rosemary Varley. Language and thought are not the same thing: evidence from neuroimaging and neurological patients. *Annals of the New York Academy of Sciences*, 1369 (1):132–153, 2016.

Gemma Team et al. Gemma 3 technical report, 2025. URL `https://arxiv.org/abs/2503.19786`.

Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Sitong Mao, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. Spatial reasoning with vision-language models in ego-centric multi-view scenes. *arXiv preprint arXiv:2509.06266*, 2025.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 3d concept learning and reasoning from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9202–9212, 2023.

Lianyu Hu, Shengqian Qin, Zeqin Liao, Qing Guo, Liang Wan, Wei Feng, and Yang Liu. Colt: Teaching multi-modal models to think with chain of latent thoughts. *arXiv preprint arXiv:2606.31986*, 2026.

Wenbo Hu, Jingli Lin, Yilin Long, Yunlong Ran, Lihan Jiang, Yifan Wang, Chenming Zhu, Runsen Xu, Tai Wang, and Jiangmiao Pang. G^2-VLM: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. *arXiv preprint arXiv:2511.21688*, 2025.

David Huang and Lianlei Shan. Dlwm: Diverse latent world models for efficient multimodal reasoning. *arXiv preprint arXiv:2606.15160*, 2026.

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

Byungwoo Jeon, Yoonwoo Jeong, Hyunseok Lee, Minsu Cho, and Jinwoo Shin. Vision-aligned latent reasoning for multi-modal large language model. *arXiv preprint arXiv:2602.04476*, 2026.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025.

Sai Srinivas Kancheti, Aditya Kanade, Rohit Sinha, Vineeth N Balasubramanian, and Tanuja Ganu. Faithful grpo: Improving visual spatial reasoning in multimodal language models via constrained policy optimization. *arXiv preprint arXiv:2604.08476*, 2026.

Phillip Y Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9241–9251, 2025.

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

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

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

Lin Li, Guikun Chen, Hanrong Shi, Jun Xiao, and Long Chen. A survey on multimodal benchmarks: In the era of large ai models. *arXiv preprint arXiv:2409.18142*, 2024.

Zongzhao Li, Zongyang Ma, Mingze Li, Songyou Li, Yu Rong, Tingyang Xu, Ziqi Zhang, Deli Zhao, and Wenbing Huang. Star-r1: Multi-view spatial transformation reasoning by reinforcing multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12041–12051, 2026c.

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

Disheng Liu, Tuo Liang, Zhe Hu, Jierui Peng, Yiren Lu, Yi Xu, Yun Fu, and Yu Yin. Spatial intelligence in vision-language models: A comprehensive survey. 2026.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. Ssr: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. *arXiv preprint arXiv:2505.12448*, 2025.

Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors. *Advances in neural information processing systems*, 37:68803–68832, 2024.

Jizheng Ma, Xiaofei Zhou, Yanlong Song, and Han Yan. Cocova: Chain of continuous vision-language thought for latent space reasoning. *arXiv e-prints*, pp. arXiv–2511, 2025a.

Weijian Ma, Shizhao Sun, Tianyu Yu, Ruiyu Wang, Tat-Seng Chua, and Jiang Bian. Thinking with blueprints: Assisting vision-language models in spatial reasoning via structured object representation. *arXiv preprint arXiv:2601.01984*, 2026.

Wufei Ma, Luoxin Ye, Celso M de Melo, Alan Yuille, and Jieneng Chen. Spatialllm: A compound 3d-informed design towards spatially-intelligent large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17249–17260, 2025b.

Mistral AI. Mistral medium 3.1. `https://docs.mistral.ai/models/mistral-medium-3-1-25-08`, August 2025. Model version: mistral-medium-2508.

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

Zhenhua Ning, Zhuotao Tian, Shaoshuai Shi, Guangming Lu, Daojing He, Wenjie Pei, and Li Jiang. Enhancing spatial reasoning in multimodal large language models through reasoning-based segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7851–7860, 2025.

AI Open. Introducing gpt-4.1 in the api, 2025.

OpenAI. Openai o3 and o4-mini system card. `https://openai.com/index/o3-o4-mini-system-card/`, April 2025. Accessed: 2026-04-21.

Tan-Hanh Pham and Chris Ngo. Multimodal chain of continuous thought for latent-space reasoning in vision-language models, 2026. URL `https://openreview.net/forum?id=UhkMZDmp4J`.

Zekun Qi, Wenyao Zhang, Yufei Ding, Runpei Dong, Xinqiang Yu, Jingwen Li, Lingyun Xu, Baoyu Li, Xialin He, Guofan Fan, et al. Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. *arXiv preprint arXiv:2502.13143*, 2025.

R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005.

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

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. *arXiv preprint arXiv:2505.23678*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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.

Dachuan Shi, Abedelkadir Asi, Keying Li, Xiangchi Yuan, Leyan Pan, Wenke Lee, and Wen Xiao. Swireasoning: Switch-thinking in latent and explicit for pareto-superior reasoning llms. In *International Conference on Learning Representations*, volume 2026, pp. 137060–137093, 2026.

Ilias Stogiannidis, Steven McDonagh, and Sotirios A Tsaftaris. Mind the gap: Benchmarking spatial reasoning in vision-language models. *arXiv preprint arXiv:2503.19707*, 2025.

Rafi Ibn Sultan, Hui Zhu, Xiangyu Zhou, Chengyin Li, Prashant Khanduri, Marco Brocanelli, and Dongxiao Zhu. Walkgpt: Grounded vision-language conversation with depth-aware segmentation for pedestrian navigation. *arXiv preprint arXiv:2603.10703*, 2026.

Guohao Sun, Hang Hua, Jian Wang, Jiebo Luo, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Latent chain-of-thought for visual reasoning. *Advances in neural information processing systems*, 38:103739–103762, 2026.

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

Jiaxu Wan, Xu Wang, Mengwei Xie, Hang Zhang, Mu Xu, Yang Han, Hong Zhang, Ding Yuan, and Yifan Yang. Eaglevision: A dual-stage framework with bev-grounding-based chain-of-thought for spatial intelligence. *arXiv preprint arXiv:2512.15160*, 2025.

Kang Wang, Xiangyu Duan, and Tianyi Du. Improving latent reasoning in llms via soft concept mixing. *arXiv preprint arXiv:2511.16885*, 2025a.

Peiyao Wang and Haibin Ling. Svqa-r1: Reinforcing spatial reasoning in mllms via view-consistent reward optimization. *arXiv preprint arXiv:2506.01371*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. Mindcube: Spatial mental modeling from limited views, 2026a. URL https://arxiv.org/abs/2506.21458.

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*, 2025b.

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

Yuxin Wang, Lei Ke, Boqiang Zhang, Tianyuan Qu, Hanxun Yu, Zhenpeng Huang, Meng Yu, Dan Xu, and Dong Yu. N3d-vlm: Native 3d grounding enables accurate spatial reasoning in vision-language models. *arXiv preprint arXiv:2512.16561*, 2025c.

Azmine Toushik Wasi, Wahid Faisal, Abdur Rahman, Mahfuz Ahmed Anik, Munem Shahriar, Mohsin Mahmud Topu, Sadia Tasnim Meem, Rahatun Nesa Priti, Sabrina Afroz Mitu, Md Iqramul Hoque, et al. Spatialab: Can vision-language models perform spatial reasoning in the wild? *arXiv preprint arXiv:2602.03916*, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Shuyu Wei, Jian Sun, Delai Qiu, Yining Wang, Shengping Liu, Jiaen Liang, Ying Fu, Wei Huang, and Jitao Sang. Taming the thinker: Conditional entropy shaping for adaptive llm reasoning. *arXiv preprint arXiv:2605.19358*, 2026a.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. Sim-cot: Supervised implicit chain-of-thought. In *International Conference on Learning Representations*, volume 2026, pp. 56721–56742, 2026b.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025a.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025b.

Junhong Wu, Jinliang Lu, Zixuan Ren, Gangqiang Hu, Zhi Wu, Dai Dai, et al. Llms are single-threaded reasoners: Demystifying the working mechanism of soft thinking. In *International Conference on Learning Representations*, volume 2026, pp. 4533–4548, 2026.

Beining Xu, Siting Zhu, Zhao Jin, Junxian Li, and Hesheng Wang. S^2-MLLM: Boosting spatial reasoning capability of MLLMs for 3D visual grounding with structural guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2557–2569, 2026a.

Xin Xu, Tong Yu, Xiang Chen, Haoliang Wang, Julian McAuley, and Saayan Mitra. Thinkrouter: Efficient reasoning via routing thinking between latent and discrete spaces. *arXiv preprint arXiv:2602.11683*, 2026b.

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. *arXiv preprint arXiv:2505.11409*, 2025.

Zhongxing Xu, Zhonghua Wang, Zhe Qian, Dachuan Shi, Feilong Tang, Ming Hu, Shiyan Su, Xiaocheng Zou, Wei Feng, Dwarikanath Mahapatra, et al. Thinking in uncertainty: Mitigating hallucinations in mlrms with latent entropy-aware decoding. *arXiv preprint arXiv:2603.13366*, 2026c.

Jiadong Yan, Ke Zhang, Chenyang Zhao, Shoushan Li, and Xizhao Luo. Grasp: Awakening latent spatial reasoning in lvlms via training-free geometric rectification. In *Forty-third International Conference on Machine Learning*, 2026.

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.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025b.

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

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 12 2024. ISSN 2095-5138. doi: 10.1093/nsr/nwae403. URL https://doi.org/10.1093/nsr/nwae403.

Zhiyuan Yu, Shijian Xiao, Cam-Tu Nguyen, Zhangyue Yin, Lekai Xing, Wenzhong Li, and Sanglu Lu. Thermometer of thoughts: Enhancing llm's exploration via attention temperature modulation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4355–4368, 2026.

Yue Zhang, Zun Wang, Han Lin, Yonatan Bitton, Idan Szpektor, and Mohit Bansal. Seeing isn't knowing: Do vlms know when not to answer spatial questions (and why)? *arXiv preprint arXiv:2605.30557*, 2026a.

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

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=84pDoCD4lH.

Baining Zhao, Ziyou Wang, Jianjie Fang, Chen Gao, Fanhang Man, Jinqiang Cui, Xin Wang, Xinlei Chen, Yong Li, and Wenwu Zhu. Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11071–11080, 2025a.

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. *arXiv preprint arXiv:2511.23075*, 2025b.

Zhi Zheng, Yu Gu, Wei Liu, Yee Whye Teh, and Wee Sun Lee. Soft-grpo: Surpassing discrete-token llm reinforcement learning via gumbel-reparameterized soft-thinking policy optimization. *arXiv preprint arXiv:2511.06411*, 2025a.

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

Shengchao Zhou, Yuxin Chen, Yuying Ge, Wei Huang, Jiehong Lin, Ying Shan, and Xiaojuan Qi. Learning to reason in 4d: Dynamic spatial understanding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9637–9646, 2026a.

Yixiao Zhou, Yang Li, Dongzhou Cheng, Hehe Fan, and Yu Cheng. Look inward to explore outward: Learning temperature policy from llm internal states via hierarchical rl. *arXiv preprint arXiv:2602.13035*, 2026b.

Yuyan Zhou, Jiarui Yu, Hande Dong, Zhezheng Hao, Hong Wang, Jianqing Zhang, and Qiang Lin. Lepo: Latent reasoning policy optimization for large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 14416–14427, 2026c.

Hanlin Zhu, Shibo Hao, Zhiting Hu, Jiantao Jiao, Stuart J Russell, and Yuandong Tian. Reasoning by superposition: A theoretical perspective on chain of continuous thought. *Advances in Neural Information Processing Systems*, 38:79931–79963, 2026.

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

A Appendix: Additional Analysis

AdaptSoft Learns to Vary Softness. As training progresses, Figure 5 shows AdaptSoft assigning temperatures over a wider range while their mean stays close to $\tau_0 = 0.5$ (0.488 to 0.510). The min–max span grows from 0.23 to 0.58, extending on both sides of the base value. AdaptSoft thus produces broader mixtures at some steps and more concentrated ones at others without shifting average softness. This spread develops over successive updates and persists later in training, revealing learned variation that the stable mean would obscure. Basic soft thinking can vary mixture weights as token probabilities change, but its temperature remains fixed. AdaptSoft also learns how soft each mixture should be; Figure 4 illustrates these adjustments within an individual soft CoT.

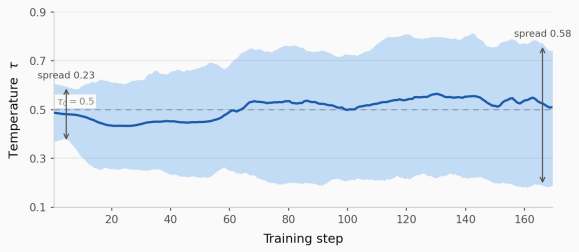


Figure 5: **AdaptSoft Temperatures During Training.** The line shows the mean temperature $\tau_{i,t}$ across soft steps in each rollout batch; shading spans their minimum and maximum. The dashed line marks $\tau_0 = 0.5$.

Predictive Uncertainty and Softness. Figure 6 plots the temperature and predictive uncertainty at each soft reasoning step across inference rollouts of the trained model. As normalized top-k entropy increases, median temperature falls from 0.590 to 0.350, with a correlation of -0.613 across all steps. When probability is spread across more candidates, these lower temperatures concentrate their influence and may reduce interference from conflicting spatial interpretations. AdaptSoft also assigns substantially different temperatures at similar uncertainty. Within every entropy bin, the 10th–90th percentile range is 0.18–0.20, nearly as large as the 0.24 change in median across the full entropy range. Even the lowest-entropy bin, containing 48.4% of steps, spans 0.470–0.670 between these percentiles. Removing the hidden-state input leaves AdaptSoft with predictive uncertainty alone and lowers average accuracy by 3.53 points (Table 4).

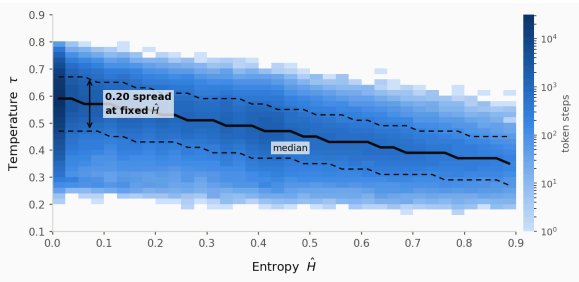


Figure 6: AdaptSoft Temperature and Predictive Uncertainty. Temperatures and predictive uncertainty are measured across soft reasoning steps during inference on the test set. Color shows the number of steps on a logarithmic scale. The solid line shows median temperature in each entropy bin; dashed lines mark the 10th and 90th percentiles.

AdaptSoft Varies Softness Within CoTs. Figure 7 compares temperature variation within individual test-set CoTs during inference with variation across their median temperatures. The median within-trace spread is 0.235, about 2.5 times the 0.093 spread across trace medians: AdaptSoft changes softness more as a CoT unfolds than it differs between CoTs. Even the smallest within-trace spread is 0.145, showing that these changes occur across the test set. Correct and incorrect traces have similar median spreads (0.233 and 0.237), so the variation is not limited to successful answers. It also persists in short and long traces (0.218 below 100 steps and 0.241 at

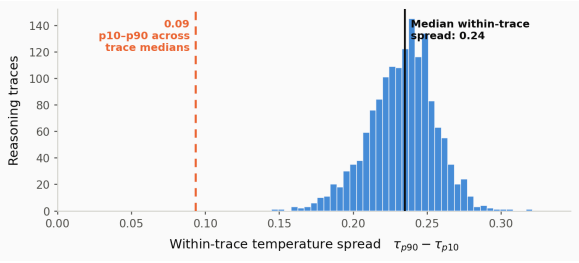


Figure 7: **AdaptSoft Varies Softness Within CoTs.** The histogram shows the 90th–10th percentile temperature spread within each test-set reasoning trace generated during inference. The solid line marks the median within-trace spread; the dashed line marks the corresponding spread across trace-median temperatures.

```
You are a spatial-reasoning assistant for visual multiple-choice
questions.

The answer options may be listed in the text, or they may appear
inside the image itself — if they are not in the text, read them
from the image. Inside <think>, briefly think step by step about
the image to work out the answer. Then close with </think> exactly
once and output exactly one capital letter: A, B, C, or D. Output
nothing after that letter.

Do not refuse. If uncertain, choose the most plausible answer based
on the image.
```

Figure 8: System prompt used to standardize the reasoning format during Soft Spatial Reasoning training.

300 steps or more), rather than arising simply because longer CoTs offer more steps over which temperature can change.

B Appendix: Implementation Details

B.1 Hyperparameter Settings

Table 5 reports the main hyperparameters used for Soft Spatial Reasoning post-training, including the GRPO training setup, optimization settings, and reward configuration.

B.2 Reasoning Format and Discrete Answer Generation

We use the fixed system prompt in Figure 8 during training to enforce a consistent reasoning format. Since the chat template of Qwen3-VL-Thinking Bai et al. (2025) automatically prefills the opening `<think>` token after the user message, the prompt does not ask the model to generate `<think>`; it only requires the model to close the reasoning segment with `</think>`.

During rollout generation, each reasoning step feeds a mixture of token embeddings back to the model. We also retain the highest-weight token at each step as a discrete *spine*, which marks where reasoning ends. When the spine emits `</think>`, subsequent steps sample discrete tokens from the full vocabulary to generate the final answer. The delimiter itself remains part of the soft reasoning phase; the switch occurs immediately afterward.

C Appendix: Technical Derivations

C.1 Derivation of the Gumbel-Reparameterized Likelihood

The likelihood in Equation 5 follows the Gumbel-reparameterized construction of SofT-GRPO Zheng et al. (2025a). The sampled variable at each soft reasoning step is the perturbed score vector $z_{i,t}$, while the soft state is constructed deterministically from $z_{i,t}$ and $\tau_{i,t}$. Conditioning on $(I, q, s_{i,<t})$ is omitted below for compactness.

During rollout, each score is obtained by adding independent standard Gumbel noise to the rollout-policy log-probability. Under the current policy, the same recorded score implies a different noise value:

$$\begin{aligned} z_{i,t,k} &= \log \bar{p}_{i,t,k} + \gamma_{i,t,k}, \qquad \gamma_{i,t,k} \overset{\text{i.i.d.}}{\sim} \text{Gumbel}(0,1), \\ f_{\text{Gum}}(\gamma) &= \exp[-\gamma - \exp(-\gamma)], \quad \widetilde{\gamma}^{\theta}_{i,t,k} = z_{i,t,k} - \log \bar{p}^{\theta}_{i,t,k}. \end{aligned} \tag{12}$$

The transformation from $\widetilde{\gamma}^{\theta}_{i,t,k}$ to $z_{i,t,k}$ is additive and therefore has unit Jacobian. Independence across vocabulary tokens gives the density and its logarithm, recovering Equation 5:

$$P_\theta(z_{i,t}) = \prod_{k=1}^{|\mathcal{V}|} f_{\mathrm{Gum}}\big(\widetilde{\gamma}^{\theta}_{i,t,k}\big), \qquad \log P_\theta(z_{i,t}) = \sum_{k=1}^{|\mathcal{V}|} \big[-\widetilde{\gamma}^{\theta}_{i,t,k} - \exp\big(-\widetilde{\gamma}^{\theta}_{i,t,k}\big)\big]. \tag{13}$$

Under the rollout policy, the implied noise equals the originally sampled value. Evaluating the same score vector under both policies therefore gives

$$\begin{gathered}\widetilde{\gamma}^{\theta_{\mathrm{old}}}_{i,t,k} = z_{i,t,k} - \log \bar{p}_{i,t,k} = \gamma_{i,t,k}, \qquad \log P_{\theta_{\mathrm{old}}}(z_{i,t}) = \sum_{k=1}^{|\mathcal{V}|} \left[-\gamma_{i,t,k} - \exp(-\gamma_{i,t,k})\right], \\ \rho^{\mathrm{soft}}_{i,t}(\theta) = \frac{P_\theta(z_{i,t})}{P_{\theta_{\mathrm{old}}}(z_{i,t})} = \exp(\log P_\theta(z_{i,t}) - \log P_{\theta_{\mathrm{old}}}(z_{i,t})).\end{gathered} \tag{14}$$

The likelihood ratio in Equation 6 evaluates the recorded perturbed scores $z_{i,t}$ under both policies. Equation 4 then constructs $s_{i,t}$ deterministically from $(z_{i,t}, \tau_{i,t})$, so the ratio requires no separate density over soft states.

C.2 Stop-Gradient Reconstruction for AdaptSoft

Training AdaptSoft requires gradients through the step-specific temperature, but retaining the language-backbone computation graph for every rollout step would be memory intensive. We therefore store the projected hidden state, standardized entropy, and AdaptSoft output during rollout. Let

$$x^{\mathrm{roll}}_{i,t} = \Big[\mathbf{P}\ \mathrm{LN}(h_{i,t}) \,\Big\|\, \widehat{H}_{i,t}\Big], \qquad u^{\mathrm{roll}}_{i,t} = f_{\phi_{\mathrm{roll}}}\big(x^{\mathrm{roll}}_{i,t}\big) \tag{15}$$

denote the stored input and output, where ϕ_{roll} denotes the AdaptSoft parameters used to generate the rollout.

During the update, AdaptSoft recomputes its output from the stored input using the current parameters ϕ. The operator $\mathrm{sg}(\cdot)$ stops gradients through recorded values; the construction below preserves the rollout temperature in the forward pass while allowing gradients to reach ϕ:

$$\begin{gathered}\widetilde{u}_{i,t} = f_\phi\big(\mathrm{sg}(x^{\mathrm{roll}}_{i,t})\big), \\ u^{\mathrm{upd}}_{i,t} = \mathrm{sg}(u^{\mathrm{roll}}_{i,t}) + \widetilde{u}_{i,t} - \mathrm{sg}(\widetilde{u}_{i,t}), \qquad \tau^{\mathrm{upd}}_{i,t} = \tau_0 + \Delta \tanh(u^{\mathrm{upd}}_{i,t}), \\ u^{\mathrm{upd}}_{i,t} = u^{\mathrm{roll}}_{i,t}, \qquad \tau^{\mathrm{upd}}_{i,t} = \tau^{\mathrm{roll}}_{i,t} \quad \text{(forward pass)}.\end{gathered} \tag{16}$$

With the recorded perturbed scores, the forward pass reproduces the rollout mixture weights and, if token embeddings are unchanged, the rollout soft state. Gradients to ϕ pass through the recomputed AdaptSoft output:

$$\nabla_\phi u^{\mathrm{upd}}_{i,t} = \nabla_\phi \widetilde{u}_{i,t}, \qquad \nabla_\phi \tau^{\mathrm{upd}}_{i,t} = \Delta\left[1 - \tanh^2\big(u^{\mathrm{roll}}_{i,t}\big)\right] \nabla_\phi \widetilde{u}_{i,t}. \tag{17}$$

Recomputing AdaptSoft from its stored, detached input gives the update a gradient path to ϕ without another backbone pass to obtain $h_{i,t}$ or backpropagation through that stored state.

C.3 Differentiable Reconstruction and AdaptSoft Gradient Propagation

AdaptSoft must preserve the temperatures used to generate each rollout while retaining a gradient path to its current parameters. Let $\mathrm{sg}(\cdot)$ denote stop-gradient, and define the recorded AdaptSoft input as $\xi_{i,t} = [\mathbf{P}\,\mathrm{LN}(h_{i,t}) \| \widehat{H}_{i,t}]$. During rollout, AdaptSoft records $\xi_{i,t}$ and its output $u^{\mathrm{roll}}_{i,t}$, produced using the rollout parameters ϕ_{roll}. At update time, the output is recomputed using the current parameters and combined with its recorded value:

$$\begin{gathered}\widetilde{u}_{i,t} = f_\phi(\mathrm{sg}(\xi_{i,t})), \qquad u^{\mathrm{upd}}_{i,t} = \mathrm{sg}\big(u^{\mathrm{roll}}_{i,t}\big) + \widetilde{u}_{i,t} - \mathrm{sg}(\widetilde{u}_{i,t}), \\ \tau^{\mathrm{upd}}_{i,t} = \tau_0 + \Delta \tanh\Big(u^{\mathrm{upd}}_{i,t}\Big).\end{gathered} \tag{18}$$

In the forward pass, $\widetilde{u}_{i,t} - \mathrm{sg}(\widetilde{u}_{i,t}) = 0$; hence $u_{i,t}^{\mathrm{upd}} = u_{i,t}^{\mathrm{roll}}$ and $\tau_{i,t}^{\mathrm{upd}} = \tau_{i,t}^{\mathrm{roll}}$. The recorded perturbed scores and candidate token identities consequently reproduce the rollout temperature and mixture weights even if ϕ has changed since rollout generation.

The stop-gradient terms vanish in the backward pass, allowing the recomputed output to carry gradients to ϕ. Together with the dependence of the mixture weights on temperature, this gives

$$\begin{aligned} \frac{\partial \tau_{i,t}^{\mathrm{upd}}}{\partial \phi} &= \Delta \left[1 - \tanh^2\left(u_{i,t}^{\mathrm{roll}}\right)\right] \frac{\partial \widetilde{u}_{i,t}}{\partial \phi}, \\ \frac{\partial p_{i,t,k}}{\partial \tau_{i,t}} &= -\frac{p_{i,t,k}}{\tau_{i,t}^2} \left(z_{i,t,k} - \sum_{j=1}^{|\mathcal{V}|} p_{i,t,j} z_{i,t,j} \right). \end{aligned} \tag{19}$$

Gradients can therefore pass through the mixture weights, the resulting soft states, and the subsequent language-model computation to the step-specific temperatures and AdaptSoft parameters. Equation 18 uses the recorded rollout temperatures in the forward pass and the recomputed AdaptSoft output to update ϕ.

Differentiable Gradient Alignment. Let $v_{i,t}$ denote the hidden representation provided to the language-model output layer, with logits $\ell_{i,t} = W v_{i,t}$. The GRPO logit residual is detached when constructing the step-specific output-layer gradient. The reference gradient is similarly detached after aggregation over the reference steps $\mathcal{R}$ included in $\mathcal{L}_{\mathrm{GRPO}}^{\mathrm{ref}}$:

$$\begin{aligned} d_{i,t} &= \mathrm{sg}\left(\frac{\partial \mathcal{L}_{\mathrm{GRPO}}}{\partial \ell_{i,t}} \right), \qquad g_{i,t} = d_{i,t} v_{i,t}^\top, \\ G_{\mathrm{ref}} &= \mathrm{sg}\left(\sum_{(j,s)\in\mathcal{R}} d_{j,s} v_{j,s}^\top \right), \quad \alpha_{i,t} = \langle g_{i,t}, G_{\mathrm{ref}} \rangle_F = d_{i,t}^\top G_{\mathrm{ref}} v_{i,t}, \\ \frac{\partial \alpha_{i,t}}{\partial v_{i,t}} &= G_{\mathrm{ref}}^\top d_{i,t}. \end{aligned} \tag{20}$$

With $d_{i,t}$ and G_{ref} detached, the alignment score provides a first-order gradient through $v_{i,t}$ to earlier temperatures, since $v_{i,t}$ precedes $s_{i,t}$.

Gradient Scaling and Centering. We scale the alignment objective using a detached moving estimate of the root-mean-square alignment score. For the optimization steps $\mathcal{T}$ in each micro-batch, the per-worker values of q_b are averaged before updating m_b:

$$\begin{aligned} q_b &= \sqrt{\frac{1}{|\mathcal{T}|} \sum_{(i,t)\in\mathcal{T}} \mathrm{sg}(\alpha_{i,t})^2}, \qquad m_b = \beta m_{b-1} + (1-\beta) q_b, \\ \kappa_b &= \min\left(\kappa_{\max}, \frac{c_{\mathrm{sc}}}{m_b + \varepsilon_{\mathrm{sc}}} \right), \quad \mathcal{L}_{\mathrm{AS}}^{\mathrm{scaled}} = \frac{\kappa_b}{|\mathcal{T}|} \mathcal{L}_{\mathrm{AS}} = -\frac{\kappa_b}{|\mathcal{T}|} \sum_{(i,t)\in\mathcal{T}} \alpha_{i,t}. \end{aligned} \tag{21}$$

With $\beta = 0.99$, $\kappa_{\max} = 10^6$, $c_{\mathrm{sc}} = 10^{-3}$, and $\varepsilon_{\mathrm{sc}} = 10^{-30}$, the resulting temperature gradients are centered within each rollout and applied to AdaptSoft through a surrogate objective:

$$\begin{aligned} g_{i,t}^\tau &= \frac{\partial \mathcal{L}_{\mathrm{AS}}^{\mathrm{scaled}}}{\partial \tau_{i,t}}, \qquad \overline{g}_i^\tau = \frac{1}{|\mathcal{T}_i|} \sum_{t\in\mathcal{T}_i} g_{i,t}^\tau, \\ \widetilde{g}_{i,t}^\tau &= \mathrm{sg}\left(g_{i,t}^\tau - \overline{g}_i^\tau \right), \quad \widetilde{\mathcal{L}}_{\mathrm{AS}}(\phi) = \sum_i \sum_{t\in\mathcal{T}_i} \widetilde{g}_{i,t}^\tau \tau_{i,t}^{\mathrm{upd}}, \\ \nabla_\phi \widetilde{\mathcal{L}}_{\mathrm{AS}} &= \sum_i \sum_{t\in\mathcal{T}_i} \widetilde{g}_{i,t}^\tau \frac{\partial \tau_{i,t}^{\mathrm{upd}}}{\partial \phi}. \end{aligned} \tag{22}$$

Centering removes the component shared by the temperature-gradient signals within a rollout, retaining their differences across steps. The reference-gradient factors are shared across workers, and

AdaptSoft gradients are averaged before its optimizer step. Only the centered alignment gradient updates ϕ: accumulated GRPO gradients on ϕ are discarded, and the alignment computation does not accumulate gradients on θ.

D Appendix: Benchmark and Evaluation Details

D.1 Benchmarks

We evaluate Soft Spatial Reasoning on OmniSpatial Jia et al. (2025), SpatiaLab Wasi et al. (2026), and MindCube Wang et al. (2026a). OmniSpatial provides the post-training and held-out evaluation splits; the other two benchmarks are used only for zero-shot evaluation.

OmniSpatial Jia et al. (2025). OmniSpatial comprises more than 8.4K question–answer pairs spanning dynamic reasoning, spatial interaction, complex spatial logic, and perspective taking. These four dimensions contain 50 fine-grained task subcategories. We use its official 6,902-sample training split for post-training and its 1,533-sample held-out test split for evaluation.

SpatiaLab Wasi et al. (2026). SpatiaLab comprises 1,400 question–answer pairs drawn from realistic, unconstrained scenes. Its six categories are relative positioning, depth and occlusion, orientation, size and scale, spatial navigation, and 3D geometry; each contains five task types. We evaluate in the multiple-choice setting. No SpatiaLab samples are used for post-training.

MindCube Wang et al. (2026a). MindCube contains 21,154 questions across 3,268 images and examines spatial mental modeling from limited views. Its Rotation, Among, and Around settings test reasoning about spatial relationships as viewpoints change and objects become partially visible. We evaluate its multiple-choice questions without using MindCube samples for post-training.

D.2 Evaluation protocol

For all three benchmarks, a response is correct when its selected option matches the ground-truth answer. We report accuracy by category for OmniSpatial and SpatiaLab and by setting for MindCube. Overall accuracy for Soft Spatial Reasoning is computed over all evaluated questions, equivalently weighting each category or setting by its sample count.

Table 5: Post-training and evaluation configuration for Soft Spatial Reasoning.

Hyperparameter	Value	Notes
Model configuration		
Backbone	Qwen3-VL-8B-Thinking	–
Language-model adaptation	Full fine-tuning	No LoRA
Numerical precision	BF16	FSDP2 mixed precision
Parameter sharding	FSDP2	Parameter and optimizer offload
Data preparation		
Training-image pixel budget	128×28^2	Aspect ratio preserved; dimensions floored to multiples of 28
Evaluation image resolution	Native	No resizing
Maximum prompt length	10,240 tokens	Overlong prompts discarded
Training configuration		
Training GPUs	8	Two nodes $\times$ four H100s
Training steps	200	≈ 2 epochs
Prompt batch size	64	Per GRPO step
Rollouts per prompt	$G = 8$	GRPO group size
Optimizer mini-batch size	32	Two prompt mini-batches per GRPO batch
Micro-batch size	4 sequences/GPU	Dynamic batching disabled
Maximum training response length	2,048 tokens	-
Policy optimization		
Optimizer	AdamW	Policy parameters
Learning rate	1×10^{-6}	–
Learning-rate schedule	Cosine	10% minimum-LR floor
Warmup ratio	0.05	–
KL coefficient	1×10^{-3}	Loss-only; frozen reference policy
Entropy coefficient	0.0	No entropy bonus
Advantage estimation	GRPO	Group-normalized
Soft thinking		
Candidate set	Top-$k = 5$	Per reasoning step
Softness temperature	Step-specific	Predicted by AdaptSoft
Gumbel-noise scale	1.0	Added to candidate log-probabilities
Application span	Reasoning tokens only	Ends at `</think>`; discrete answer
AdaptSoft f_ϕ		
Projection dimension	$d_p = 8$	Fixed random projection
AdaptSoft architecture	Two-layer MLP	Width 256; GELU
Temperature map	$\tau_{i,t} = 0.5 + 0.4 \tanh(u_{i,t})$	$\tau_{i,t} \in (0.1, 0.9)$
AdaptSoft initialization	$a = 0,\ b = 0$	Starts at $\tau = 0.5$
Entropy standardization	$\mu_H = 0.173,\ \sigma_H = 0.224$	Estimated from the initial policy and held fixed
AdaptSoft optimizer	AdamW, 1×10^{-3}	Separate from policy optimizer
Gradient centering	Per rollout	Across soft reasoning steps
Reference subset	25% of each batch	Disjoint and all-gathered
Reward		
Answer-reward weight	1.0	Exact answer match
Format-reward weight	0.2	One `<think>` block; final letter A–D
Evaluation		
Samples per question	8	Mean@8
Decoding temperature	0.6	–
Decoding top-k	5	–
Maximum response length	3,072 tokens	–
Infrastructure		
RL framework	`verl` 0.8.0	GRPO; FSDP2
Rollout engine	SGLang 0.5.12	TP $= 1$; DP $= 8$
GPU memory fraction	0.4	Rollout engine
Orchestration	Ray and Slurm	Multi-node execution