VAST: VISUALLY ANCHORED SOFT THINKING FOR MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) increasingly rely on chain-of-thought reasoning, yet autoregressive decoding commits to a single discrete token at each step, limiting exploration of alternative reasoning directions. Soft thinking relaxes this constraint by propagating probability-weighted mixtures of token embeddings, but we find that naively applying it to MLLMs can substantially degrade accuracy due to unnecessary uncertainty. We identify two requirements for effective multimodal soft reasoning: *when* to activate exploration and *how* to keep it anchored in visual evidence. We introduce VAST (Visually Anchored Soft Thinking), a training-free framework that addresses both through a position-aware entropy schedule and visual contrastive reweighting. VAST selectively activates soft reasoning during the intermediate reasoning phase and steers the resulting representations toward image-dependent candidates. Across Qwen2.5-VL and LLaVA-NeXT models from 7B to 32B parameters and five multimodal benchmarks, VAST achieves the highest average accuracy on every model, with ablations confirming the contribution of both components. Our results show that effective multimodal soft reasoning requires selective exploration that remains anchored in visual evidence.

1 INTRODUCTION

Multimodal large language models (MLLMs) have made rapid progress on tasks requiring joint reasoning over images and text Liu et al. (2024a); Bai et al. (2025), driven in large part by chain-of-thought (CoT) prompting Wei et al. (2022); Zhang et al. (2023); Chen et al. (2024b). However, CoT inherits a fundamental limitation of autoregressive decoding: at every step, the model maintains a distribution over many plausible continuations but commits to a single token, collapsing alternative reasoning directions before their consequences can be explored. We refer to this as a *commitment bottleneck*, where an early suboptimal choice can redirect the entire reasoning chain leading to an incorrect answer.

Human reasoning is less constrained by this discrete process. When reasoning about a visual scene, people can consider multiple interpretations of a scene at once, exploring alternatives before committing to a verbal explanation, and repeatedly return to the image, directing attention toward relevant regions Yarbus (1967). Recent works have introduced similar flexibility to text-only large language models (LLMs), either by training models to reason in continuous space Hao et al. (2024); Xu et al. (2025) or, without additional training, through *soft thinking* Zhang et al. (2026); Zhuang et al. (2026). Instead of feeding back the embedding of a single sampled token, soft thinking propagates a continuous mixture of candidate token embeddings, allowing the model to implicitly explore alternative reasoning directions without committing to a single token at every step.

A straightforward extension of soft thinking to MLLMs, however, does not reliably provide this benefit. While some reasoning steps involve competing alternatives, others are near-deterministic. Applying soft thinking uniformly can therefore introduce noise rather than useful exploration, as demonstrated by the accuracy degradation observed across our experiments (Table 1). Consistent with findings in text-only models Shi et al. (2026); Fu & Luo (2026), this raises the first question: *when should the model think softly?*

Selective activation alone, however, does not address the multimodal nature of the problem. Soft thinking reweights high-probability candidate tokens, but in MLLMs these probabilities may reflect

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

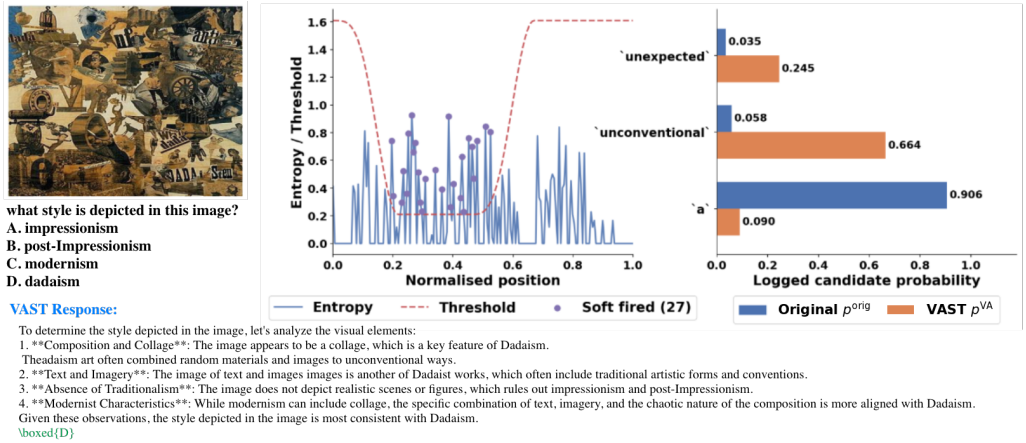


Figure 1: **VAST on an example from MMBench from LLaVA-NeXT.** *Left:* Input image and question. *Middle:* Soft reasoning fires (purple) only when top-K entropy (blue) crosses the valley-shaped threshold (red), so boundary steps remain discrete. *Right:* At one activated step, visual contrast shifts mass from the generic "a" (p^{orig}) to the image-dependent "unconventional" and "unexpected" (p^{VA}). *Bottom:* "Unconventional" surfaces in the reasoning, and VAST correctly answers Dadaism.

linguistic priors rather than visual evidence Leng et al. (2024); Favero et al. (2024). Figure 1 (right) illustrates this issue: when identifying the style of a collage, the model assigns 0.906 of its candidate mass to the generic token "a", while contrasting against a visually degraded input shifts the mass toward the image-dependent descriptors "unconventional" (0.664) and "unexpected" (0.245). This raises the second question: *how can soft reasoning remain anchored in visual evidence?*

These observations motivate a simple design principle: **explore selectively, and anchor that exploration in the visual input**. Guided by this principle, we introduce **VAST** (**V**isually **A**nchored **S**oft **T**hinking), a training-free inference framework for selective and visually anchored soft reasoning in MLLMs.

VAST addresses the first question through *position-aware selective activation*. Multimodal reasoning trajectories exhibit a coarse positional structure: early tokens primarily interpret and organize the problem, intermediate tokens perform the main reasoning, and later tokens consolidate the reasoning into an answer Venhoff et al. (2025); Bogdan et al. (2025). Soft exploration is most useful during the intermediate phase, where the model is more likely to weigh competing reasoning directions, and less useful near the boundaries, where it can blur the problem representation or destabilize the final answer. VAST captures this structure with a valley-shaped entropy schedule: the activation threshold is higher near the beginning and end of the trajectory and lower in the intermediate region (Figure 1, middle). Rather than learning a reasoning policy, this lightweight schedule provides a positional prior that concentrates exploration where it is most likely to help.

VAST addresses the second question through *visual contrastive reweighting*. At each position selected by the gate, VAST compares the model's prediction under the original image and a spatially corrupted counterpart. The corruption disrupts coherent visual structure while preserving low-level appearance information such as color and texture. The difference between the two predictions provides a signal for distinguishing candidates supported by the visual content from those driven primarily by linguistic context. VAST uses this signal to reweight the candidate distribution before constructing the soft embedding, steering continuous exploration toward visually supported reasoning directions.

VAST thus combines two complementary mechanisms:

$$\underbrace{\text{position-aware selective activation}}_{\text{when to think softly}} + \underbrace{\text{visual contrastive reweighting}}_{\text{how to think softly}}. \tag{1}$$

This separation also keeps VAST efficient: the additional forward pass for visual contrast is needed only at gated steps, roughly 6–20% of reasoning steps on average, so most decoding remains stan-

dard discrete CoT. Across Qwen2.5-VL and LLaVA-NeXT models (7B–32B) and five multimodal benchmarks, VAST achieves the highest average accuracy on every model, improving over greedy CoT by up to 1.46 points and over standard soft thinking by up to 5.68 points, whereas naive soft thinking often underperforms standard CoT. Ablations confirm that both the position-aware schedule and visual contrastive reweighting contribute to these gains.

To summarize, our contributions are as follows: **(i) A multimodal formulation of soft thinking.** We show that uniformly applying soft thinking can degrade MLLM accuracy and identify two requirements for effective multimodal soft reasoning: exploration should be activated selectively over the generation trajectory, and the resulting continuous representations should remain anchored in visual evidence. **(ii) A training-free inference framework.** We introduce VAST, which combines a position-aware uncertainty schedule with visual contrastive reweighting to enable sparse, visually anchored soft reasoning without modifying model parameters or requiring additional training. **(iii) Evaluation across models, scales, and tasks.** We evaluate VAST across diverse MLLMs and multimodal benchmarks, isolate the contributions of selective activation and visual anchoring through controlled ablations, and analyze how visual contrast reshapes the soft candidate distribution. Taken together, our results suggest that effective soft reasoning in multimodal models requires more than preserving uncertainty: the model must explore at the right moments, and that exploration must remain connected to what it sees. Throughout the text, we use *visual contrastive reweighting* and *visually anchored logit adjustment* interchangeably.

2 Related Work

Soft and latent reasoning. To move beyond discrete CoT (Wei et al., 2022), training-based methods reason over hidden states (Hao et al., 2024) or learned soft thoughts (Xu et al., 2025), while training-free Soft Thinking (Zhang et al., 2026) and Mixture of Inputs (Zhuang et al., 2026) feed back probability-weighted token mixtures. Such mixtures tend to collapse toward the top candidate (Wu et al., 2026), and SwiReasoning (Shi et al., 2026) and SeLaR (Fu & Luo, 2026) activate them selectively based on token-level uncertainty. These methods are text-only and use local uncertainty to determine *when* to maintain alternatives. VAST extends this paradigm to MLLMs by introducing a position-aware activation schedule and visually anchored soft representation.

Latent reasoning in MLLMs. Latent reasoning has been explored in MLLMs, with several approaches using additional training to encourage reasoning through latent visual representations (Wang et al., 2026; Yang et al., 2026; Zhang et al., 2025; Li et al., 2026). Training-free alternatives switch between discrete and soft reasoning: LEAD (Xu et al., 2026) gates on entropy and injects a fixed, image-agnostic visual anchor, while AGS (Kang et al., 2026) gates on vision-to-text attention, noting that entropy-gated soft reasoning can underperform CoT in MLLMs. Both are developed largely around reasoning models with long thinking traces. VAST instead targets instruction-tuned MLLMs, gates with a positional prior, and uses image-specific visual contrast to reweight the soft mixture itself.

Contrastive decoding. Contrastive decoding (Li et al., 2023a) improves generation by contrasting the predictions of an expert model with those of a weaker model. In MLLMs, VCD (Leng et al., 2024) adapts this principle by contrasting predictions from the original image and a visually distorted counterpart. Subsequent approaches introduce alternative negative inputs, such as text-only contexts (Favero et al., 2024) and perturbed instructions (Wang et al., 2024), primarily to mitigate multimodal hallucination. These methods use contrastive predictions to modify discrete token selection. VAST instead applies visual contrast selectively at soft-thinking steps, using the resulting distribution to construct a continuous representation rather than directly selecting the next token.

3 Method

We introduce **VAST**, a training-free framework for selectively activating soft reasoning in vision-language models (VLMs). As illustrated in Figure 2, VAST operates in two stages. First, a position-aware entropy gate identifies uncertain reasoning steps at which retaining multiple candidate tokens may be useful. At each selected step, VAST then contrasts the model's predictions under the original image and a perturbed image, reweights the candidate distribution toward visually supported

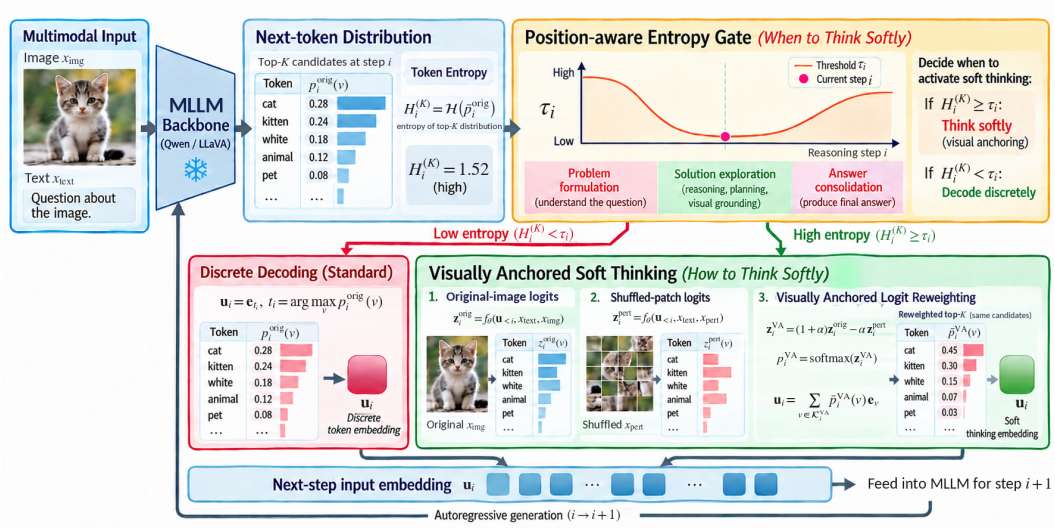


Figure 2: **Overview of VAST.** At each step, a valley-shaped, position-aware entropy threshold decides the decoding mode: low-entropy steps decode greedily, while high-entropy steps trigger visually anchored soft thinking. For these steps, VAST contrasts logits under the original and a patch-shuffled image, reweights the top-K candidates, and feeds their mixture back as a soft embedding.

directions, and constructs a soft embedding for the next decoding step. The entropy gate is the only hard activation decision; visual contrast determines how the selected soft embedding is formed.

3.1 Preliminaries

Discrete chain-of-thought reasoning. Let x_{text} be the textual input, comprising the CoT instruction and the question, x_{img} the input image, and $\mathcal{V}$ the vocabulary of size $|\mathcal{V}|$. Let $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d}$ be the token-embedding matrix and $\mathbf{e}_v = \mathbf{E}[v] \in \mathbb{R}^d$ the embedding of token v. An MLLM first generates an reasoning trace $t_{1:m} = (t_1, \ldots, t_m)$ and then a final answer $y_{1:n} = (y_1, \ldots, y_n)$, where $m \leq M$ for a maximum reasoning horizon M. At reasoning step i, the model produces logits and a next-token distribution conditioned on the input text and the original input image:

$$\mathbf{z}_i^{\text{orig}} = f_\theta(\mathbf{u}_{<i}, x_{\text{text}}, x_{\text{img}}), \qquad \mathbf{p}_i^{\text{orig}} = \text{softmax}\left(\mathbf{z}_i^{\text{orig}}\right), \tag{2}$$

where $\mathbf{u}_{<i} = (\mathbf{u}_1, \ldots, \mathbf{u}_{i-1})$ denotes the embeddings previously fed to the decoder. Standard CoT selects a single token with a decoding rule $\mathcal{D}$ (e.g. greedy decoding or sampling) and feeds its embedding to the next step: $t_i = \mathcal{D}\left(\mathbf{p}_i^{orig}\right), \mathbf{u}_i = \mathbf{e}_{t_i}$ Thus, each reasoning step commits to one token even when several continuations receive substantial probability.

Soft thinking in text-only models. Soft thinking Zhang et al. (2026) relaxes this commitment by replacing a sampled intermediate token with a probability-weighted mixture of token embeddings. Given a next-token distribution $\mathbf{p}_i$, the corresponding soft embedding is

$$\widetilde{\mathbf{e}}_i = \sum_{v \in \mathcal{V}} p_i^{\text{orig}}(v)\, \mathbf{e}_v, \tag{3}$$

typically restricted to the top candidates with renormalized probabilities. VAST adopts this construction but applies it only at selected steps (3.2) and derives the weights from a visually contrasted distribution (3.3).

3.2 Position-Aware Entropy Gating

VAST first uses predictive uncertainty to identify candidate steps for soft reasoning. Let K be the candidate width and $\mathcal{K}_i^{\text{orig}} \subset \mathcal{V}$ the K most probable tokens under $\mathbf{p}_i^{\text{orig}}$. We renormalize

their probability mass as $\overline{p}_i^{\text{orig}}(v) = \frac{p_i^{\text{orig}}(v)}{\sum_{w \in \mathcal{K}_i^{\text{orig}}} p_i^{\text{orig}}(w)}, v \in \mathcal{K}_i^{\text{orig}}$, and compute the entropy of the resulting top-K distribution: $H_i^{(K)} = -\sum_{v \in \mathcal{K}_i^{\text{orig}}} \overline{p}_i^{\text{orig}}(v) \log \overline{p}_i^{\text{orig}}(v)$. Restricting the entropy to the top-K candidates has two advantages: it measures competition among exactly the tokens that would form the soft mixture, and it is bounded, $H_i^{(K)} \in [0, \log K]$, which makes thresholds comparable across steps. Low entropy indicates that one candidate dominates, whereas entropy near $\log K$ indicates that the candidates are nearly tied.

Valley-shaped threshold schedule. Entropy alone is an incomplete signal for when to think softly: applying soft thinking at every step degrades accuracy in MLLMs (Table 1), and even entropy-gated soft reasoning can underperform standard CoT (Kang et al., 2026). The value of exploration also depends on where a step lies in the trajectory, since reasoning traces tend to progress from problem setup, through intermediate computation, to answer consolidation (Venhoff et al., 2025; Bogdan et al., 2025). Early decisions strongly influence the rest of the trajectory (Xu et al., 2026), so blurring them into mixtures risks propagating an ambiguous problem setup, while exploration near the end can destabilize an otherwise settled answer. Intermediate steps, where the model weighs competing directions, are where retaining alternatives is most likely to help. VAST therefore uses a valley-shaped threshold that is high at the boundaries and low in the intermediate phase.

Let $s_i = i - 1$ be the zero-indexed reasoning step, S the expected reasoning length used by the schedule, and $\nu \in [0, 1]$ the fraction of S occupied by the flat central valley. Each transition has length $D = \lfloor \frac{1-\nu}{2} S \rfloor$, the final transition begins at $R = S - D$, and $D_+ = \max\{D, 1\}$. Let τ_{low} be the threshold in the central region and τ_{high} the threshold at the boundaries. Using the cosine ramp $c(r) = \frac{1-\cos(\pi r)}{2}$, the threshold at step i is

$$\tau_i = \begin{cases} \tau_{\text{high}} + (\tau_{\text{low}} - \tau_{\text{high}})\, c\left(\dfrac{s_i}{D_+}\right)^2, & 0 \le s_i \le D, \\ \tau_{\text{low}}, & D < s_i < R, \\ \tau_{\text{low}} + (\tau_{\text{high}} - \tau_{\text{low}})\, c\left(\dfrac{s_i - R}{D_+}\right)^2, & R \le s_i \le S, \\ \tau_{\text{high}}, & s_i > S. \end{cases} \tag{4}$$

The threshold descends smoothly from τ_{high}, stays at τ_{low} through the central region, and rises again toward the end of the trajectory (Figure 2, top) Since $H_i^{(K)} \le \log K$, setting $\tau_{\text{high}} = \log K$ restricts soft reasoning at the boundaries to steps where the top-K candidates are almost exactly tied.

Soft reasoning is activated when the top-K entropy meets or exceeds the position-dependent threshold: $g_i = \mathbb{I}\left[H_i^{(K)} \ge \tau_i\right], g_i \in \{0, 1\}$. When $g_i = 0$, VAST applies the discrete update. When $g_i = 1$, it constructs and applies visually anchored logit adjustments as described next.

3.3 Visually Anchored Logit Adjustment

Entropy identifies competition among candidate tokens but not whether that competition is informed by the image. In MLLMs, candidate probabilities can be dominated by linguistic priors (Leng et al., 2024; Favero et al., 2024), so a soft embedding built directly from $\mathbf{p}_i^{\text{orig}}$ may preserve uncertainty while drifting away from the visual evidence. At each step selected by the gate, VAST therefore contrasts the model's prediction on the original image with its prediction on a perturbed counterpart x_{pert}, keeping the text and reasoning history unchanged (Figure 2, bottom):

$$\mathbf{z}_i^{\text{pert}} = f_\theta(\mathbf{u}_{<i}, x_{\text{text}}, x_{\text{pert}})\,. \tag{5}$$

By default, x_{pert} is obtained by randomly shuffling the patches of x_{img}. Shuffling destroys spatial structure and object-level content while preserving low-level statistics such as color and texture, so the contrast isolates candidates that depend on what the image depicts rather than on the mere presence of an image. VAST then forms visually adjusted logits:

$$\mathbf{z}_i^{\text{VA}} = \mathbf{z}_i^{\text{orig}} + \alpha \left(\mathbf{z}_i^{\text{orig}} - \mathbf{z}_i^{\text{pert}}\right), \tag{6}$$

	Qwen2.5-VL-7B				LLaVA-NeXT-8B			
Benchmark	**CoT**	**ST**	**VAST (Ours)**	ρ_{soft} (%)	**CoT**	**ST**	**VAST (Ours)**	ρ_{soft} (%)
M3CoT	65.73	60.53	**66.73**	6.96	50.13	49.46	**50.26**	21.91
MMMU	44.67	39.79	**46.53**	6.73	36.20	35.46	**37.93**	21.24
MMBench	82.53	73.80	**83.27**	9.61	**76.53**	75.20	75.53	15.37
MMStar	57.08	54.41	**58.60**	10.98	40.73	40.26	**41.33**	24.42
POPE	83.73	84.13	**85.90**	3.18	**83.67**	83.20	**83.67**	18.02
Average	66.75	62.53	**68.21**	7.49	57.45	56.72	**57.74**	20.19

Table 1: **Benchmark results across MLLMs.** Exact-match accuracy (%) for Greedy CoT, Standard Soft Thinking (ST), and VAST. **Bold** indicates the best result among the three methods for each benchmark and model. The **Average** row reports the mean performance across the five benchmarks.

where $\alpha \geq 0$ controls the strength of the visual contrast. The adjustment to each token v is proportional to $z_{i,v}^{\text{orig}} - z_{i,v}^{\text{pert}}$: candidates better supported by the original image than by its perturbation are amplified, while candidates driven mainly by the text context receive little relative boost. This encourages image-dependent candidates.

3.4 Visually Anchored Soft Embedding

The adjusted logits distribution is $\mathbf{p}_i^{\text{VA}} = \text{softmax}(\mathbf{z}_i^{\text{VA}})$. Let $\mathcal{K}_i^{\text{VA}} \subset \mathcal{V}$ contain the K most probable tokens under $\mathbf{p}_i^{\text{VA}}$. We renormalize their probability mass as $\widehat{p}_i^{\text{VA}}(v) = \frac{\mathbf{p}_i^{\text{VA}}(v)}{\sum_{w \in \mathcal{K}_i^{\text{VA}}} \mathbf{p}_i^{\text{VA}}(w)}, v \in \mathcal{K}_i^{\text{VA}}$. The visually anchored soft embedding is

$$\widetilde{\mathbf{e}}_i^{\text{VA}} = \sum_{v \in \mathcal{K}_i^{\text{VA}}} \widehat{p}_i^{\text{VA}}(v) \mathbf{e}_v. \tag{7}$$

Like standard soft thinking (Eq. (3)), this representation retains multiple candidate directions, but weights them by the visually adjusted distribution $\widehat{\mathbf{p}}^{\text{VA}}$ rather than by $\mathbf{p}_i^{\text{orig}}$ alone. Combining the discrete and soft branches, the next decoder input is

$$\mathbf{u}_i = \begin{cases} \mathbf{e}_{t_i}, & H_i^{(K)} < \tau_i, \\ \widetilde{\mathbf{e}}_i^{\text{VA}}, & H_i^{(K)} \geq \tau_i. \end{cases} \tag{8}$$

where $t_i = \mathcal{D}(\mathbf{p}_i^{\text{orig}})$ as in section 3.1. VAST therefore retains discrete CoT as its default behavior and introduces a visually anchored soft embeddings only at steps selected by the entropy gate. Once reasoning terminates, the final answer $y_{1:n}$ is generated through standard discrete decoding.

3.5 Overall Inference Procedure

Algorithm A.3 summarizes VAST. We maintain an embedding history $\mathcal{U}$ because it may contain both discrete token embeddings and continuous soft embeddings.

Inference cost. Let m be the realized number of reasoning steps and $\rho_{\text{soft}} = \frac{1}{m} \sum_{i=1}^{m} g_i$ the soft-activation rate. The original-image forward pass runs at every reasoning step, whereas the perturbed-image pass runs only when $g_i = 1$. VAST therefore performs $m + \sum_{i=1}^{m} g_i$ forward passes during intermediate reasoning. Ignoring implementation-dependent caching effects, its average forward-pass factor relative to standard CoT is $1 + \rho_{\text{soft}}$. The additional multimodal computation is thus restricted only to the steps selected by the entropy gate.

4 Experiments

We evaluate the utility of VAST across model families, scales, and benchmarks to answer three questions: (i) Does VAST improve multimodal reasoning over discrete CoT and Standard Soft Thinking? (ii) Do these gains transfer across model families and scales? and (iii) How much do position-aware selection and visual anchoring contribute individually? Additionally, we also analyze how a training-free method like VAST compare with training-based latent reasoning.

Benchmark	Qwen2.5-VL-32B			
	CoT	ST	VAST	ρ_{soft} (%)
M3CoT	72.00	**72.80**	**72.80**	5.12
MMBench	85.30	84.70	**86.20**	7.10
MMStar	65.51	65.20	**67.13**	6.40
Average	74.27	74.23	**75.38**	6.21

Table 2: **Results on Qwen2.5-VL-32B.** Exact-match accuracy (%) for Greedy CoT, Standard Soft Thinking (ST), and VAST. **Bold** indicates the best result among the three methods for each benchmark. The **Average** row reports the mean across the three benchmarks.

Schedule	M3CoT		MMBench	
	Acc.	ρ_{soft} (%)	Acc.	ρ_{soft} (%)
Fixed / Flat	61.9	22.5	79.8	29.6
Cosine	62.1	9.2	80.8	9.3
Cosine^2	62.3	11.9	79.9	11.7
Exponential	61.4	14.7	80.2	15.7
Linear	61.4	9.1	80.8	8.6
Position-Aware Valley (Ours)	**63.6**	10.9	**83.9**	12.4

Table 3: **Effect of the threshold schedule.** All configurations use the same visually anchored soft representation and differ only in the entropy schedule.

4.1 Experimental Setup

Benchmarks. We conduct a comprehensive evaluation across a diverse set of benchmarks, specifically, M3CoT Chen et al. (2024b), MMMU Yue et al. (2024), MMBench Liu et al. (2024b), MMStar Chen et al. (2024a), and POPE Li et al. (2023b) spanning multimodal reasoning, visual understanding, and hallucination detection. Unless otherwise stated, Main results use a random subset of 1,500 samples per benchmark.

Models. We use Qwen2.5-VL-7B and Qwen2.5-VL-32B (Bai et al., 2025) and LLaVA-NeXT-8B (Liu et al., 2024a), covering two model families and two scales.

Baselines. We compare VAST against *Greedy CoT*, which decodes the full reasoning trace discretely, and *Standard Soft Thinking (ST)* Zhang et al. (2026), which applies soft reasoning at every step. For component analysis, we also evaluate entropy-gated soft thinking with a fixed threshold (ST-Fixed) and with the valley-shaped threshold but without visual contrast (ST-U).

Evaluation. Unless otherwise specified, we report exact-match accuracy (%). For each benchmark, predictions are compared against the corresponding ground-truth answers, with a prediction counted as correct only when it exactly matches the expected answer.

Hyperparameters. Unless otherwise specified, we use $\nu = 0.35$, $\tau_{\text{low}} = 0.2$, $\alpha \in \{0.1, 0.3\}$, $K = 3$, batch size 1, and shuffled-patch sizes in $\{1, 2, 3\}$. The upper threshold is fixed at $\tau_{\text{high}} = 1.10$.

4.2 Main Results

Comparison with CoT and soft thinking. Table 1 reports results on Qwen2.5-VL-7B and LLaVA-NeXT-8B. Standard soft thinking often falls below greedy CoT, sometimes substantially, whereas VAST improves over both and achieves the highest average accuracy on both models. Replacing discrete tokens with continuous mixtures is therefore not beneficial by itself; the gains come from *how* VAST introduces soft reasoning.

Generalization across families and scale. The trend holds across model families and scales. Gains are largest on Qwen2.5-VL-7B and smaller but consistent on LLaVA-NeXT-8B, where VAST is best or tied-best on all benchmarks except MMBench. On Qwen2.5-VL-32B (Table 2), VAST remains the best method on every benchmark, while ST continues to behave inconsistently relative to CoT. Increased capacity alone therefore does not make naive soft reasoning reliable, and VAST's benefits are not tied to a particular model or scale.

Soft-activation rate. VAST activates soft reasoning at 7.49% of reasoning steps on Qwen2.5-VL-7B, 6.21% on Qwen2.5-VL-32B, and 20.19% on LLaVA-NeXT-8B, compared with every step for ST. Larger models are more confident and trigger the gate less often compared to their smaller counterpart. Notably, LLaVA-NeXT-8B activates soft reasoning most often yet gains least, suggesting that *when* soft reasoning is applied matters more than how often.

Benchmark	Qwen2.5-VL-7B ST	ST-Fixed	ST-U	VAST	LLaVA-NeXT-8B ST	ST-Fixed	ST-U	VAST
M3CoT	60.53	63.65	64.00	**66.73**	49.46	49.33	49.80	**50.26**
MMMU	39.79	43.67	**46.60**	46.53	35.46	36.00	35.86	**37.93**
MMBench	73.80	73.73	81.53	**83.27**	75.20	**77.20**	76.33	75.53
MMStar	54.41	55.10	57.60	**58.60**	40.26	40.86	37.93	**41.33**
Average	57.13	59.04	62.43	**63.78**	50.10	50.85	49.98	**51.26**

Table 4: **Component ablation.** Each column adds one component to the previous one. ST: soft thinking at every step. ST-Fixed: adds entropy gating with a fixed threshold. ST-U: replaces the fixed threshold with the valley-shaped schedule (VAST without visual contrast). VAST: adds visual contrastive reweighting. Accuracy (%); **bold** marks the best result per benchmark and model.

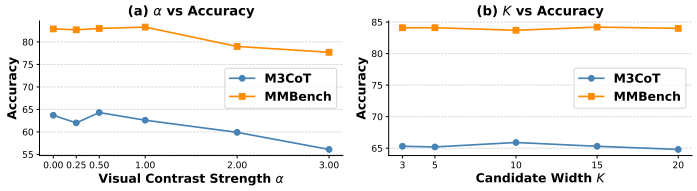


Figure 3: **Sensitivity analysis on Qwen2.5-VL-7B.** (a) Accuracy as a function of the contrast strength α; $\alpha = 0$ removes visual contrast. (b) Accuracy as a function of the candidate width K.

Perturbation	MMBench
Blank image	81.83
Gaussian blur	83.00
Shuffled patches	**84.00**

Table 5: **Effect of visual perturbation.** Results use Qwen2.5-VL-7B with $\alpha = 1.0$; all other hyperparameters are fixed.

4.3 Ablation Studies

To systematically assess the contribution of each design choice and the sensitivity of VAST to its key hyperparameters, we conduct a focused set of ablation studies. All ablations are performed on **Qwen2.5-VL-7B** using **M3CoT** and **MMBench**, varying one component at a time with all other settings fixed. These two benchmarks provide a compact test bed for analyzing the behavior of VAST while avoiding an expensive sweep across every model and dataset.

Contribution of each component. Since VAST's gating stage combines entropy-based selection with a valley-shaped threshold, we ablate its components incrementally (Table 4), starting from standard soft thinking (ST) and adding, in turn, entropy gating with a fixed threshold (ST-Fixed), the valley-shaped threshold (ST-U), and visual contrastive reweighting (VAST). On Qwen2.5-VL-7B, every component helps, with the valley-shaped threshold yielding the largest gain. On LLaVA-NeXT-8B, the valley-shaped threshold alone does not help, but adding visual contrast reverses this and produces the best average accuracy. The two stages are therefore complementary: position-aware gating determines where exploration occurs, and visual anchoring determines whether that exploration is productive.

Threshold schedule. To test whether concentrating exploration in the intermediate phase is beneficial, Table 3 compares the valley-shaped schedule with a flat threshold and with cosine, squared-cosine, exponential, and linear schedules (Appendix A.6), keeping all other settings fixed. This is a targeted check of the valley schedule rather than a search for an optimal scheduling function. The valley schedule achieves the best accuracy on both benchmarks while activating soft reasoning about half as often as the flat threshold. Since the alternative schedules activate at similar rates but reach lower accuracy, the gain comes from *where* exploration is placed rather than from activating it less.

Visual perturbation. Table 5 compares three contrast inputs that remove different aspects of the image: a blank image removes all visual information, Gaussian blur removes fine detail but keeps coarse layout, and patch shuffling destroys spatial structure but keeps low-level statistics such as color and texture. Patch shuffling performs best, followed by blur and the blank image. This ordering is consistent with our motivation (Section 3.3): contrasting against a blank image partly measures the mere presence of an image, whereas a perturbation that preserves low-level statistics isolates candidates that depend on what the image actually depicts.

Sensitivity to hyperparameters. We examine VAST's sensitivity to its two main hyperparameters, the contrast strength α and the candidate width K, varying each individually. The coefficient α

(Eq. (6)) controls the strength of visual contrast, with $\alpha = 0$ removing it entirely. Figure 3(a) shows that moderate values ($\alpha \in [0.5, 1]$) work best, while larger values degrade accuracy sharply suggesting that visual contrast is most effective as a controlled adjustment to the original distribution. The candidate width K determines how many tokens form each soft representation and also changes the entropy range used by the gate. VAST is robust to this choice: accuracy remains nearly flat from $K = 3$ to 20 (Figure 3(b)), even as the activation rate drops by more than half, so VAST does not require careful tuning of K.

Comparison with training-based latent reasoning. We also compare VAST with supervised fine-tuning and Latent Visual Reasoning (LVR) (Li et al., 2026), a training-based latent reasoning method (Appendix A.7). Despite requiring no training, VAST performs comparably to LVR on most tasks and outperforms supervised fine-tuning on five of seven, suggesting that much of the benefit of latent visual reasoning can be obtained at inference time alone.

4.4 Qualitative Analysis

How VAST reshapes soft reasoning: Figure 1 traces VAST on a single example. Soft reasoning fires in the intermediate phase of the trajectory, while comparably high-entropy steps near the end remain discrete (middle panel).At one activated step, visual contrast shifts mass from the generic continuation "a" to descriptors that characterize the collage's style, such as "unconventional" (right panel). This descriptor reappears in the reasoning, where the model describes Dadaist works as combining materials in unconventional ways, and the answer is grounded in visual features of the image, such as its collage composition and mixed text and imagery, leading VAST to correctly identify the style as Dadaism.

Comparison with CoT and soft thinking Figure 4 compares the three decoding strategies on a POPE example with LLaVA-NeXT-8B. Greedy CoT and standard soft thinking both answer incorrectly: CoT restricts the claim to the area outside the bus, and soft thinking mentions the driver but then contradicts itself. VAST keeps its reasoning consistent with the visual evidence and answers correctly. Additional examples are provided in Appendix.

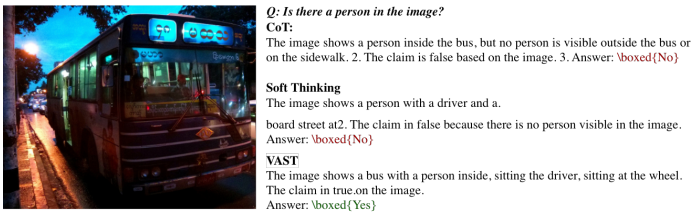


Figure 4: **Qualitative comparison on POPE with LLaVA-NeXT-8B.** Only VAST answers correctly. At soft steps, the displayed text shows the highest-probability token under the visually adjusted distribution, which accounts for occasional disfluencies.

5 Limitation and Future Work

The limitations and future work are described in the appendix A.1.

6 Conclusion

In this work, we introduced **VAST**, a training-free soft-thinking framework for MLLMs. VAST combines two complementary components: a position-aware entropy schedule, which determines *when* soft reasoning is activated, and visual contrastive reweighting, which determines *how* the soft representation is formed. Across model families and scales, VAST achieves the highest average accuracy, improving over both greedy CoT and standard soft thinking, and remains competitive with training-based latent visual reasoning without any additional training. Ablations confirm that both components contribute, that concentrating exploration in the intermediate phase outperforms alternative schedules, and that VAST is robust to the choice of perturbation and candidate width. Because soft reasoning is activated at only a small fraction of steps, these gains come with limited additional computation. Overall, our results suggest that effective multimodal soft reasoning depends not only on introducing continuous representations, but on controlling both *when* and *how* they are used.

AI USE STATEMENT

In this work, we used generative AI tools for designing or providing feedback on research methodology or experiments, implementing methods, and interpreting results. We have not used generative AI tools to support qualitative and thematic data analysis. Generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, assisting with translation, and cleaning and reformatting datasets do not apply to this work. Additionally, we used generative AI tools for creating or modifying scientific figures or images, creating or editing software code, drafting parts of a research paper, brainstorming, sourcing/searching for information, identifying relevant literature. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artefacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All the details of the models and hyperparameters are given in section 4.1 and appendix A.2. The code will be released upon acceptance.

REFERENCES

Shuai Bai, Ke qin Chen, Xue-Jing Liu, Jia-Lin 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*, abs/2502.13923, 2025. URL `https://api.semanticscholar.org/CorpusID:276449796`.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which llm reasoning steps matter? *ArXiv*, abs/2506.19143, 2025. URL `https://api.semanticscholar.org/CorpusID:280000432`.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024a.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8199–8221, 2024b.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312. IEEE, 2024.

Renyu Fu and Guibo Luo. Selar: Selective latent reasoning in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7073–7087, 2026.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

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.

Haoqian Kang, Liupeng Li, Kuofeng Gao, Jinpeng Wang, Zhenyu Lu, Bin Chen, Ke Chen, and Yaowei Wang. Balancing efficiency and efficacy: Training-free attention-guided switching between explicit and latent thoughts for mllms. *arXiv preprint arXiv:2608.03450*, 2026.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

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, 2026.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 12286–12312, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

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.

Constantin Venhoff, Iv'an Arcuschin, Philip H. S. Torr, Arthur Conmy, and Neel Nanda. Understanding reasoning in thinking language models via steering vectors. *ArXiv*, abs/2506.18167, 2025. URL https://api.semanticscholar.org/CorpusID:278340657.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond image and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12030–12040, 2026.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the association for computational linguistics: ACL 2024*, pp. 15840–15853, 2024.

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.

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.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. Softcot: Soft chain-of-thought for efficient reasoning with llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23336–23351, 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11064–11075, 2026.

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.

Alfred L. Yarbus. Eye movements and vision. In *Springer US*, 1967. URL https://api.semanticscholar.org/CorpusID:40057049.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Huanyu Zhang, Wenshan Wu, Chengzu Li, Ning Shang, Yan Xia, Yangyu Huang, Yifan Zhang, Li Dong, Zhang Zhang, Liang Wang, et al. Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in mllms. *arXiv preprint arXiv:2510.24514*, 2025.

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

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alexander J. Smola. Multimodal chain-of-thought reasoning in language models. *Trans. Mach. Learn. Res.*, 2024, 2023. URL https://api.semanticscholar.org/CorpusID:256504063.

Yufan Zhuang, Liyuan Liu, Chandan Singh, Jingbo Shang, and Jianfeng Gao. Mixture of inputs: Text generation beyond discrete token sampling. *Advances in Neural Information Processing Systems*, 38:26363–26387, 2026.

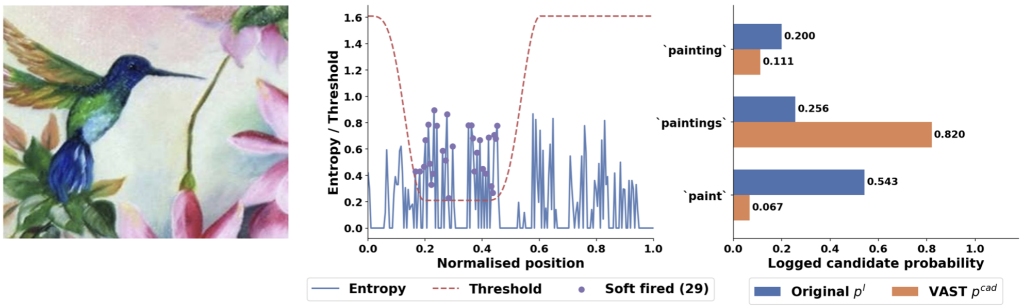
A Appendix

A.1 Limitations and Future work

Although VAST enables continuous latent reasoning, its intermediate representations are still constructed from vocabulary-token embeddings and therefore remain bounded by the language representation space. An important direction for future work is to develop a training-free extension of VAST that performs latent reasoning jointly in both visual and vocabulary representation spaces, allowing intermediate reasoning to branch not only over alternative linguistic tokens but also over alternative visual latent states.

A.2 Implementation Details

We use publicly available Hugging Face model weights without any fine-tuning. Our approach requires no architectural modifications and can be directly integrated into the standard decoding pipeline of MLLMs that provide a Hugging Face interface. All experiments are conducted on one NVIDIA RTX Ada GPU. For all the main paper results, we have used 1500 samples. And all the ablations are don on 1000 samples.



Identify the art style of this image.\nA. oil paint\nB. vector art\nC. Baroque\nD. watercolor

*"To determine the art style of the image, let's analyze the characteristics:\n\n1. **Oil paint**: Oil paint often has a rich, textured appearance with visible brush strokesrokes. The the image does show a vibrant depth and it does't appear to have the thick, layered texture typical of oil paintings.\n\n2. **Vector art**: Vector art is digital and typically lacks clean, clean lines with no visible. The image does not have the crisp precision and in vector art.\n\n3. **Baroque**: Baroque art is known for its dramatic use of light and shadow, often with exaggerated forms and movement. This image does not exhibit those dramatic Baroque characteristics.\n\n4. **Watercolor**: Watercolor paintings are characterized by their soft, translucent quality and the blending of colors. The image shows a soft, blended effect with smooth transitions between colors, which is typical of watercolor.\n\nGiven the analysis, the art style of the image is most consistent with watercolor.\n\n\boxed{D}*

Figure 5: **VAST on an example from MMBench from LLaVA-NeXT.** *Left:* Input image and question. *Middle:* Soft reasoning fires (purple) only when top-K entropy (blue) crosses the valley-shaped threshold (red), so boundary steps remain discrete.

A.3 VAST Algorithm

At each step, VAST compares the top-K entropy with the position-aware threshold to choose between standard discrete decoding and visually anchored soft reasoning. When activated, it uses a perturbed-image forward pass to reweight the logits and feeds the resulting soft embedding into the next step; otherwise, decoding proceeds normally. **?** contains the step-by-step algorithm.

A.4 Appendix: Sensitivity to visual reweighting strength

We examine how sensitive VAST is to the visual contrast strength α in $\mathbf{z}_i^{VA} = \mathbf{z}_i^I + \alpha\left(\mathbf{z}_i^I - \mathbf{z}_i^{perturbed}\right)$. The coefficient α determines how strongly the contrastive visual signal influences the reweighted distribution at selected soft-thinking steps. Setting $\alpha = 0$ removes visual reweighting entirely and therefore serves as a direct control for isolating the contribution of visual

Algorithm 1 VAST: Visually Anchored Soft Thinking

Require: MLLM f_θ; textual input x_{text}; image x_{img}; perturbed image x_{pert}
Require: Maximum reasoning horizon M; schedule length S; valley fraction ν; lower threshold τ_{low}; candidate width K; contrast strength α
Ensure: Final answer $y_{1:n}$

Initialize reasoning history $\mathcal{U} \leftarrow [\,]$
for $i = 1, \ldots, M$ **do**
  Stage 1: Position-aware gating
  $\tau_i \leftarrow \textsc{ValleyThreshold}(i, S, \nu, \tau_{\text{low}}, \log K)$
  $\mathbf{z}_i^{\text{orig}} \leftarrow f_\theta(\mathcal{U}, x_{\text{text}}, x_{\text{img}})$
  $\mathbf{p}_i^{\text{orig}} \leftarrow \text{softmax}(\mathbf{z}_i^{\text{orig}})$
  $\mathcal{K}_i^{\text{orig}} \leftarrow \text{TopK}(\mathbf{p}_i^{\text{orig}}, K)$
  $\overline{\mathbf{p}}_i^{\text{orig}} \leftarrow \textsc{Renormalize}(\mathbf{p}_i^{\text{orig}}, \mathcal{K}_i^{\text{orig}})$
  $H_i^{(K)} \leftarrow -\sum_{v \in \mathcal{K}_i^{\text{orig}}} \overline{p}_i^{\text{orig}}(v) \log \overline{p}_i^{\text{orig}}(v)$
  if $H_i^{(K)} < \tau_i$ **then** ▷ Confident: discrete reasoning
    $t_i \leftarrow \mathcal{D}(\mathbf{p}_i^{\text{orig}})$
    $\mathbf{u}_i \leftarrow \mathbf{e}_{t_i}$
  else ▷ Uncertain: visually anchored soft reasoning
    Stage 2: Visual reweighting
    $\mathbf{z}_i^{\text{pert}} \leftarrow f_\theta(\mathcal{U}, x_{\text{text}}, x_{\text{pert}})$
    $\mathbf{z}_i^{\text{VA}} \leftarrow (1+\alpha)\mathbf{z}_i^{\text{orig}} - \alpha \mathbf{z}_i^{\text{pert}}$
    $\mathbf{p}_i^{\text{VA}} \leftarrow \text{softmax}(\mathbf{z}_i^{\text{VA}})$
    $\mathcal{K}_i^{\text{VA}} \leftarrow \text{TopK}(\mathbf{p}_i^{\text{VA}}, K)$
    $\widehat{\mathbf{p}}_i^{\text{VA}} \leftarrow \textsc{Renormalize}(\mathbf{p}_i^{\text{VA}}, \mathcal{K}_i^{\text{VA}})$
    $\widetilde{\mathbf{e}}_i^{\text{VA}} \leftarrow \sum_{v \in \mathcal{K}_i^{\text{VA}}} \widehat{p}_i^{\text{VA}}(v) \mathbf{e}_v$
    $\mathbf{u}_i \leftarrow \widetilde{\mathbf{e}}_i^{\text{VA}}$
  end if
  $\mathcal{U} \leftarrow \mathcal{U} \,\|\, \mathbf{u}_i$ ▷ Append the selected embedding
end for
$y_{1:n} \leftarrow \textsc{DiscreteAnswer}(f_\theta, \mathcal{U}, x_{\text{text}}, x_{\text{img}})$
return $y_{1:n}$

anchoring. This ablation tests whether VAST depends on a narrowly tuned contrast strength or remains effective across a broader range of values. Moderate values of α should provide useful visual guidance, whereas overly large values may distort an otherwise reliable next-token distribution.

Observation: From fig 3(a), moderate contrast strengths yield the strongest performance, with M3CoT peaking at $\alpha = 0.5$ and MMBench at $\alpha = 1.0$. Performance is relatively stable for small values of α, but degrades noticeably as the contrast strength increases, suggesting that visual anchoring is most effective when used as a controlled adjustment to the original distribution rather than as an overly strong reweighting.

A.5 Appendix: Sensitivity to candidate width

We study the sensitivity of VAST to the candidate width K, which determines the number of top-ranked token candidates used to construct each soft representation. Because K affects both the composition of the candidate mixture and the maximum entropy $\log K$ considered by the activation rule, changes in K can influence both reasoning behavior and soft-activation frequency. This ablation tests whether VAST depends on a narrowly chosen candidate width or remains stable across a broader range of K values.

Observation: From Fig 3(b), VAST remains relatively stable across a broad range of candidate widths. As K varies from 3 to 20, accuracy remains within 64.8-65.9 on M3CoT and 83.7-84.2 on MMBench. These results indicate that VAST does not rely on a narrowly tuned candidate width and maintains comparable performance despite substantial changes in activation frequency.

Method	V*		BLINK				
	DA	**RP**	**Counting**	**IQ-Test**	**Jigsaw**	**Relative Reflectance**	**Spatial Relation**
Qwen2.5-VL-7B	78.26	76.32	66.70	26.00	52.67	37.31	88.11
SFT	82.60	73.70	67.50	26.70	45.30	33.60	88.80
LVR (4 steps)	**84.40**	76.30	69.20	28.70	**52.70**	42.50	**89.50**
LVR (8 steps)	**84.40**	77.60	70.00	**29.30**	52.00	42.50	86.00
LVR (16 steps)	81.70	**79.00**	**70.80**	27.30	**52.70**	41.80	87.40
VAST (training-free)	80.00	77.63	68.33	28.67	50.00	**45.52**	73.23

Table 6: **Comparison with training-based latent visual reasoning** on Qwen2.5-VL-7B. SFT and LVR require additional training; VAST does not. LVR is reported with 4, 8, and 16 latent reasoning steps. **Bold** marks the best result per task.

A.6 Thresholding Schedules

For the schedule-shape ablation, we fix the lower threshold at $\tau_{\text{low}} = 0.2$, the upper threshold at $\tau_{\text{high}} = 1.10$ with $K = 3$, and the schedule horizon at $S = 200$ generation steps. We compare Flat, Linear, Cosine, Cosine^2, Exponential, and our Position-Aware Valley schedule; for the latter, we use a valley fraction of $\nu = 0.35$, yielding a central low-threshold region surrounded by smooth high-to-low and low-to-high transitions.

A.7 Comparison with Training-Based Latent Reasoning

Several multimodal latent reasoning methods require additional training. To assess how a training-free approach compares, Table 6 evaluates VAST against supervised fine-tuning (SFT) and Latent Visual Reasoning (LVR) (Li et al., 2026), a representative training-based method. We train both SFT and LVR on Qwen2.5-VL-7B following the protocol of Li et al. (2026), and evaluate on the vision-centric benchmarks used in that work. V* (Wu & Xie, 2024) measures fine-grained visual detail search and relative spatial reasoning through its Direct Attributes (DA) and Relative Position (RP) subsets. From BLINK (Fu et al., 2024), a benchmark of expert-annotated, perception-heavy tasks, we use Counting (object enumeration), IQ-Test (visual pattern reasoning), Jigsaw (image reconstruction from fragments), Relative Reflectance (pixel-level albedo comparison), and Spatial Relation (object-relation understanding within a scene).

Despite requiring no training, VAST performs comparably to LVR on most tasks. It improves over both the base model and SFT on five of seven tasks, matches or surpasses individual LVR variants on V* RP and IQ-Test, and achieves the best result overall on Relative Reflectance. Excluding Spatial Relation, where VAST falls below the base model, its average accuracy is within 0.6 points of the best LVR variant. These results show that a substantial portion of the benefit of training-based latent reasoning can be obtained at inference time alone.