

SPOT-E: Test-Time Entropy Shaping with Visual Spotlights for Frozen VLMs

Bo Yin¹, Xiaobin Hu^{1*}, Chengming Xu², Ruolin Shen³, Mo Yang⁴,
Jiangning Zhang⁵, Peng-Tao Jiang⁶, Cheng Tan⁷, and Shuicheng Yan¹

¹ National University of Singapore, Singapore

² Fudan University, Shanghai, China

³ Technical University of Munich, Munich, Germany

⁴ Sagenic Tech, China

⁵ Zhejiang University, Hangzhou, China

⁶ vivo, China

⁷ Shanghai Artificial Intelligence Laboratory, Shanghai, China

Abstract. Vision-language models (VLMs) often underperform on evidence intensive tasks because decisive visual evidence are small, localized, and easy to overlook, leading to failures in evidence readout even when high-level reasoning is intact. Prior inference-time visual interventions can improve grounding without retraining, but they are largely open-loop and lack a mechanism to verify whether highlighted evidence is actually used. We study answer-span prediction entropy as a model-internal feedback signal and show that naive entropy minimization is ambiguous, since low entropy may arise from evidence-grounded confidence or short-cut collapse. To resolve this ambiguity, we introduce low-entropy anchors and an entropy-shaping objective that reduces answer uncertainty while preserving baseline high-confidence tokens. We instantiate this principle in SPOT-E, a plug-and-play test-time method that produces question-conditioned spotlights, optimized per instance via light-weight tuning based on Group Relative Policy Optimization (GRPO). Across all benchmarks and different VLM families, SPOT-E yields consistent gains and improved robustness under visual corruptions. Code is publicly available at: <https://github.com/YinBo0927/SPOT-E>

Keywords: Vision-language models · Test-time adaptation · Entropy

1 Introduction

Vision-language models have made rapid progress in multimodal understanding, yet they remain unreliable on evidence-intensive tasks such as chart reading and document parsing [14, 21, 23, 27, 28, 42, 48]. In these settings, the decisive evidence are often small and localized. A model may describe a correct reasoning plan, for example it may say “read the y-axis value and then compare the two bars”, but still misread the underlying number, causing the final answer to fail [10, 25,

* Corresponding author.

26, 37, 50–52, 55]. This pattern exposes a bottleneck that we call the evidence utilization gap: the model can reason about what evidence is needed, but cannot reliably extract and focus on the fine-grained visual evidence that determines the answer [34]. Closing this gap for frozen, already-deployed VLMs, without costly retraining or task-specific annotation, is practically urgent and methodologically challenging [45]. **How can we improve evidence utilization at inference time while keeping the backbone frozen?**

A natural direction is to intervene on the visual input at test time so that decisive regions become more accessible to the model [3, 33, 39, 41, 54]. Methods such as FGVP instantiate this idea and report gains without modifying model weights [40, 49]. However, these approaches are open-loop. They apply a fixed intervention but do not provide a mechanism to verify whether the model actually relied on the emphasized evidence. When the selected region misses the decisive evidence, or when the intervention degrades the evidence itself, the failure remains invisible to the method and therefore cannot be corrected. This raises a key question: **how can we design a test-time intervention that is both instance-adaptive and self-verifying?**

These limitations motivate a closed-loop feedback signal that can be obtained from a frozen VLM during inference [18, 31]. We find that answer entropy, computed from the logits over the tokens that form a structured final answer, tracks evidence usability. Entropy is low when the decisive visual cues are clear, and it increases when they are obscured [19, 20, 38, 43]. However, entropy reduction is inherently ambiguous. Both evidence-grounded confidence and shortcut behaviors can produce low entropy [4, 9, 44], as shown in Fig. 2. As a result, **minimizing entropy alone can drive the model into confident-but-wrong shortcuts**, creating a systematic failure mode in the absence of labels.

To resolve this ambiguity, we introduce an **entropy-shaping** principle that reduces answer entropy while preserving the model’s baseline high-confidence predictions on the unmodified input. Specifically, we identify low-entropy anchors as token positions where the frozen VLM is already near-deterministic at baseline, and we penalize interventions that disrupt these anchors. This yields a label-free objective that favors evidence-supported confidence over shortcut-induced collapse. We instantiate the principle in **SPOT-E**, a test-time visual adaptation framework that augments a frozen VLM with a lightweight, question-conditioned visual spotlight module that produces spotlights. For each instance, we optimize only the spotlight-module LoRA [13] parameters using Group Relative Policy Optimization [30] under the entropy-shaping reward.

Our contributions are summarized as follows:

- **Entropy Signal.** We identify answer entropy as a label-free signal for evidence utilization in frozen VLMs, and show that entropy reduction is inherently ambiguous.
- **Entropy Shaping.** We propose low-entropy anchors and an entropy-shaping reward to disambiguate evidence-supported confidence from shortcut collapse.

- **SPOT-E Framework.** We present SPOT-E, a plug-and-play test-time framework that keeps the VLM frozen while optimizing a question-conditioned visual spotlight via per-instance GRPO to leverage low-entropy anchors.
- **Broad Evaluation.** We conduct extensive evaluations across both open-source and closed-source VLM families and multiple backbones, demonstrating consistent gains on diverse benchmarks, with particularly strong improvements on evidence-intensive tasks and improved robustness under corruptions.

2 Related Work

2.1 Inference-Time Visual Interventions and Visual Prompting

Inference-time visual interventions improve grounding by manipulating the visual evidence presented to a frozen VLM at test time [2, 5]. Common approaches include overlaying marks or masks to highlight regions, as in FGVP and Set-of-Mark prompting, selectively cropping or zooming into candidate areas to preserve small details, as in ViCrop, and spatial transformations that reallocate resolution toward query-relevant evidence, as in AttWarp [8, 39, 40, 54]. Another practical direction uses attention- or score-guided prompting strategies such as API-style prompting to steer the model toward informative regions without re-training [47]. These methods are appealing because they are lightweight and model-agnostic, but they often depend on fixed heuristics, external region proposals, or discrete choices that may be brittle across instances. Our method fits this paradigm and uses a question-conditioned spotlight for adaptive evidence emphasis under a frozen backbone.

2.2 Entropy and Uncertainty for Evidence Localization

Uncertainty signals have long been used to diagnose and steer model behavior at inference time [12, 19]. Entropy over output distributions is a common proxy for confidence and has been used for calibration and selective prediction, self-consistency and re-ranking, as well as entropy- and confidence-driven decoding heuristics [11, 20, 36]. In vision and multimodal reasoning, uncertainty is also closely tied to evidence localization, since failures in fine-grained grounding often appear as high uncertainty concentrated on a small set of answer tokens [15, 22]. Recent multimodal work further leverages entropy or confidence to trigger additional perception, guide region selection, and filter visually unsupported generations [45, 46]. Our method combines an entropy-shaping reward with a question-conditioned visual spotlight to encourage decisive, visually supported answers without updating the frozen VLM.

3 Motivation

In this section, we motivate an entropy-centric view of visual adaptation in VLMs. When decisive visual evidence is usable, the model commits more consistently to a single final answer, making answer-span entropy a proxy for evidence

use. We show that this effect is spatially grounded and can change sharply with the visibility of the decisive region. However, entropy reduction is ambiguous and may also reflect shortcut behaviors that suppress hard evidence, yielding confident-but-wrong predictions. To resolve this, we introduce low-entropy anchors and an entropy-shaping principle that reduces answer entropy while preserving anchor stability.

3.1 Notation

Consider a frozen VLM F_ϕ , which typically comprises a visual encoder V , a multimodal connector F_c , and an LLM M . Given an input image x and a user instruction q , the model receives a multimodal sequence of visual tokens $\{v_1, v_2, \dots, v_n\}$ and text tokens $\{t_1, t_2, \dots, t_m\}$. During decoding, F_ϕ processes the concatenated multimodal context $\{v_1, \dots, v_n, t_1, \dots, t_m\}$ followed by previously generated tokens $\{y_1, \dots, y_{k-1}\}$ to predict the next token y_k .

Next-token distribution. At decoding step k , the model outputs a distribution over the vocabulary $\mathcal{W}$:

$$p_k(w) = p_\phi(w \mid x, q, y_{<k}), \quad w \in \mathcal{W}, \quad (1)$$

where $y_{<k} = (y_1, \dots, y_{k-1})$ is the prefix.

Entropy. We quantify predictive uncertainty using Shannon entropy of the next-token distribution:

$$H_k(x, q) = - \sum_{w \in \mathcal{W}} p_k(w) \log p_k(w). \quad (2)$$

Unless otherwise specified, $\log(\cdot)$ denotes the natural logarithm. We compute entropy under the baseline decoding trajectory.

Answer entropy. Long-form generations contain many tokens weakly related to visual evidence. To focus on evidence-relevant uncertainty, we compute entropy on a final answer span. We enforce a structured output format (e.g., **Final answer**: ...) and extract the token indices of the answer span as $\mathcal{T}_{\text{ans}}$. We define answer entropy as

$$H_{\text{ans}}(x, q) = \frac{1}{|\mathcal{T}_{\text{ans}}|} \sum_{k \in \mathcal{T}_{\text{ans}}} H_k(x, q). \quad (3)$$

For an intervention producing a modified input $\tilde{x}$, we measure entropy reduction by

$$\Delta H_{\text{ans}} = H_{\text{ans}}(\tilde{x}, q) - H_{\text{ans}}(x, q). \quad (4)$$

Low-entropy anchors. To characterize token positions that the base model is already confident about, we define a set of low-entropy anchors under the baseline input. Concretely, we select the K positions with the smallest next-token entropies:

$$\mathcal{I}_{\text{low}}(x, q) = \text{TopK}_{\text{small}}(\{H_k(x, q)\}_{k=1}^T), \quad (5)$$

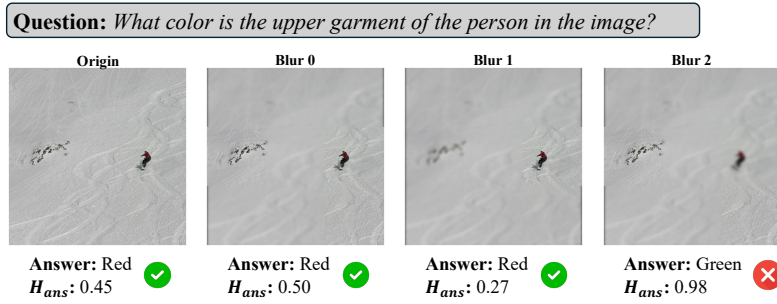


Fig. 1: Localized evidence controls answer entropy. We apply region-level blur to a subset of grid regions while keeping all other pixels unchanged, and measure $H_{ans}(\tilde{x}_S, q)$.

where T is the output length. In addition, we evaluate $H_k(\tilde{x}, q)$ by conditioning on the baseline token prefix to align anchor positions. These anchors are later used to distinguish desirable entropy reduction from shortcut behaviors.

3.2 Visual Evidence Shapes Answer Entropy

A natural intuition is that visual evidence affects a VLM’s answer mainly by changing how certain the model can be at the point of committing to the final answer. If the decisive evidence for q is clear, the model should concentrate probability mass on a consistent answer; if the evidence is weak or obscured, the model should remain uncertain. We operationalize this intuition using the answer entropy $H_{ans}(x, q)$ in Eq. (3), which measures uncertainty on the final answer.

We validate the intuition with a simple spatial sensitivity analysis. We partition the image into a coarse grid of regions $\mathcal{R} = \{r_1, \dots, r_M\}$ and apply localized interventions while keeping the rest of the input unchanged. For a given subset size n , we sample subsets $S \subseteq \mathcal{R}$ with $|S| = n$, construct an intervened image $\tilde{x}_S$ by applying a fixed region-level transformation only within regions in S such as blurring, and compute $H_{ans}(\tilde{x}_S, q)$. Fig. 1 shows a representative instance where suppressing the localized evidence required by q increases H_{ans} , eventually leading to incorrect predictions.

3.3 Entropy Reduction is Ambiguous

The previous subsection suggests that lowering H_{ans} often correlates with making decisive evidence more usable. However, entropy reduction alone is ambiguous: an intervention may decrease uncertainty not by improving evidence quality, but by attenuating hard evidence and steering the model toward priors or salient distractors, yielding confident-but-wrong answers with smaller H_{ans} .

We illustrate this ambiguity within the same subset-based intervention framework, but now restricting attention to the most entropy-sensitive singleton subset. Concretely, for any subset $S \subseteq \mathcal{R}$, we define the entropy change induced by

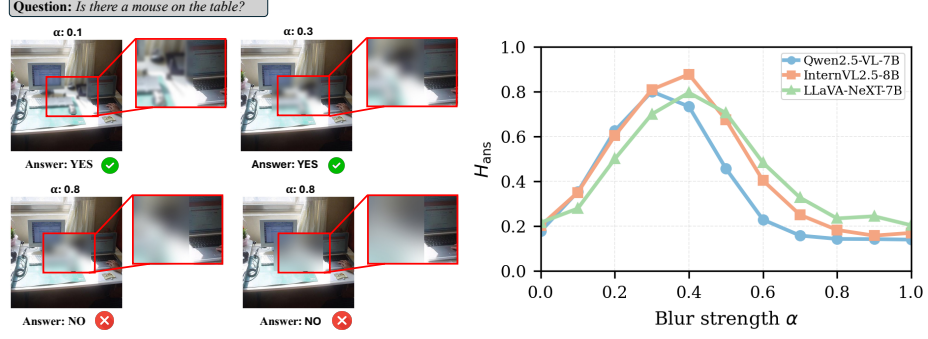


Fig. 2: Entropy reduction can be misleading. We blur only the most entropy-sensitive singleton subset S^* with strength α . H_{ans} often rises as evidence becomes ambiguous, but may drop again when evidence is erased.

intervening on S as

$$\Delta H_{\text{ans}}(S) = H_{\text{ans}}(x, q) - H_{\text{ans}}(\tilde{x}_S, q), \quad (6)$$

where $\tilde{x}_S$ modifies only regions in S and leaves all other pixels unchanged. We identify

$$S^* = \arg \max_{|S|=1} |\Delta H_{\text{ans}}(S)|, \quad (7)$$

i.e., the single-region subset whose intervention yields the largest magnitude of entropy change. We then keep the rest of the image fixed and construct a family of intervened inputs by blurring only within S^* with increasing strength α :

$$\tilde{x}_{S^*}^{(\alpha)} = \mathcal{A}_\alpha(x; S^*), \quad \alpha \in [0, 1], \quad (8)$$

where $\mathcal{A}_\alpha(\cdot; S^*)$ applies blur of level α to regions in S^* and leaves all other regions unchanged.

Sweeping α typically yields a non-monotonic profile: H_{ans} is low when the evidence is clear (small α), peaks when the evidence becomes ambiguous (intermediate α), and can decrease again once the evidence is effectively erased (large α), as the model collapses to a prior- or distractor-driven answer. Fig. 2 shows that low H_{ans} at large α can also arise from prior-driven overconfidence.

3.4 Entropy Shaping with Low-Entropy Anchors

Answer entropy H_{ans} provides a convenient scalar readout of how concentrated the model is on the final answer. The difficulty is that the same decrease in H_{ans} can be produced by different mechanisms: the intervention may genuinely expose missing evidence, or it may remove hard evidence and let the model settle on priors or distractors. To disambiguate these cases, we look beyond the answer entropy itself and ask whether an intervention is non-destructive, namely



Fig. 3: Low-entropy anchors reveal destructive shortcuts.

whether it preserves parts of the decoding process that the base model was already confident about. As illustrated in Fig. 3, two interventions may achieve a similar drop in answer entropy, yet only the evidence-revealing one preserves the baseline’s low-entropy tokens, whereas the destructive shortcut inflates their entropy motivating our anchor disruption measure.

We use low-entropy anchors $\mathcal{I}_{\text{low}}(x, q)$ in Eq. (5) to represent such stable positions under the baseline input. Given an intervened input $\tilde{x}$, we measure anchor disruption by the average entropy increase on anchor positions,

$$\Delta H_{\text{low}}(\tilde{x}) = \frac{1}{|\mathcal{I}_{\text{low}}|} \sum_{k \in \mathcal{I}_{\text{low}}} \max(0, H_k(\tilde{x}, q) - H_k(x, q)). \quad (9)$$

Interventions with comparable reductions in H_{ans} can behave very differently under this criterion. Evidence-revealing interventions tend to keep ΔH_{low} small, while shortcut interventions often reduce H_{ans} at the cost of increasing entropy on anchors. This motivates an entropy-shaping principle that favors reducing H_{ans} while preserving low-entropy anchors.

4 SPOT-E: Visual Spotlighting for Entropy Shaping

Building on Sec. 3, we introduce **SPOT-E**. As shown in Fig. 4, SPOT-E performs entropy-guided test-time visual adaptation by keeping the VLM F_ϕ frozen and optimizing a lightweight, question-conditioned visual spotlight. Given an image x and instruction q , the spotlight produces an intervened image $\tilde{x} = \mathcal{S}(x; m)$, where $\mathcal{S}$ is the spotlighting operator and m is a soft pixel mask. We feed $\tilde{x}$ into F_ϕ to compute the answer entropy $H_{\text{ans}}(\tilde{x}, q)$ and anchor entropies on $\mathcal{I}_{\text{low}}(x, q)$. At test time, SPOT-E runs a short per-instance episode: it samples candidate spotlights, scores them with an entropy-shaping objective, and updates only the spotlight parameters via GRPO [30]. The final prediction is chosen by Best-of- N over candidates, and the spotlight is reset after each instance.

4.1 Visual Spotlight

As illustrated in Fig. 5, SPOT-E introduces a CLIP-based [29] visual spotlight module to produce question-conditioned visual spotlights. Given instruction q ,

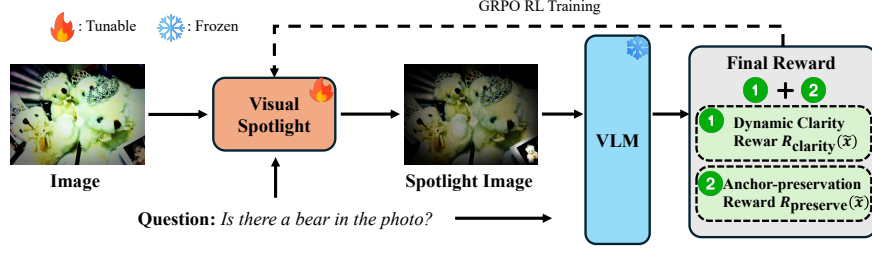


Fig. 4: SPOT-E overview. SPOT-E freezes the VLM and optimizes a lightweight visual spotlight at test time to generate an intervened image, scored by answer-entropy clarity and anchor-preservation.

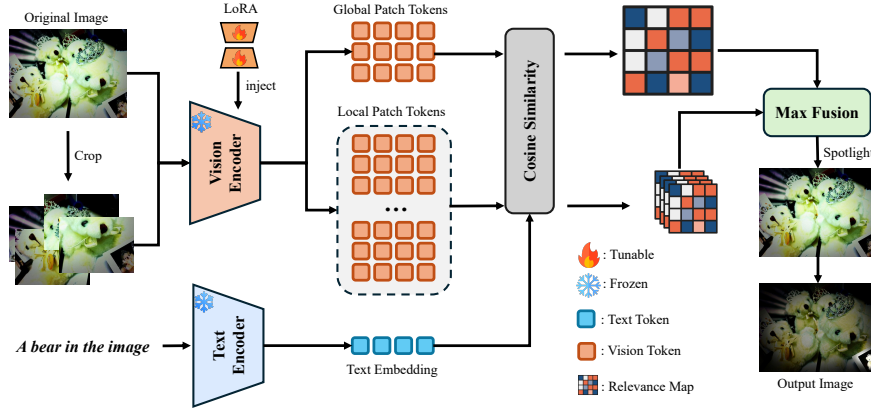


Fig. 5: SPOT-E visual spotlight module. Both the image encoder and the text encoder are CLIP. The module computes patch-text similarities on the global view and local crops, then fuses multi-view relevance maps via max pooling to produce the final spotlight mask.

we extract a compact visual phrase $\bar{q}$ by retaining the key entities and attributes relevant to visual grounding. Then the module extracts global patch tokens and local crop tokens with a CLIP vision encoder with LoRA [13] adapter, matches them to the frozen CLIP text embedding via patch-text similarity to obtain relevance maps, and fuses multi-view evidence by max pooling to form the final spotlight mask.

Global view and crop max fusion. We run the CLIP on the full image and on multiple crops to avoid missing small evidence. Let $x^{(0)} = x$ be the global view and $\{x^{(i)}\}_{i=1}^{N_c}$ be crops. For each view $x^{(i)}$, CLIP outputs patch tokens $\{p_j^{(i)}\}_{j=1}^{N_i}$ and a text embedding $t(\bar{q})$. We compute patch relevance by cosine similarity:

$$u_j^{(i)} = \left\langle \text{norm}(p_j^{(i)}), \text{norm}(t(\bar{q})) \right\rangle. \quad (10)$$

We reshape $\{u_j^{(i)}\}$ to a 2D grid and map each crop grid back to the full-image coordinates via $\mathcal{W}_i(\cdot)$, which warps the i -th crop map to the global image coordinates. The fused relevance map is obtained by max fusion:

$$u = \max\left(u^{(0)}, \max_{i=1,\dots,N_c} \mathcal{W}_i\left(u^{(i)}\right)\right). \quad (11)$$

Soft mask and spotlighting operator. We upsample the fused relevance map u to the image resolution and obtain a soft pixel mask

$$m = \sigma\left(\frac{1}{\tau} u^\uparrow\right) \in [0, 1]^{H \times W}, \quad (12)$$

where $u^\uparrow$ denotes bilinear upsampling of u and τ controls mask sharpness. We then form the intervened input via

$$\tilde{x} = \mathcal{S}(x; m) = m \odot x + (1 - m) \odot \mathcal{B}(x), \quad (13)$$

where $\odot$ denotes element-wise multiplication and $\mathcal{B}(\cdot)$ is a fixed background-degrading transform (background dimming).

4.2 Entropy-Shaping Reward

Given a candidate spotlight mask m and intervened input $\tilde{x} = \mathcal{S}(x; m)$, we score it using an entropy-shaping reward

$$R(\tilde{x}) = R_{\text{clarity}}(\tilde{x}) + R_{\text{preserve}}(\tilde{x}). \quad (14)$$

Both terms are computed from the frozen VLM logits, while gradients update only the visual spotlight module.

Dynamic clarity reward. We encourage the model to become more decisive on the final answer span by reducing answer entropy:

$$\Delta H_{\text{ans}}(\tilde{x}) = H_{\text{ans}}(x, q) - H_{\text{ans}}(\tilde{x}, q). \quad (15)$$

To avoid over-optimizing when the base model is already confident, we apply a dynamic scaling factor based on the baseline answer entropy:

$$R_{\text{clarity}}(\tilde{x}) = \gamma(x, q) \cdot \Delta H_{\text{ans}}(\tilde{x}), \quad \gamma(x, q) = \frac{H_{\text{ans}}(x, q)}{H_{\text{ans}}(x, q) + c}, \quad (16)$$

where $c > 0$ is a small constant. When the baseline is uncertain, γ increases and entropy reduction is rewarded more; when the baseline is already confident, γ suppresses blind exploration.

Anchor-preservation reward. Entropy reduction can be achieved by shortcut behaviors that disrupt tokens the base model was already confident about. We therefore penalize entropy increases on low-entropy anchor positions $\mathcal{I}_{\text{low}}(x, q)$ (Eq. (5)) using the anchor disruption measure $\Delta H_{\text{low}}(\tilde{x})$ (Eq. (9)):

$$R_{\text{preserve}}(\tilde{x}) = -\lambda \cdot \Delta H_{\text{low}}(\tilde{x}), \quad (17)$$

where λ controls the strength of anchor preservation.

4.3 Test-Time Optimization with GRPO

SPOT-E runs a short per-instance test-time optimization episode to adapt the visual spotlight module, while keeping the VLM F_ϕ fully frozen. Let θ denote the visual spotlight parameters (LoRA adapters in CLIP vision encoder attention layers), initialized to θ_0 for each instance.

Group sampling and scoring. At each iteration, we sample a group of N candidate masks by injecting Gaussian noise into the visual spotlight and obtain intervened inputs $\{\tilde{x}^{(n)}\}_{n=1}^N$. Each candidate is scored by the total reward $R(\tilde{x}^{(n)})$ (Eq. (14)).

Group-relative advantages. We compute standardized advantages within the group:

$$\mu_R = \frac{1}{N} \sum_{n=1}^N R(\tilde{x}^{(n)}), \quad \sigma_R = \sqrt{\frac{1}{N} \sum_{n=1}^N (R(\tilde{x}^{(n)}) - \mu_R)^2}. \quad (18)$$

$$A^{(n)} = \frac{R(\tilde{x}^{(n)}) - \mu_R}{\sigma_R + \epsilon}. \quad (19)$$

where ϵ is a small constant.

GRPO update and reset. We apply a standard GRPO clipped policy update on θ using $\{A^{(n)}\}$, with a KL regularizer to keep the visual spotlight close to its initialization.

$$\bar{r}^{(n)}(\theta) = \text{clip}(r^{(n)}(\theta), 1 - \delta, 1 + \delta). \quad (20)$$

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{N} \sum_{n=1}^N \min(r^{(n)}(\theta)A^{(n)}, \bar{r}^{(n)}(\theta)A^{(n)}) + \beta \text{KL}(\pi_\theta \| \pi_{\theta_0}). \quad (21)$$

where δ is the clipping threshold and β controls the KL strength. We update only θ by gradient descent on Eq. (21), and reset $\theta \leftarrow \theta_0$ after each instance to avoid cross-sample drift.

5 Experiments

Overview. We evaluate SPOT-E from four complementary angles. First, we report main results across a broad set of frozen backbones, covering both open-source model families and closed-source VLM APIs, to test generality. Second, we compare against strong inference-time visual prompting baselines under matched decoding settings. Third, we assess out-of-distribution robustness under controlled visual corruptions and analyze confidence behavior through answer entropy. Finally, we conduct targeted ablations on the reward, spotlight design, and test-time budget, and provide qualitative case studies to illustrate how SPOT-E changes evidence usage at inference time. Due to space constraints, additional experimental results, such as those using larger backbones, are provided in the Appendix.

Table 1: Applying SPOT-E to closed-source and open-source backbones.

Base Model	TextVQA	DocVQA	ChartQA	MathVista	MMMU	GQA	MMBench	POPE
<i>Closed-Source</i>								
GPT-4o [17]	77.4	91.1	86.7	63.5	69.2	73.0	83.1	86.9
+ SPOT-E (Ours)	79.9 +2.5↑	92.3 +1.2↑	88.2 +1.5↑	65.5 +2.0↑	70.4 +1.2↑	73.8 +0.8↑	83.9 +0.8↑	87.9 +1.0↑
GPT-4o-mini [17]	70.0	86.0	80.0	55.0	60.0	68.0	78.0	84.0
+ SPOT-E (Ours)	73.5 +3.5↑	88.0 +2.0↑	82.5 +2.5↑	58.0 +3.0↑	62.0 +2.0↑	69.0 +1.0↑	79.2 +1.2↑	85.2 +1.2↑
Gemini-2.5-Flash [7]	80.0	91.5	84.0	68.0	70.0	72.0	82.0	86.0
+ SPOT-E (Ours)	82.8 +2.8↑	93.0 +1.5↑	85.8 +1.8↑	70.2 +2.2↑	71.5 +1.5↑	72.8 +0.8↑	82.8 +0.8↑	86.9 +0.9↑
<i>Open-Source</i>								
Qwen2.5-VL-7B [35]	84.9	85.7	87.3	67.8	55.0	64.0	82.6	86.4
+ SPOT-E (Ours)	86.9 +2.0↑	86.5 +0.8↑	88.5 +1.2↑	70.8 +3.0↑	58.5 +3.5↑	65.0 +1.0↑	83.5 +0.9↑	87.4 +1.0↑
Qwen3-VL-8B [1]	86.0	86.2	88.0	70.5	58.0	65.5	83.8	87.2
+ SPOT-E (Ours)	87.8 +1.8↑	86.8 +0.6↑	89.0 +1.0↑	73.3 +2.8↑	61.0 +3.0↑	66.4 +0.9↑	84.6 +0.8↑	88.1 +0.9↑
LLaVA-NeXT-7B [23]	78.5	80.0	79.0	47.0	38.0	63.0	75.0	85.0
+ SPOT-E (Ours)	84.7 +6.2↑	82.0 +2.0↑	81.5 +2.5↑	50.5 +3.5↑	41.0 +3.0↑	64.5 +1.5↑	76.5 +1.5↑	86.5 +1.5↑
LLaVA-OV-7B [23]	80.0	81.0	80.5	48.5	39.5	63.5	76.0	85.5
+ SPOT-E (Ours)	85.0 +5.0↑	82.8 +1.8↑	82.8 +2.3↑	51.8 +3.3↑	42.5 +3.0↑	64.9 +1.4↑	77.4 +1.4↑	86.9 +1.4↑
InternVL2.5-8B [6]	81.0	82.0	83.0	66.0	56.0	62.0	80.5	89.0
+ SPOT-E (Ours)	84.5 +3.5↑	83.5 +1.5↑	85.0 +2.0↑	69.5 +3.5↑	59.5 +3.5↑	63.2 +1.2↑	81.5 +1.0↑	90.0 +1.0↑
InternVL3-8B [6]	80.2	82.7	86.6	71.6	62.7	61.0	81.7	91.1
+ SPOT-E (Ours)	82.4 +2.2↑	83.7 +1.0↑	88.1 +1.5↑	74.4 +2.8↑	65.2 +2.5↑	62.2 +1.2↑	82.5 +0.8↑	92.0 +0.9↑

5.1 Implementation Details

Models. To verify the effectiveness of our approach, we apply SPOT-E to multiple frozen open-source VLM backbones spanning three representative families: Qwen-VL [35], LLaVA [23], and InternVL [6], and three proprietary VLM APIs that expose token-level log probabilities: GPT-4o, GPT-4o-mini [17], and Gemini-2.5-Flash [7]. Unless otherwise stated, all backbones are kept fully frozen and SPOT-E updates only the CLIP-based eye module at test time with per-instance reset.

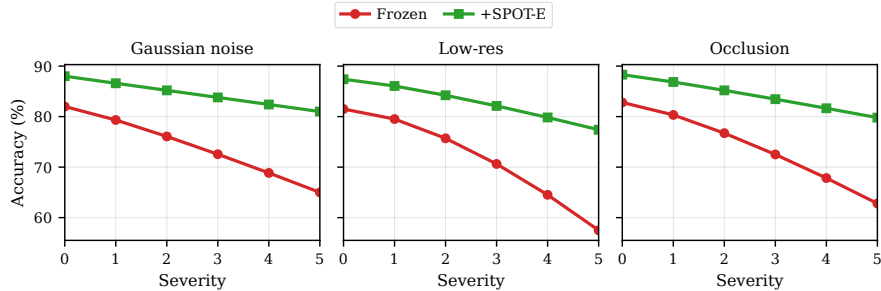
Evaluation. Our evaluation comprises multiple benchmarks that stress fine-grained visual grounding and localized evidence usage, spanning text-centric grounding (TextVQA [32], DocVQA [28], ChartQA [27]), compositional VQA and general multimodal understanding (GQA [16], MMBench [24]), knowledge- and reasoning-intensive tasks (MathVista [26], MMMU [53]), and hallucination-oriented evaluation (POPE [22]). We follow the standard evaluation protocols and report the official metrics for each benchmark.

5.2 Main Results

Consistent Improvements Across Frozen Backbones. We compare the frozen base model and its SPOT-E augmented version across both closed-source VLM APIs and open-source backbones which cover three backbone families with two released variants per family. Table 1 shows that SPOT-E yields consistent gains across all evaluated models. The improvements are most pronounced on evidence-intensive benchmarks such as TextVQA, DocVQA, ChartQA, and MathVista, where answers depend on small text, numbers, or localized symbols.

Table 2: Comparison with inference-time visual evidence manipulation baselines.

Method	TextVQA	GQA	MMMU	POPE	DocVQA
FGVP-Mask [40]	77.3	55.8	46.0	84.4	56.6
FGVP-RBM [40]	72.3	55.8	46.5	81.3	38.6
SoM [39]	61.5	47.8	45.1	75.8	57.4
API [47]	81.6	61.1	47.4	85.8	68.4
ViCrop [54]	83.8	60.6	47.1	86.7	82.5
AttWarp [8]	84.7	64.0	50.4	87.4	84.1
SPOT-E	86.9	65.0	58.5	87.4	86.5

**Fig. 6:** Out-of-distribution evaluation.

On broader multimodal reasoning benchmarks (GQA, MMBench, and MMMU), SPOT-E still provides positive but typically smaller gains, suggesting that suppressing distractors and amplifying decisive regions complements backbone reasoning capacity rather than replacing it. Finally, on POPE, SPOT-E tends to improve factual consistency by steering generation toward visually supported answers, indicating that entropy-guided spotlighting can mitigate confident-but-unsupported responses even without modifying the underlying VLM.

Comparison with Visual Prompting Baselines. Since several strong baselines improve grounding by manipulating visual evidence at inference time, we compare SPOT-E with representative methods including FGVP [40], SoM [39], API [47], ViCrop [54], and AttWarp [8]. All methods are evaluated on the frozen Qwen2.5-VL-7B backbone under the same decoding configuration, following each baseline’s standard inference-time procedure. As shown in Table 2, SPOT-E is competitive with these visual intervention baselines and yields further improvements across benchmarks, with particularly strong gains on evidence-intensive tasks where small, localized cues are critical.

Out-of-distribution Robustness under Visual Corruptions. To evaluate robustness under domain shift, we test on TextVQA with three synthetic corruptions applied at inference time: Gaussian noise, low-resolution downsampling, and local occlusion. We sweep corruption severity and plot accuracy curves for the frozen Qwen2.5-VL-7B baseline and +SPOT-E under the same decoding setting. As shown in Fig. 6, SPOT-E consistently reduces performance drop across severities, indicating improved robustness to corrupted visual evidence.

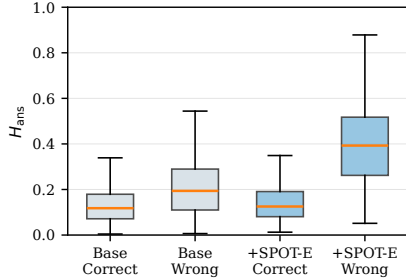


Fig. 7: Confidence calibration boxplot.

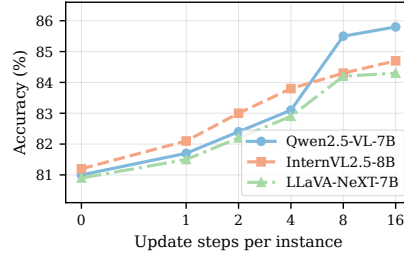


Fig. 8: Test-time budget discussion.

Confidence Calibration via Answer Entropy. We analyze how SPOT-E affects overconfident errors by measuring the answer entropy on each example. Fig. 7 reports boxplots of H_{ans} for correct and incorrect predictions under the frozen baseline and +SPOT-E. SPOT-E increases entropy on incorrect cases while maintaining low entropy on correct ones, reducing confident-but-unsupported responses and improving the separation between correct and wrong predictions.

5.3 Ablation Studies

We ablate three factors that govern SPOT-E: (i) the reward design, (ii) the spotlight configuration, and (iii) the test-time update budget (number of adaptation steps per instance). Unless otherwise stated, all ablations are conducted on frozen Qwen2.5-VL-7B under the same decoding setup as the main results, and we report the official benchmark metrics. For fair comparison, we keep the spotlight operator, learning rate, and evaluation prompts fixed, varying only the targeted component in each study.

Reward Design. We ablate the entropy-shaping reward in Sec. 4.2 to quantify the contribution of each term in Eq. (14). Keeping the spotlight mechanism and the test-time budget fixed, we compare: (i) *Clarity-only*, using R_{clarity} alone; (ii) *Preserve-only*, using R_{preserve} alone; (iii) *Full reward*, using $R_{\text{clarity}} + R_{\text{preserve}}$; and (iv) *w/o dynamic scaling*, where we replace $\gamma(x, q)$ in Eq. (16) with a constant factor. Tab. 3 shows that combining clarity and anchor preservation yields the most consistent gains, while removing either term degrades performance. These results suggest that R_{clarity} and R_{preserve} are complementary. The clarity term encourages decisiveness on the answer span, while the preservation term discourages shortcut updates that disrupt already-reliable evidence.

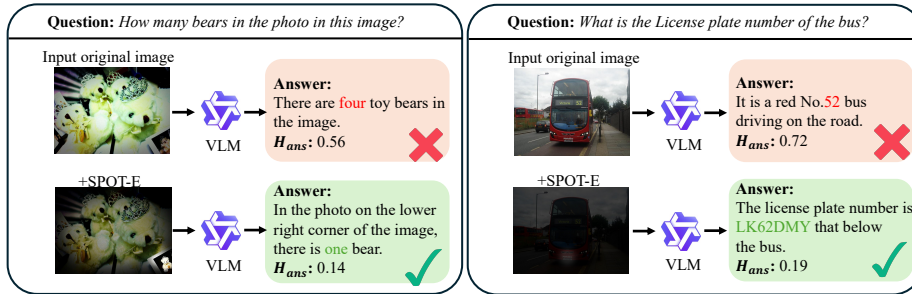
Visual Spotlight Design. We ablate the spotlight module in Sec. 4.1 to assess the impact of multi-view fusion and the spotlighting operator. Keeping the reward and test-time budget fixed, we compare: (i) GLOBAL (only the global view; $N_c=0$), (ii) MEANFUSE (average fusion; $\max \rightarrow \text{mean}$ in Eq. (11)), and (iii) NOBGDEG (no background degradation; $\mathcal{B}(x)=x$ in Eq. (13)). Table 4 shows that the default design (DEFAULT) provides the most consistent improvements, while removing crops or disabling background degradation reduces the benefit of evidence localization.

Table 3: Reward design ablation.

Variant	TextVQA	MathVista	POPE
R_{clarity} only	85.8	69.4	86.9
R_{preserve} only	84.9	68.6	87.1
w/o dynamic	86.4	70.1	87.2
Ours	86.9	70.8	87.4

Table 4: Spotlight design ablation.

Variant	TextVQA	MathVista	POPE
GLOBAL	84.1	66.9	86.3
MEANFUSE	85.7	68.5	86.9
NOBGDEG	85.2	67.8	86.7
Ours	86.9	70.8	87.4

**Fig. 9:** Qualitative case studies comparing the frozen baseline and +SPOT-E with the same inference setup.

Test-Time Budget. We vary the test-time adaptation budget by sweeping the number of eye-module update steps per instance in 0, 1, 2, 4, 8, 16, fixing the reward, spotlight design, learning rate, and decoding, where 0 is the frozen baseline. We evaluate on Qwen2.5-VL-7B, InternVL2.5-8B, and LLaVA-NeXT-7B. As shown in Fig. 8, accuracy increases with more steps and then saturates; we use 8 steps by default as it captures most gains with modest overhead.

5.4 Case Studies

We provide qualitative comparisons to show how SPOT-E changes visual evidence usage at inference time. Fig. 9 contrasts the frozen baseline and +SPOT-E under the same decoding setup, showing the original input, the spotlight-intervened image, and the resulting outputs. In both examples, the baseline is distracted by salient but irrelevant regions and answers incorrectly with higher H_{ans} , whereas SPOT-E suppresses distractors, amplifies the decisive evidence, and produces a visually supported answer with lower H_{ans} .

6 Conclusion

We present SPOT-E, a plug-and-play test-time method that strengthens fine-grained visual evidence utilization in frozen VLMs via lightweight per-instance adaptation of a question-conditioned visual spotlight module. Across diverse backbones and benchmarks, SPOT-E delivers consistent gains and stronger robustness under visual corruptions without retraining the base model, and analyses highlight remaining failures on extremely small or inherently ambiguous evidence.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., Efros, A.: Visual prompting via image inpainting. *Advances in neural information processing systems* **35**, 25005–25017 (2022)
3. Brown, B., Juravsky, J., Ehrlich, R., Clark, R., Le, Q.V., Ré, C., Mirhoseini, A.: Large language monkeys: Scaling inference compute with repeated sampling. arXiv preprint arXiv:2407.21787 (2024)
4. Carter, B., Jain, S., Mueller, J.W., Gifford, D.: Overinterpretation reveals image classification model pathologies. *Advances in Neural Information Processing Systems* **34**, 15395–15407 (2021)
5. Chen, A., Yao, Y., Chen, P.Y., Zhang, Y., Liu, S.: Understanding and improving visual prompting: A label-mapping perspective. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 19133–19143 (2023)
6. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 24185–24198 (2024)
7. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
8. Dalal, D., Vashishtha, G., Mishra, U., Kim, J., Kanda, M., Ha, H., Lazebnik, S., Ji, H., Jain, U.: Constructive distortion: Improving mllms with attention-guided image warping. arXiv preprint arXiv:2510.09741 (2025)
9. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**(11), 665–673 (2020)
10. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al.: Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 14375–14385 (2024)
11. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: *International conference on machine learning*. pp. 1321–1330. PMLR (2017)
12. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. *Iclr* **1**(2), 3 (2022)
14. Hu, X., Qian, Y., Yu, J., Liu, J., Ji, X., Xu, C., Tang, P., Xu, C., Tang, P., Liu, J., et al.: The landscape of medical agents: A survey (2026)
15. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13418–13427 (2024)

16. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
17. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
18. Jian, P., Wu, J., Sun, W., Wang, C., Ren, S., Zhang, J.: Look again, think slowly: Enhancing visual reflection in vision-language models. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 9262–9281 (2025)
19. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al.: Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221 (2022)
20. Kuhn, L., Gal, Y., Farquhar, S.: Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. arXiv preprint arXiv:2302.09664 (2023)
21. Li, Q., Yin, B., Huang, W., Liu, R., Zou, B., Yu, R., Ye, J., Yu, W., Wang, X.: Vision-language-action safety: Threats, challenges, evaluations, and mechanisms. arXiv preprint arXiv:2604.23775 (2026)
22. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 292–305 (2023)
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *Advances in neural information processing systems* **36**, 34892–34916 (2023)
24. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
25. Liu, Y., Tang, H., Peng, J., Zhang, J., Ji, X., He, Q., Wu, W., Luo, D., Gan, Z., Zhu, J., et al.: Human-mme: A holistic evaluation benchmark for human-centric multimodal large language models. arXiv preprint arXiv:2509.26165 (2025)
26. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
27. Masry, A., Do, X.L., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of the association for computational linguistics: ACL 2022. pp. 2263–2279 (2022)
28. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
30. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
31. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., Yao, S.: Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems* **36**, 8634–8652 (2023)

32. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
33. Snell, C., Lee, J., Xu, K., Kumar, A.: Scaling llm test-time compute optimally can be more effective than scaling model parameters. arXiv preprint arXiv:2408.03314 (2024)
34. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9568–9578 (2024)
35. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
36. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
37. Xing, Y., Hu, X., He, Q., Zhang, J., Yan, S., Lu, S., Jiang, Y.G.: Boosting reasoning in large multimodal models via activation replay. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19229–19240 (2026)
38. Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., Hooi, B.: Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. arXiv preprint arXiv:2306.13063 (2023)
39. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv preprint arXiv:2310.11441 (2023)
40. Yang, L., Wang, Y., Li, X., Wang, X., Yang, J.: Fine-grained visual prompting. *Advances in Neural Information Processing Systems* **36**, 24993–25006 (2023)
41. Yin, B., Hu, X., Zhou, X., Jiang, P.T., Liao, Y., Zhu, J., Zhang, J., Tai, Y., Wang, C., Yan, S.: Fera: Frequency-energy constrained routing for effective diffusion adaptation fine-tuning. arXiv preprint arXiv:2511.17979 (2025)
42. Yin, B., Li, Q., Wang, X.: On-policy self-evolution via failure trajectories for agentic safety alignment. arXiv preprint arXiv:2605.11882 (2026)
43. Yin, B., Li, Q., Yu, R., Wang, X.: Refinement provenance inference: Detecting llm-refined training prompts from model behavior. arXiv preprint arXiv:2601.01966 (2026)
44. Yin, B., Yang, X., Wang, X.: Don’t forget the nonlinearity: Unlocking activation functions in efficient fine-tuning. arXiv preprint arXiv:2509.13240 (2025)
45. Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., Chen, E.: Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences* **67**(12), 220105 (2024)
46. Yin, Z., Sun, Q., Guo, Q., Wu, J., Qiu, X., Huang, X.J.: Do large language models know what they don’t know? In: Findings of the association for Computational Linguistics: ACL 2023. pp. 8653–8665 (2023)
47. Yu, R., Yu, W., Wang, X.: Attention prompting on image for large vision-language models. In: European Conference on Computer Vision. pp. 251–268. Springer (2024)
48. Yu, X., Chen, Z., He, Y., Fu, T., Yang, C., Xu, C., Ma, Y., Hu, X., Cao, Z., Xu, J., et al.: The latent space: Foundation, evolution, mechanism, ability, and outlook. arXiv preprint arXiv:2604.02029 (2026)

49. Yu, X., Xu, C., Chen, Z., Yin, B., Yang, C., He, Y., Hu, Y., Zhang, J., Tan, C., Hu, X., et al.: Dual latent memory for visual multi-agent system. arXiv preprint arXiv:2602.00471 (2026)
50. Yu, X., Xu, C., Chen, Z., Zhang, Y., Lu, S., Yang, C., Zhang, J., Yan, S., Hu, X.: Visual document understanding and reasoning: A multi-agent collaboration framework with agent-wise adaptive test-time scaling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12300–12311 (2026)
51. Yu, X., Xu, C., Zhang, G., Chen, Z., Zhang, Y., He, Y., Jiang, P.T., Zhang, J., Hu, X., Yan, S.: Vismem: Latent vision memory unlocks potential of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 31544–31555 (2026)
52. Yu, X., Xu, C., Zhang, G., He, Y., Chen, Z., Xue, Z., Zhang, J., Liao, Y., Hu, X., Jiang, Y.G., et al.: Visual multi-agent system: Mitigating hallucination snowballing via visual flow. arXiv preprint arXiv:2509.21789 (2025)
53. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., 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)
54. Zhang, J., Khayatkhoei, M., Chhikara, P., Ilievski, F.: Visual cropping improves zero-shot question answering of multimodal large language models. In: R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models (2023)
55. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 (2023)