ReflectCAP: Detailed Image Captioning with Reflective Memory

Kyungmin Min[1], Minbeom Kim[1], Kang-il Lee[2], Seunghyun Yoon[3], Kyomin Jung[1,2⋆]

[1] IPAI, Seoul National University
[2] Dept. of ECE, Seoul National University
[3] Adobe Research
{kyungmin97, kjung}@snu.ac.kr

Abstract. Detailed image captioning demands both factual grounding and fine-grained coverage, yet existing methods have struggled to achieve them simultaneously. We address this tension with Reflective Note-Guided Captioning (ReflectCAP), where a multi-agent pipeline analyzes what the target large vision-language model (LVLM) consistently hallucinates and what it systematically overlooks, distilling these patterns into reusable guidelines called Structured Reflection Notes. At inference time, these notes steer the captioning model along both axes—what to avoid and what to attend to—yielding detailed captions that jointly improve factuality and coverage. Applying this method to 8 LVLMs spanning the GPT-4.1 family, Qwen series, and InternVL variants, ReflectCAP reaches the Pareto frontier of the trade-off between factuality and coverage, and delivers substantial gains on CapArena-Auto, where generated captions are judged head-to-head against strong reference models. Moreover, ReflectCAP offers a more favorable trade-off between caption quality and compute cost than model scaling or existing multi-agent pipelines, which incur 21–36% greater overhead. This makes high-quality detailed captioning viable under real-world cost and latency constraints.

Keywords: Detailed Image Captioning · Large Vision-Language Models · Multi-Agent Systems

1 Introduction

Hyper-detailed captions capture not only salient objects but also their attributes, orientations, spatial relations, background context, and subtle visual states, forming a comprehensive textual representation of an image. Such captions have become a key ingredient in downstream multimodal systems—improving prompt fidelity for text-to-image and text-to-video generation, and serving as a reasoning aid for compositional and grounded decision-making [1,2,8,9,12,24]. Large vision-language models (LVLMs) can produce these descriptions fluently [5, 19, 20, 40], yet they frequently hallucinate—generating text that is not grounded in the

⋆ Corresponding authors

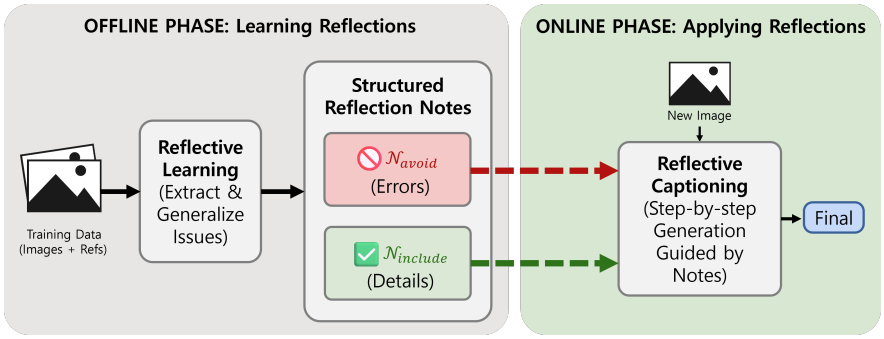

Fig. 1: Overview of ReflectCAP. In the offline phase, a multi-agent reflective learning pipeline distills a target LVLM's recurring captioning errors and omissions into Structured Reflection Notes. In the online phase, these notes guide caption generation for new images, producing captions that better balance factuality and coverage.

image. This limitation is widely attributed to the tendency of language priors to progressively dominate over visual evidence as generation length increases, leading the model to describe what is statistically probable rather than what is actually depicted [14, 15, 18, 25]. Hyper-detailed captioning, which inherently requires such extended generation, thus remains a critical bottleneck for reliable deployment.

The straightforward remedy is supervised fine-tuning on human-authored detailed captions [8, 26] to improve LVLMs' intrinsic performance; however, as we demonstrate in Section 5.1, such captions often exceed the model's perceptual capacity, even increasing hallucinations well beyond the base model. Moreover, this approach requires not only expensive human annotation but also additional training, further limiting its practicality. An alternative is inference-time correction, where the model iteratively revises its own output without additional training—an approach that has proven effective for LLMs [13, 22]. However, recent studies show that LVLMs struggle to self-correct during inference without external feedback, as they tend to confirm rather than rectify their own errors [10, 38]. Moreover, iterative revision lengthens the text context, further amplifying language-prior reliance over visual evidence [15, 25]. Taken together, these limitations suggest that a single LVLM alone is unlikely to fully address the detail–faithfulness tension, and that external guidance from a multi-agent system is needed.

To this end, we propose **ReflectCAP** (Reflective Note-Guided Captioning), a gradient-free framework that distills a target LVLM's recurring errors into an agentic memory called *Structured Reflection Notes* and leverages them for inference-time steering (Figure 1). ReflectCAP operates in two distinct phases. In the offline phase, a multi-agent pipeline critiques the target model's captions against a small set of human-annotated references to diagnose its systematic error patterns, separately encoding 1) **hallucination patterns** and 2) **omission patterns** into generalized guideline notes. In the online phase, each set

of notes serves a distinct role in steering generation: one produces a grounded base caption that suppresses recurring hallucinations, and the other produces a detail-focused caption covering typically neglected visual elements. A final merge step combines both with the image as a reference, producing a comprehensive caption that improves factuality and coverage simultaneously.

Empirically, ReflectCAP sets new pareto frontier under the factuality–coverage evaluation proposed by Lee et al. [15], substantially expanding coverage while preserving precision and achieving the highest F1 across model families. This improvement also holds on CapArena-Auto [3], a pairwise benchmark that evaluates captions holistically and is closely aligned with human preference. On 600 images evaluated with CapArena-Auto, ReflectCAP yields substantial win-rate improvements of +32.2 and +21.9 points over the corresponding baselines for the GPT and open-source model families, respectively. Furthermore, ReflectCAP can elevate smaller models beyond flagship baselines; for instance, GPT-4.1-mini with our method surpasses GPT-5.2 in CapArena-Auto win rate. Beyond performance comparisons, ReflectCAP consistently offers a more compute-efficient path to improving caption quality than either scaling model size or scaling inference-time computation; for example, InternVL3.5-4B with ReflectCAP achieves a comparable factuality–coverage F1 to InternVL3.5-38B while requiring approximately 8 times lower inference TFLOPs, and outperforms existing multi-agent pipelines while incurring 21–36% lower compute overhead, suggesting that ReflectCAP can serve as a practical, cost-effective alternative to both model scaling and inference-time computation scaling for detailed captioning.

2 Related Work

2.1 Detailed Image Captioning

Obtaining large-scale detailed image captions is increasingly important, as they serve as training signals for text-to-image and text-to-video generation and as reasoning aids for compositional vision–language tasks [1, 2, 8, 9]. Approaches to scaling detailed captions broadly follow two directions. The first relies on human-authored dense captions (e.g., DCI [32], DOCCI [26], IIW [8]), which offer high fidelity and coverage but are prohibitively expensive to scale beyond limited datasets. The second approach employs LVLMs as automated captioners. However, in long-form generation, these models often over-rely on language priors, which results in hallucinated but unsupported details [14, 25] and the omission of subtle visual attributes [7, 23, 28]. Many existing mitigation methods primarily improve factuality (precision) without comparably improving descriptive coverage (recall), leaving the two principal quality axes of detailed captioning in tension [6, 11, 16, 34, 39, 41]. Our work addresses this trade-off by distilling reflection notes that explicitly capture recurring hallucination patterns and missing-detail patterns, enabling separate control of hallucination suppression and detail recovery at inference time.

2.2 Reflective Memory in Agentic Frameworks

Recent advancements in LLM-based agents have successfully leveraged reflective memory to refine behavior in long-horizon tasks. By analyzing historical trajectories and inference-time feedback, these text-based models synthesize reusable reasoning strategies and deploy them at appropriate moments to guide subsequent multi-step decision-making [27, 29, 31, 33, 42].

However, extending this long-horizon reflection paradigm to LVLMs is fundamentally problematic. Unlike pure text generation, LVLMs struggle to reliably extract error patterns over extended reasoning chains [10, 38]. As the number of inference steps increases, the visual evidence becomes increasingly diluted, and the models fall prey to severe language prior phenomena—relying more on text-induced hallucinations than on the grounded visual input [4, 17, 25, 30]

To address these intrinsic limitations, we shift from online, long-horizon trajectory tracking to an offline, bottom-up distillation process. By systematically aggregating image-specific feedback across diverse samples, we identify and distill the target LVLM's recurring hallucination and omission patterns into a generalized reflection memory. This distilled memory is then utilized to proactively steer the model during inference, bypassing the need for costly step-by-step refinement.

3 Reflective Note-Guided Captioning Framework

We introduce Reflective Note-Guided Captioning (ReflectCAP), a framework that 1) distills systematic error patterns of a target LVLM into compact, reusable directives, and 2) leverages them as guidance to improve both factuality and detailedness in image captioning. The key insight is that large vision-language models exhibit predictable failure modes, including recurring hallucinations and consistent blind spots. Once surfaced, these patterns can be counteracted through lightweight prompt-level intervention rather than costly retraining or multi-agent inference at test time. ReflectCAP operationalizes this insight in two stages: an **offline phase** that analyzes a small exemplar set through a multi-agent pipeline to construct structured reflection notes encoding the model's characteristic errors, and an **online phase** that injects these notes into the generation context for new images, achieving improved factuality and coverage at negligible additional cost. Figure 2 illustrates the overall pipeline.

3.1 Offline Phase: Constructing Structured Reflection Notes

The goal of the offline phase is to surface the target LVLM's systematic error patterns and encode them as reusable guidance. Given a small exemplar set $\mathcal{D}_{\text{train}} = \{(x_i, y_i^*)\}_{i=1}^{M}$ of images paired with human-written reference captions, where x_i denotes the i-th input image, y_i^* its corresponding reference caption, and M the total number of exemplars, we run a three-agent pipeline that progressively moves from raw errors to generalizable directives.

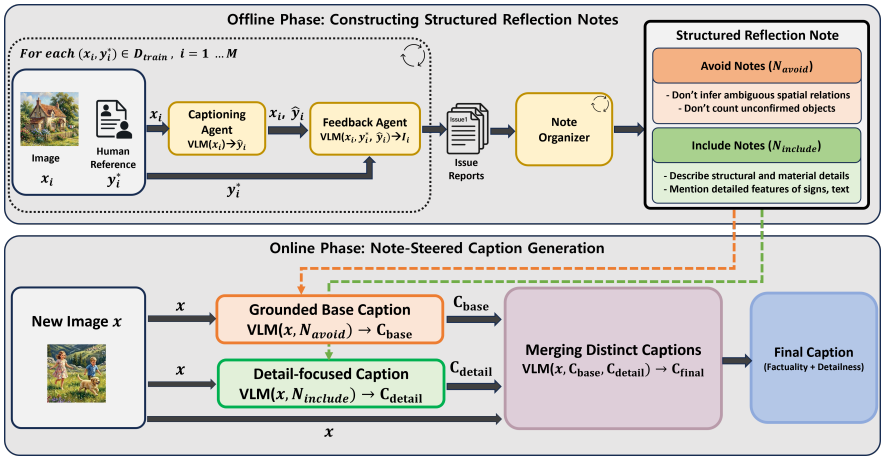

Fig. 2: ReflectCAP framework. In the **offline phase**, a multi-agent pipeline analyzes a small exemplar set to distill recurring errors and omissions of the target LVLM into *Structured Reflection Notes*. In the **online phase**, these notes guide caption generation: Avoid Notes suppress hallucinations, Include Notes encourage missing details, and a final merge integrates grounded and detail-focused captions into the final output.

Captioning Agent. The pipeline begins by letting the target LVLM caption each image x_i in a zero-shot manner, producing a candidate caption $\hat{y}_i$ with no additional guidance. This is intentional: the resulting captions faithfully reflect the model's default behavior—including its characteristic hallucinations and omissions—providing an unbiased basis for the diagnosis that follows.

Feedback Agent. Each candidate caption is then critiqued against two sources of evidence: the image itself and the human reference. The Feedback Agent cross-references $\hat{y}_i$ against both x_i and y_i^*, producing a structured issue report $\mathcal{I}_i$ for each example. Reports are organized into two categories:

- **Hallucinations**: details in $\hat{y}_i$ that are factually incorrect or not visible in x_i.
- **Missing Details**: important details present in y_i^* that are absent from $\hat{y}_i$.

For example, a hallucination might be *"the caption states two people are sitting, but the image shows three,"* while a missing detail might be *"the caption does not mention the wooden railing visible in the foreground."* The Feedback Agent has access to the image during critique, ensuring that its judgments are visually grounded. However, at this stage, every diagnosis is tied to a particular image and caption, making it difficult to apply directly at inference time.

Note Organizer. Instance-specific diagnoses help explain individual failures, but reusable guidance requires identifying *patterns*—errors that recur across images. The Note Organizer performs this consolidation. Because diagnoses collected across images can easily exceed the LVLM context window, the organizer processes them incrementally, consuming batches and updating a running set of

Algorithm 1 Offline: Constructing Structured Reflection Notes

Require: Training set $\mathcal{D}_{\text{train}} = \{(x_i, y_i^*)\}_{i=1}^{M}$, max items K, batch size B
$\mathcal{N} \leftarrow \emptyset$
for $i = 1$ to M **do**
  $\hat{y}_i \leftarrow \text{CaptioningAgent}(x_i)$ ▷ Zero-shot captioning
  $\mathcal{I}_i \leftarrow \text{FeedbackAgent}(x_i, \hat{y}_i, y_i^*)$ ▷ Instance-specific critique
end for
for each batch $\mathcal{B} \subset \{\mathcal{I}_1, \ldots, \mathcal{I}_M\}$ of size B **do**
  $\mathcal{N} \leftarrow \text{NoteOrganizer}(\mathcal{B}, \mathcal{N}, K)$ ▷ Cross-instance generalization
end for
return $\mathcal{N} = (\mathcal{N}_{\text{avoid}}, \mathcal{N}_{\text{include}})$

Algorithm 2 Online: Note-Steered Caption Generation

Require: Test image x, Structured Reflection Notes $\mathcal{N} = (\mathcal{N}_{\text{avoid}}, \mathcal{N}_{\text{include}})$
Step 1: Grounded Base Caption
$c_{\text{base}} \leftarrow \text{VLM}(x,\ \mathcal{N}_{\text{avoid}})$ ▷ Suppress known hallucination patterns
Step 2: Detail-Focused Caption
$c_{\text{detail}} \leftarrow \text{VLM}(x,\ \mathcal{N}_{\text{include}})$ ▷ Attend to typically neglected details
Step 3: Merging Distinct Captions
$c_{\text{final}} \leftarrow \text{VLM}(x,\ c_{\text{base}},\ c_{\text{detail}})$ ▷ Merge with image as reference
return c_{final}

notes after each step. During each update, it merges semantically similar issues and abstracts them into broadly applicable rules. By imposing an upper bound of K items, the note set retains only patterns corresponding to frequently recurring mistakes or commonly omitted details, thereby pruning redundant or overly narrow entries. The result is a compact, prioritized set of notes that we call **Structured Reflection Notes**. It consists of two complementary components:

- **Avoid Notes** $\mathcal{N}_{\text{avoid}}$: directives that suppress recurrent hallucination patterns (e.g., *"Do not infer object colors when they are ambiguous"*).
- **Include Notes** $\mathcal{N}_{\text{include}}$: directives that enforce frequently omitted details (e.g., *"Describe visible architectural details such as structural supports and railings"*).

This progression from instance-level diagnosis to cross-instance generalization is what allows the notes to capture the model's systematic tendencies rather than one-off mistakes. In practice, M=30 exemplar images and K=5 items per category are sufficient, as shown in our ablation study (§5.3). Algorithm 1 summarizes the full offline procedure.

3.2 Online Phase: Note-Steered Caption Generation

Once constructed, the Structured Reflection Notes replace the multi-agent pipeline entirely. Given a new image x and the pre-computed notes $\mathcal{N} = (\mathcal{N}_{\text{avoid}}, \mathcal{N}_{\text{include}})$, the online phase generates a caption through at most three LVLM calls.

Step 1: Grounded Base Caption. $\mathcal{N}_{\text{avoid}}$ is injected into the captioning prompt, directing the LVLM to suppress its known hallucination patterns during generation. This produces a grounded base caption c_{base} that is more factually reliable than a zero-shot caption while preserving the model's natural descriptive ability. Since this step requires exactly one forward pass, almost identical in cost to zero-shot inference, it can serve as a standalone variant for captioning pipelines where factuality is the primary concern.
Step 2: Detail-Focused Caption. A second call uses $\mathcal{N}_{\text{include}}$ to direct the LVLM's attention toward the specific types of details it typically misses, such as material textures, background elements, and spatial arrangements, producing a detail-focused caption c_{detail}. This caption captures the descriptive details that c_{base} trades off in favor of factual grounding.
Step 3: Merging Distinct Captions. The final stage merges c_{base} and c_{detail} into a unified caption c_{final}, using the image as a grounding reference. To prevent the integration of spurious details, the merge strategy is conservative: c_{base} is treated as the primary source of truth, while c_{detail} serves as a supplementary source. In cases of conflict, the model prioritizes c_{base}, as it is generated under hallucination-suppressing guidance designed for factual grounding. We refer to this full three-step pipeline as **ReflectCAP**. The complete procedure is summarized in Algorithm 2.

4 Experiments

We evaluate ReflectCAP along three dimensions. (1) **Fine-grained evaluation**: fine-grained factuality and coverage, which examines the trade-off between these two axes, (2) **Holistic evaluation**: holistic caption quality, which tests whether fine-grained gains translate into perceived overall quality, and (3) **Cost-efficiency**: computational cost analysis which examines whether the method is practical enough for downstream deployment.

4.1 Experimental Settings

Evaluation Metrics. We adopt two complementary evaluation suites, both well-aligned with human judgments. For fine-grained evaluation, we evaluate on IIW-400 dataset [8] using the factuality and coverage metrics proposed by Lee et al. [15]. *Factuality* (Precision) decomposes each caption into atomic propositions and verifies each against the image and ground-truth; *Coverage* (Recall) is measured via curated VQA items associated with each IIW-400 image, answered using only the generated caption. We report *F1 score* as the harmonic mean of factuality and coverage. For holistic evaluation, we adopt CapArena-Auto [3], a pairwise benchmark scored by average win-rate margin against three reference models.
Baseline LVLMs. We evaluate across eight LVLMs spanning closed-source and open-source families at a range of scales: two proprietary models (GPT-4.1-mini, GPT-4.1-nano) and six open-weight instruction-tuned models, comprising three

Table 1: Factuality and Coverage on IIW-400. Precision measures the ratio of verified-true propositions. Recall measures the ratio of correctly answered VQA questions. F1 is the harmonic mean. Best per model in **bold**. Δ denotes F1 change from Zero-shot.

Model	Method	P	R	F1	Δ	Model	Method	P	R	F1	Δ
GPT-4.1-mini	Zero-shot	83.6	68.1	75.1	—	Qwen2.5-7B	Zero-shot	68.3	57.9	62.7	—
	Few-shot	81.9	71.5	76.3	+1.2		Few-shot	49.9	56.6	53.1	−9.6
	Self-Corr.	82.6	69.3	75.4	+0.3		Self-Corr.	63.8	57.3	60.4	−2.3
	CapMAS	**84.2**	67.7	75.1	0.0		CapMAS	**72.0**	57.4	63.9	+1.2
	ReflectCAP	83.8	**72.0**	**77.5**	**+2.4**		**ReflectCAP**	68.8	**62.3**	**65.4**	**+2.7**
GPT-4.1-nano	Zero-shot	78.9	62.6	69.8	—	Qwen2.5-32B	Zero-shot	71.9	64.0	67.7	—
	Few-shot	77.1	67.1	71.8	+2.0		Few-shot	64.9	64.5	64.7	−3.0
	Self-Corr.	79.2	62.6	70.0	+0.2		Self-Corr.	70.2	66.2	**68.1**	**+0.4**
	CapMAS	**83.1**	57.2	67.8	−2.0		CapMAS	**73.5**	63.1	67.9	+0.2
	ReflectCAP	78.0	**67.2**	**72.2**	**+2.4**		**ReflectCAP**	69.9	**66.4**	**68.1**	**+0.4**
InternVL-4B	Zero-shot	64.1	54.9	59.1	—	Qwen3-8B	Zero-shot	76.3	69.2	72.6	—
	Few-shot	49.4	52.8	51.1	−8.0		Few-shot	71.3	**72.2**	71.8	−0.8
	Self-Corr.	62.1	54.1	57.8	−1.3		Self-Corr.	76.3	70.0	73.0	+0.4
	CapMAS	**72.2**	53.3	61.3	+2.2		CapMAS	**79.3**	68.8	73.7	+1.1
	ReflectCAP	64.8	**61.3**	**63.0**	**+3.9**		**ReflectCAP**	77.0	71.2	**74.0**	**+1.4**
InternVL-38B	Zero-shot	72.5	57.8	64.3	—	Qwen3-32B	Zero-shot	79.2	72.5	75.7	—
	Few-shot	63.2	62.1	62.6	−1.7		Few-shot	73.4	**74.4**	73.9	−1.8
	Self-Corr.	73.7	58.2	65.0	+0.7		Self-Corr.	78.6	73.6	76.0	+0.3
	CapMAS	**78.6**	57.5	66.4	+2.1		CapMAS	**81.0**	72.0	76.2	+0.5
	ReflectCAP	73.7	**64.2**	**68.6**	**+4.3**		**ReflectCAP**	79.2	73.8	**76.4**	**+0.7**

smaller models (InternVL3.5-4B-Instruct, Qwen2.5-VL-7B-Instruct, Qwen3-VL-8B-Instruct) and three larger models (InternVL3.5-38B-Instruct, Qwen2.5-VL-32B-Instruct, Qwen3-VL-32B-Instruct).

Baseline Methods. For each model, we compare ReflectCAP against four captioning strategies. *Zero-shot* uses a minimal prompt ("Describe this image in detail"). *Few-shot* prepends three randomly sampled human-annotated caption exemplars. *Self-Correction* first generates a zero-shot caption, then revises it after re-examining the image. *CapMAS* [15], a multi-agent baseline, decomposes a caption into atomic propositions via specialized agents, verifies each against the image, and rewrites the caption by removing unverified content. For fair comparison, ReflectCAP uses the target model itself for both the offline phase (note construction) and the online phase (caption generation), ensuring that no external model contributes to the final output. In the offline phase, we use images and reference captions from IIW-Eval that are not included in IIW-400 to construct the notes.

4.2 Fine-Grained Evaluation

Table 1 presents our main results. Existing methods tend to improve one axis at the cost of the other: few-shot prompting increases coverage by imitating human-

Table 2: CapArena-Auto scores. Score denotes the average win-rate margin (higher is better; range $[-100, 100]$). For CapMAS and ReflectCAP, we report *score* with the change relative to zero-shot in parentheses. Rows marked with † use zero-shot reference values taken from the provided CapArena caption.

Model	Zero-shot	CapMAS	ReflectCAP (Ours)
Leaderboard anchors (zero-shot only)			
GPT-5.2	70.0	—	—
Gemini-1.5-Pro†	62.3	—	—
GPT-4o-0806†	44.3	—	—
Qwen2.5VL-72B†	39.7	—	—
Claude-3.5-Sonnet†	29.7	—	—
GPT-4.1-mini	57.7	53.7 (−4.0)	**90.0** (+32.3)
GPT-4.1-nano	21.2	−14.7 (−35.9)	**51.3** (+30.1)
InternVL3.5-4B	−54.3	−46.7 (+7.6)	**−18.7** (+35.6)
InternVL3.5-38B	−24.0	−15.0 (+9.0)	**12.0** (+36.0)
Qwen2.5-VL-7B	−33.7	−28.0 (+5.7)	**−7.3** (+26.4)
Qwen2.5-VL-32B	5.7	8.0 (+2.3)	**25.3** (+19.6)
Qwen3-VL-8B	76.0	74.0 (−2.0)	**85.3** (+9.3)
Qwen3-VL-32B	87.3	87.0 (−0.3)	**91.7** (+4.4)

authored demonstrations but pushes the model beyond its perceptual boundary, causing factuality to drop. Self-correction yields only marginal gains regardless of model scale, indicating that models struggle to identify and fix errors in their own generated captions through revision alone. CapMAS that improves factuality by removing unreliable content from existing captions inevitably sacrifices coverage in the process. In contrast, ReflectCAP substantially boosts coverage while incurring minimal loss in factuality, achieving the highest F_1 across all eight models. This demonstrates that the Structured Reflection Notes effectively balance the inherent trade-off: coverage guidance encourages the model to describe more, which inevitably risks lowering factuality, while factuality guidance separately constrains this degradation. By controlling each objective through its own dedicated guidance, ReflectCAP achieves the highest F_1 across all models, advancing the Pareto frontier between the two objectives.

4.3 Holistic Evaluation

Fine-grained metrics measure factuality and coverage in isolation, but it is also necessary to evaluate overall caption quality. CapArena-Auto (Table 2) tests this by pitting each method's captions against fixed three reference models[4] in head-to-head comparisons judged by GPT-4.1-mini.

ReflectCAP improves the average win-rate margin by +33.2 points for the GPT family and +21.9 points for open-source models over their zero-shot baselines, confirming that the fine-grained gains in §4.2 translate directly into holistic

[4] GPT-4o-0806, CogVLM2-llama3-chat-19B, and MiniCPM-V2.6-8B.

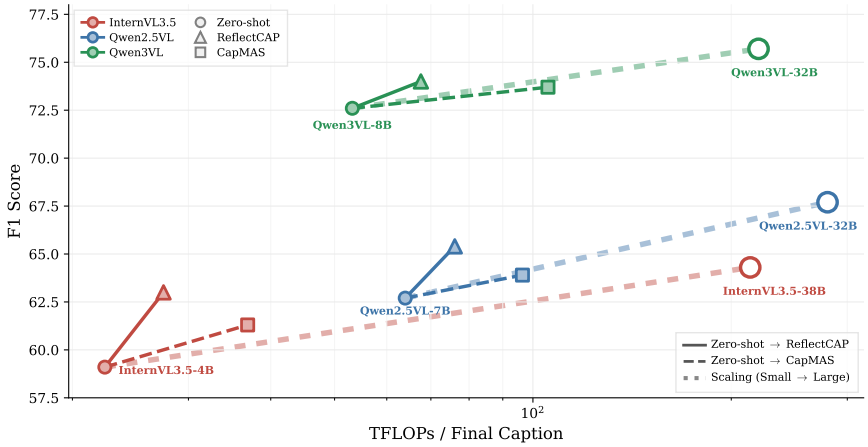

Fig. 3: Solid and **dash-dotted lines** denote improvements from zero-shot to ReflectCAP and CapMAS, respectively. ReflectCAP achieves higher F1 scores while requiring 21–36% less compute than CapMAS. **Light dashed lines** denote performance gains from model parameter scaling. Compared to simply increasing model size, ReflectCAP achieves comparable quality at up to 8× lower compute cost, enabling high-quality, detailed captioning more practical under real-world cost and latency constraints.

caption quality. Notably, GPT-4.1-mini with ReflectCAP (90.0) surpasses even GPT-5.2 (70.0), suggesting that structured reflection notes can compensate for inherent model capacity differences in overall caption quality. Additionally, ReflectCAP yields consistent positive gains over the zero-shot baseline across all models, whereas CapMAS tends to degrade performance when applied to models that already exhibit strong zero-shot capabilities.

4.4 Cost-Efficiency

Beyond caption quality, practical deployment requires efficient inference. To examine this, we measure cost-efficiency across methods and open-source models in terms of the total TFLOPs required to produce a final caption at inference time.[5]

Figure 3 plots F_1 against inference TFLOPs per final caption for all open-source models. Two trends emerge. First, ReflectCAP improves caption quality more compute-efficiently than simply scaling model size. For example, InternVL3.5-4B with ReflectCAP achieves an F_1 of 63.0, approaching InternVL3.5-38B zero-shot (64.3) while requiring roughly 7.8× fewer TFLOPs (27.5 vs. 213.7). Similarly, Qwen3-VL-8B with ReflectCAP maintains a comparable performance gap relative to Qwen3-VL-32B zero-shot, while generating captions at approximately

[5] We approximate inference cost as $C \approx 2NT$, where N is the number of non-embedding parameters and T is the total token count. For methods with multiple calls per image, image tokens are counted only once via KV caching.

Table 3: Impact of SFT on detailed captioning factuality.

Model	Method	Fact.	Cov.	F1
InternVL3.5-4B	Zero-shot	64.1	54.9	59.1
	SFT w/ Human-authored	58.2	56.5	57.4
	SFT w/ ReflectCAP-generated	**66.5**	**61.6**	**63.9**
Qwen2.5-VL-7B	Zero-shot	68.3	57.9	62.7
	SFT w/ Human-authored	57.1	58.0	57.6
	SFT w/ ReflectCAP-generated	**69.9**	**63.5**	**66.5**

3.3× lower computational cost. Second, ReflectCAP is also more compute-efficient than existing inference-time baselines within the same model. Across three models— InternVL3.5-4B, Qwen2.5-VL-7B, and Qwen3-VL-8B—ReflectCAP reduces inference cost by 21%–36% TFLOPs compared to CapMAS, while consistently achieving higher F_1 scores. This inefficiency stems from CapMAS applying its multi-agent pipeline directly at inference time, and its focus on factuality alone limits overall caption quality. These results highlight ReflectCAP as a practical alternative to both model scaling and compute-heavy inference-time pipelines.

5 Analysis

5.1 Supervised Fine-Tuning on Detailed Image Captioning

The most intuitive approach to improving detailed captioning is supervised fine-tuning (SFT) on human-authored detailed captions. To examine this, we fine-tune InternVL3.5-4B and Qwen2.5-VL-7B on 9,647 DOCCI human-annotated captions using LoRA. As shown in Table 3, SFT substantially degrades factuality compared to the zero-shot baseline for both models (evaluation follows the same protocol as Section 4), lending further support to recent findings that training on annotations exceeding the model's visual capabilities amplifies hallucinations [36, 37]. An interesting finding is that replacing human captions with ReflectCAP-generated captions on the same DOCCI images for training maintains factuality while improving coverage. This suggests that ReflectCAP can serve as a scalable pipeline for constructing training data that improves recall while staying within the model's visual boundary, without manual annotation effort.

5.2 Generalization with Frozen Reflection Notes

We further examine whether the reflection notes learned from the original benchmark transfer to external detailed-captioning benchmarks. Specifically, we evaluate ReflectCAP on CaptionQA [35] and CAPability [21] using the same frozen IIW-derived notes, without re-mining notes for each target benchmark. CaptionQA measures downstream caption utility across four domains, while CAPability evaluates caption correctness and thoroughness across multiple captioning

Table 4: Non-homologous generalization. CaptionQA accuracy (%; Nat./Doc./E-com/Emb.=natural/document/e-commerce/embodied); CAPability reports P=precision, H=hit, Avg=(P+H)/2 over 9 dims.

		CaptionQA					**CAPability**		
Backbone	Method	Nat.	Doc.	E-com	Emb.	Ovr.	P	H	Avg
GPT-4.1-mini	ZS	70.8	81.3	83.4	64.1	74.5 (-)	**85.2**	51.9	68.6 (-)
	CapMAS	70.8	78.7	82.6	63.6	73.6 (-0.9)	83.9	49.9	66.9 (-1.7)
	Ours	**81.9**	**84.7**	**88.4**	**72.1**	**81.8 (+7.3)**	83.8	**58.2**	**71.0 (+2.4)**
Qwen3-VL-8B	ZS	76.1	68.7	85.0	68.3	74.9 (-)	80.9	58.1	69.5 (-)
	CapMAS	75.2	69.3	83.9	66.1	74.0 (-0.9)	82.7	57.7	70.2 (+0.7)
	Ours	**80.6**	**74.7**	**86.1**	**68.9**	**78.0 (+3.1)**	**84.2**	**59.7**	**71.9 (+2.4)**

dimensions. For evaluation, we sample 50 images per domain from CaptionQA (4 domains, 200 images total) and 100 images per dimension from CAPability (9 dimensions, 900 images total) using seed 42.

Table 4 shows that ReflectCAP consistently improves over zero-shot and CapMAS on CaptionQA. On GPT-4.1-mini, ReflectCAP improves the overall CaptionQA accuracy from 74.5 to 81.8, and on Qwen3-VL-8B, it improves the score from 74.9 to 78.0. ReflectCAP also improves the average CAPability score on both backbones. These results suggest that the reflection notes capture reusable model-level error tendencies rather than benchmark-specific artifacts.

5.3 Ablation Study

Effect of Grounded Base Caption. As shown in Figure 4, the Grounded Base Caption, which applies only the hallucination-suppression notes, improves factuality over the zero-shot baseline in nearly all models. However, the magnitude of this improvement varies with the model's instruction-following capability. Models with stronger instruction-following abilities, such as GPT-4.1-mini and GPT-4.1-nano, generate well-grounded base captions where hallucination patterns are effectively suppressed, whereas the Qwen2.5-VL family shows marginal improvement or even degradation compared to the zero-shot baseline. This suggests that as instruction-following capabilities of LVLMs continue to advance, ReflectCAP can achieve even greater improvements without any modification to the framework.

Separate vs. Combined Injection. Table 5 compares two strategies on 100 images sampled from IIW-400: applying Avoid and Include notes in separate generation passes versus injecting both into a single prompt. Across all four models, the separated approach consistently outperforms the combined variant. This effect is particularly pronounced in InternVL3.5-4B, where the combined injection causes a severe F1 drop (64.4 $\rightarrow$ 57.3), indicating that overloading the prompt with too many directives can be detrimental, especially for models with limited instruction-following capability. These results confirm that hallucination suppression and detail recovery are better handled as separate objectives.

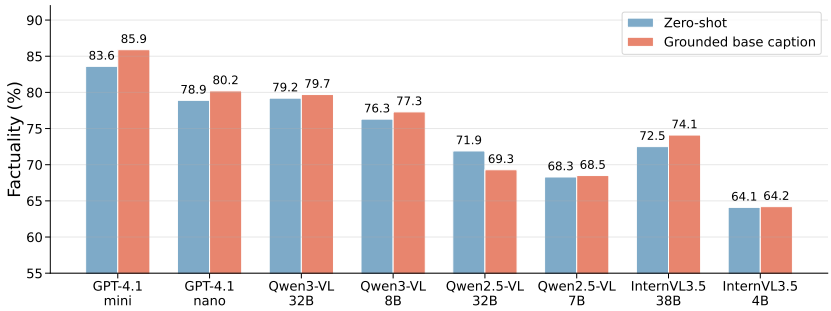

Fig. 4: Factuality comparison between Zero-shot and Grounded Base Caption across all models. Models with stronger instruction-following capabilities show larger gains.

Table 5: Separate vs. Combined note injection. Separate-Merge applies Avoid and Include notes in separate passes with merging; Combined injects both into a single prompt.

Model	Method	Fact.	Cov.	F1
GPT-4.1-mini	Separate-Merge	84.8	72.8	**78.3**
	Combined	84.8	70.7	77.1
GPT-4.1-nano	Separate-Merge	78.1	68.6	**73.1**
	Combined	78.4	67.0	72.2
Qwen3-VL-8B	Separate-Merge	77.6	72.4	**74.9**
	Combined	74.4	71.2	72.7
InternVL3.5-4B	Separate-Merge	66.6	62.4	**64.4**
	Combined	55.8	58.8	57.3

Number of Exemplars and Note Items. Figure 5 analyzes two key parameters of the offline phase using 100 images sampled from IIW-400 for efficient evaluation.

For the number of exemplar images N (Figure 5(a)), all models show a substantial improvement from zero-shot to $N = 10$, with performance largely plateauing around $N = 30$ and slightly declining at $N = 100$. Qualitative analysis of GPT-4.1-mini suggests that larger exemplar pools shift reflection notes from model-specific guidance toward generic instructions. For example, notes generated with $N = 30$ contain targeted rules such as *"Do not add unsupported details to signs, logos, or symbols,"* whereas those from $N = 100$ become broader directives like *"Avoid subjective or interpretive descriptions not clearly supported by the image or reference."* This indicates that systematic error patterns can be reliably surfaced from a modest exemplar set, while larger pools introduce variation that dilutes the corrective signal.

For the maximum items per category K (Figure 5(b)), even K=1 provides a substantial performance gain. Under this setting, models tend to produce a single reflection note that aggregates multiple corrective signals rather than a narrowly scoped rule. For example, a reflection note generated at K=1 states: *"Include*

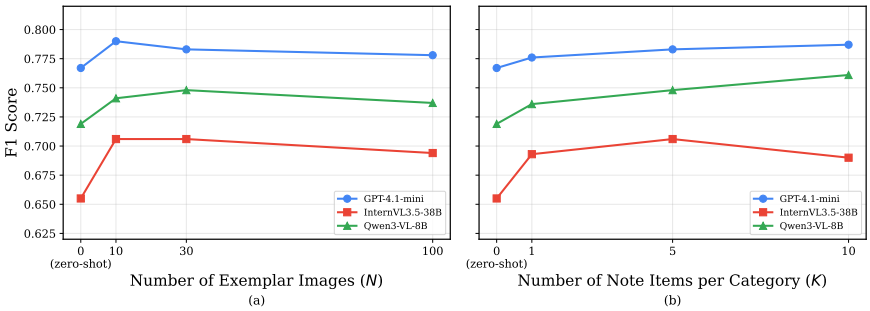

Fig. 5: Ablation on note construction parameters. **(a)** F1 vs. the number of exemplar images N. Performance saturates at $N{\approx}30$, indicating that systematic error patterns can be surfaced from a modest exemplar set. **(b)** F1 vs. the maximum number of note items per category K. Even $K{=}1$ already yields strong gains, with performance improving slightly further at $K{=}5$.

precise visible details of object features, spatial distributions, lighting and shadow effects, background elements, and compositional angles to ensure completeness and accuracy." As K increases (e.g., $K{=}5$), these aggregated instructions are decomposed into multiple more specialized reflection notes that introduce more specific guidelines, which further improves performance over the $K{=}1$ setting. While some models continue improving up to $K{=}10$, others peak around $K{=}5$ and slightly decline thereafter, suggesting that the optimal number of guidelines varies across models and may depend on their ability to incorporate multiple instructions.

5.4 Case Study

Figure 6 illustrates a case study examining each pipeline stage of ReflectCAP using GPT-4.1 mini. In the zero-shot caption, hallucinations are observed in fine-grained details such as roof shape, window count and arrangement, and sign appearance. By applying our hallucination avoidance patterns, ReflectCAP corrects the window count from six to five and omits unverifiable arrangements, yielding a more factually grounded description. Furthermore, the missing detail recovery patterns enable the model to capture previously overlooked visual elements, including the lion's head fountain spout and cast shadows. These results demonstrate that ReflectCAP effectively encodes common error patterns into structured reflection notes and leverages them to steer caption generation toward greater accuracy and visual fidelity.

6 Conclusion

We presented ReflectCAP, a tuning-free framework that distills a target LVLM's recurring hallucination and omission patterns into Structured Reflection Notes.

1) Zeroshot Caption
The image shows a small, single-story building with a beige exterior and a red-tiled triangular roof. The building has six small, white-framed windows on the visible side, arranged in two rows of three. In front of the building, there is an overgrown grassy patch with various wild plants and weeds. On the right side of the building, there is a paved sidewalk with a street sign that has two signs, one triangular and one square, both facing away from the camera. The sidewalk curves around the corner, where multiple parked cars are visible along the street. The street has a designated bike lane marked with white paint. ...

2) ReflectCAP-base (Hallucination Avoidance)
The image shows a street scene with a small, single-story building situated in the center. The building has a yellowish-beige facade and a red-tiled roof with some visible dark patches. There are five small windows evenly spaced along the building's visible side. In front of the building, there is an area of overgrown grass and wild plants, with a large ceramic pot and a stone fountain structure attached to the building wall. There are large green trees surrounding the area, with one prominent tree on the left side of the building and another on the right near the sidewalk. ...

Error note (Hallucination Patterns)	Hallucination → Correction	Corrected Description
"Do not add or alter visible object features, colors, or counts without confirmation."	triangular roof → no triangular roof	red-tiled roof with some visible dark patches
	six small windows → five small windows	five small windows
	in two rows of three → one row of five	No arrangement mentioned

3) ReflectCAP (Missing Detail Recovery)
The image shows a street scene with a small, single-story building situated in the center. The building has a yellowish-beige facade and a red-tiled roof with some visible dark patches. There are five small windows evenly spaced along the building's visible side. In front of the building, there is an area of overgrown grass and wild plants, with a large ceramic pot and a stone fountain structure attached to the building wall, featuring a lion's head spout. ... The pole casts a sharp shadow on the sidewalk, ... There are large green trees surrounding the area, including a large leafy tree on the left side of the building and coniferous trees behind it, as well as another tree on the right near the sidewalk. ...

Error note (Missing Detail Patterns)	Grounded base caption (omitted)	Recovery details
"Include precise object condition, orientation, posture, and spatial positioning as visible."	No mentioned about fountain feature	fountain structure attached to the building wall, featuring a lion's head spout
"Note lighting effects, shadows, and surface details"	No mentioned about lighting effects and shadows	The pole casts a sharp shadow on the sidewalk
"Describe all visible background elements, environmental context"	No mentioned about trees behind the building	coniferous trees

Fig. 6: Case study of our pipeline. **Top**: Zero-shot Caption **Middle**: ReflectCAP-Base suppresses hallucinations via Avoid notes. **Bottom**: ReflectCAP-Full recovers embossed text details guided by Include notes. Red denotes hallucinated expressions, blue denotes hallucination-corrected descriptions, and green denotes recovered fine-grained details.

By leveraging these notes at caption generation time, ReflectCAP steers the model separately for each pattern type—suppressing hallucinations and recovering missing details—then merges the resulting captions into a single, comprehensive description. Across 8 LVLMs, ReflectCAP consistently achieves the highest F1 score on factuality–coverage evaluation, and substantially outperforms all baselines on CapArena-Auto. Furthermore, from a compute-efficiency perspective, ReflectCAP is more effective than both scaling up model size and scaling inference-time computation for improving caption quality.

Acknowledgements

This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (RS-2024-00348233), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] & RS-2021-II212068, Artificial Intelligence Innovation Hub (Artificial Intelligence Institute, Seoul National University)], and the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2024. K. Jung is with ASRI, Seoul National University, Korea.

References

1. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023)
2. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024)
3. Cheng, K., Song, W., Fan, J., Ma, Z., Sun, Q., Xu, F., Yan, C., Chen, N., Zhang, J., Chen, J.: CapArena: Benchmarking and analyzing detailed image captioning in the LLM era. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics: ACL 2025. pp. 14077–14094. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.findings-acl.724, https://aclanthology.org/2025.findings-acl.724/
4. Chung, J., Kim, J., Kim, S., Lee, J., Kim, M.S., Yu, Y.: v1: Learning to point visual tokens for multimodal grounded reasoning (2026), https://arxiv.org/abs/2505.18842
5. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
6. Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., Soatto, S.: Multi-modal hallucination control by visual information grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14303–14312 (2024)
7. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024)
8. Garg, R., Burns, A., Karagol Ayan, B., Bitton, Y., Montgomery, C., Onoe, Y., Bunner, A., Krishna, R., Baldridge, J.M., Soricut, R.: ImageInWords: Unlocking hyper-detailed image descriptions. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 93–127. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.6, https://aclanthology.org/2024.emnlp-main.6/
9. Gutflaish, E., Kachlon, E., Zisman, H., Hacham, T., Sarid, N., Visheratin, A., Huberman, S., Davidi, G., Bukchin, G., Goldberg, K., et al.: Generating an image from 1,000 words: Enhancing text-to-image with structured captions. arXiv preprint arXiv:2511.06876 (2025)
10. He, J., Lin, H., Wang, Q., Fung, Y.R., Ji, H.: Self-correction is more than refinement: A learning framework for visual and language reasoning tasks. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 6405–6421 (2025)
11. 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)
12. Ju, X., Gao, Y., Zhang, Z., Yuan, Z., Wang, X., Zeng, A., Xiong, Y., Xu, Q., Shan, Y.: Miradata: A large-scale video dataset with long durations and structured captions. Advances in Neural Information Processing Systems **37**, 48955–48970 (2024)

13. Kamoi, R., Zhang, Y., Zhang, N., Han, J., Zhang, R.: When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. Transactions of the Association for Computational Linguistics **12**, 1417–1440 (2024). https://doi.org/10.1162/tacl_a_00713, https://aclanthology.org/2024.tacl-1.78/
14. Lee, K.i., Kim, M., Yoon, S., Kim, M., Lee, D., Koh, H., Jung, K.: VLind-bench: Measuring language priors in large vision-language models. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Findings of the Association for Computational Linguistics: NAACL 2025. pp. 4129–4144. Association for Computational Linguistics, Albuquerque, New Mexico (Apr 2025). https://doi.org/10.18653/v1/2025.findings-naacl.231, https://aclanthology.org/2025.findings-naacl.231/
15. Lee, S., Yoon, S., Bui, T., Shi, J., Yoon, S.: Toward robust hyper-detailed image captioning: A multiagent approach and dual evaluation metrics for factuality and coverage. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=REnIf3dCsI
16. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13872–13882 (2024)
17. Li, Z., Shi, H., Gao, Y., Liu, D., Wang, Z., Chen, Y., Liu, T., Zhao, L., Wang, H., Metaxas, D.N.: The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=7BKcLeHQsm
18. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. arXiv preprint arXiv:2306.14565 (2023)
19. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
20. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
21. Liu, Z., Xie, C.W., Wen, B., Yu, F., Li, P., Zhang, B., Yang, N., Gao, Z., Zheng, Y., Xie, H.: Capability: A comprehensive visual caption benchmark for evaluating both correctness and thoroughness. Advances in Neural Information Processing Systems **38** (2026)
22. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B.P., Hermann, K., Welleck, S., Yazdanbakhsh, A., Clark, P.: Self-refine: Iterative refinement with self-feedback (2023), https://arxiv.org/abs/2303.17651
23. Marsili, D., Mehta, A., Lin, R.Y., Gkioxari, G.: Same or not? enhancing visual perception in vision-language models. arXiv preprint arXiv:2512.23592 (2025)
24. Merchant, N., de Ocáriz Borde, H.S., Popescu, A.C., Suarez, C.G.J.: Structured captions improve prompt adherence in text-to-image models (re-laion-caption 19m) (2025), https://arxiv.org/abs/2507.05300
25. Min, K., Kim, M., Lee, K.i., Lee, D., Jung, K.: Mitigating hallucinations in large vision-language models via summary-guided decoding. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Findings of the Association for Computational Linguistics: NAACL 2025. pp. 4183–4198. Association for Computational Linguistics, Albuquerque, New Mexico (Apr 2025). https://doi.org/10.18653/v1/2025.findings-naacl.235, https://aclanthology.org/2025.findings-naacl.235/

26. Onoe, Y., Rane, S., Berger, Z., Bitton, Y., Cho, J., Garg, R., Ku, A., Parekh, Z., Pont-Tuset, J., Tanzer, G., Wang, S., Baldridge, J.: Docci: Descriptions of connected and contrasting images (2024), https://arxiv.org/abs/2404.19753
27. Ouyang, S., Yan, J., Hsu, I.H., Chen, Y., Jiang, K., Wang, Z., Han, R., Le, L.T., Daruki, S., Tang, X., Tirumalashetty, V., Lee, G., Rofouei, M., Lin, H., Han, J., Lee, C.Y., Pfister, T.: Reasoningbank: Scaling agent self-evolving with reasoning memory (2025), https://arxiv.org/abs/2509.25140
28. Rahmanzadehgervi, P., Bolton, L., Taesiri, M.R., Nguyen, A.T.: Vision language models are blind: Failing to translate detailed visual features into words. arXiv preprint arXiv:2407.06581 (2024)
29. 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)
30. Sun, H.L., Sun, Z., Peng, H., Ye, H.J.: Mitigating visual forgetting via take-along visual conditioning for multi-modal long CoT reasoning. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5158–5171. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.acl-long.257, https://aclanthology.org/2025.acl-long.257/
31. Tan, Z., Yan, J., Hsu, I.H., Han, R., Wang, Z., Le, L., Song, Y., Chen, Y., Palangi, H., Lee, G., et al.: In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8416–8439 (2025)
32. Urbanek, J., Bordes, F., Astolfi, P., Williamson, M., Sharma, V., Romero-Soriano, A.: A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions (2024), https://arxiv.org/abs/2312.08578
33. Wan, G., Ling, M., Ren, X., Han, R., Li, S., Zhang, Z.: Compass: Enhancing agent long-horizon reasoning with evolving context (2025), https://arxiv.org/abs/2510.08790
34. Wang, X., Pan, J., Ding, L., Biemann, C.: 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)
35. Yang, S., Liu, Y., Zhai, B., Sun, X., Liu, Z., Barsoum, E., Li, M., Xu, C.: Captionqa: Is your caption as useful as the image itself? (2026), https://arxiv.org/abs/2511.21025
36. Yanuka, M., Ben-Kish, A., Bitton, Y., Szpektor, I., Giryes, R.: Bridging the visual gap: Fine-tuning multimodal models with knowledge-adapted captions. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 10497–10518. Association for Computational Linguistics, Albuquerque, New Mexico (Apr 2025). https://doi.org/10.18653/v1/2025.naacl-long.527, https://aclanthology.org/2025.naacl-long.527/
37. Yue, Z., Zhang, L., Jin, Q.: Less is more: Mitigating multimodal hallucination from an EOS decision perspective. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 11766–11781. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.633, https://aclanthology.org/2024.acl-long.633/

38. Zhang, L., Zeng, X., Li, K., Yu, G., Chen, T.: Sc-captioner: Improving image captioning with self-correction by reinforcement learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23145–23155 (2025)
39. Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., Yao, H.: Analyzing and mitigating object hallucination in large vision-language models. arXiv preprint arXiv:2310.00754 (2023)
40. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
41. Zhu, L., Ji, D., Chen, T., Xu, P., Ye, J., Liu, J.: Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1624–1633 (2025)
42. Zhu, X., Cai, Y., Liu, Z., Zheng, B., Wang, C., Ye, R., Chen, J., Wang, H., Wang, W.C., Zhang, Y., et al.: Toward ultra-long-horizon agentic science: Cognitive accumulation for machine learning engineering. arXiv preprint arXiv:2601.10402 (2026)