Iterative Caption Refinement through Vision Language Feedback and Reasoning

Mahyar Ghazanfari
Department of Computer Science
University of Maryland, College Park
mghazanf@umd.edu

Kazem Faghih
Department of Computer Science
University of Maryland, College Park
kazemf@umd.edu

Zahra Sodagar
Department of Computer Science
University of Maryland, College Park
zsodagar@umd.edu

Abstract

Iterative refinement is a simple yet powerful mechanism for improving generated outputs without retraining the generator. In image captioning, this paradigm is commonly explored either when a single model both generates and critiques, or when the critic provides rich free-form feedback that may inadvertently leak the target description. We instead study a more constrained and controlled setting in which a text-only generator never observes the image and must rely on an object inventory together with minimal, structured feedback from a vision–language critic that is explicitly forbidden from quoting the candidate caption or providing a full correct caption. Within this framework, we investigate two complementary refinement strategies. In the first, the generator receives sparse visual hints from a vision–language model (VLM) critic in the form of non-revealing JSON feedback, guiding iterative caption updates. In the second, we replace the standard generator with models designed for explicit reasoning, using chain-of-thought prompting and reasoning-aligned language models to analyze why a given caption receives a particular score and to contrastively deduce more optimized variants. We implement a two-model loop that produces an initial COCO-style caption from extracted objects and iteratively refines it over multiple steps. Experiments on 1,000 MS COCO validation images show consistent improvements in BLEU, ROUGE-L, METEOR, and SPICE relative to the initial caption, with the largest gains occurring in early refinement steps followed by saturation or mild decay. We further analyze win rates against the initial and best-so-far captions, demonstrating that simple selection strategies can yield additional gains. Finally, we discuss evaluation pitfalls that can lead to misleading cross-paper comparisons, particularly for CIDEr and BLEU-4 scaling. The code for this project is publicly available at Iterative Caption Refinement.

1 Introduction

Large vision language models have made image captioning increasingly reliable by conditioning directly on pixels and generating fluent long range descriptions. However, many practical deployments do not allow a caption generator to access images at inference time. Privacy restrictions, latency constraints, or modular system design may require the generator to operate on intermediate signals such as detected objects, attributes, or compact descriptors produced by upstream vision modules.

This separation creates a gap between what the generator can express and what is actually present in the image, often leading to generic captions, omission of salient entities, or hallucinated details.

A natural way to address this mismatch is to treat captioning as an iterative process rather than a one shot prediction. Iterative refinement has a long history in language generation and has recently been explored in self improvement style frameworks, including MILS [1] like approaches, where candidates are repeatedly revised using feedback from a judge or critic. The promise of these methods is attractive: they can improve outputs without finetuning, are easy to plug into existing pipelines, and provide a mechanism to correct errors that are hard to avoid in a single decoding pass.

In this work, we adopt the high level idea of iterative refinement but place the system under strict constraints that are closer to a modular deployment setting. Our generator is text only and never observes the image. It is given only an object inventory extracted from the image and a sequence of minimal, structured hints produced by a vision language critic. The critic is required to avoid quoting the candidate caption and is forbidden from producing a full correct caption or a comprehensive scene description. These restrictions are intended to prevent the refinement process from collapsing into a trivial "copy the answer" behavior. Instead, the generator must use the object list and limited guidance to produce short, COCO style captions that remain faithful to the image.

Beyond this base refinement loop, we further investigate whether introducing explicit reasoning mechanisms into the text generator can improve the effectiveness of iterative caption refinement. Specifically, we experiment with two complementary reasoning-based setups. In the first, we apply chain-of-thought prompting to the generator, encouraging it to internally reason about how the provided object inventory and structured hints relate to the current caption and to potential failure modes indicated by the critic. In the second, we replace the base generator with a reasoning-aligned model that shares the same underlying architecture but has been distilled from DeepSeek-R1, a model trained to emphasize structured reasoning and deliberation. Importantly, in both cases, the reasoning process remains internal to the generator and does not grant access to the image or to richer feedback from the critic.

These reasoning-based variants allow us to contrast iterative refinement driven primarily by external guidance (via VLM hints) with refinement driven by stronger internal reasoning capabilities in the generator itself. By comparing standard prompting, chain-of-thought prompting, and a distilled reasoning model under otherwise identical conditions, we aim to better understand when reasoning helps, when it saturates, and how it interacts with minimal, non-revealing feedback.

We evaluate on 1000 random samples from MS COCO validation images across multiple refinement steps. The results show consistent improvements over the initial caption in BLEU, ROUGE L, METEOR, and SPICE. The performance curves indicate rapid gains in the first few steps followed by saturation, suggesting that the most common failure modes can be corrected quickly with minimal guidance. We further analyze the fraction of images that improve relative to step 0 and the best so far trajectory, demonstrating that selecting the best candidate across steps can yield additional gains. We conclude by discussing why naive cross paper comparisons can look inconsistent in practice, and focus on these inconsistencies.

2 Method

2.1 Overview

As demonstrated in Figure 1, our pipeline consists of two pretrained components: a vision–language model (VLM) that serves as an object inventory extractor and as a visual critic, and a text-only large language model (LLM) that generates and refines captions. Given an image, the VLM first outputs a list of salient visible objects. The text-only LLM then produces an initial COCO-style caption using only this object inventory. The refinement loop runs for a fixed number of steps. At each step, the critic examines the image and the current caption and returns a structured hint object. The generator rewrites the caption using the object list and these hints while adhering to strict style constraints.

This design intentionally differs from standard captioning pipelines that directly condition the generator on pixels. It also differs from many iterative self-improvement methods that allow the critic to provide rich free-form textual critiques. By restricting the critic to minimal, non-revealing

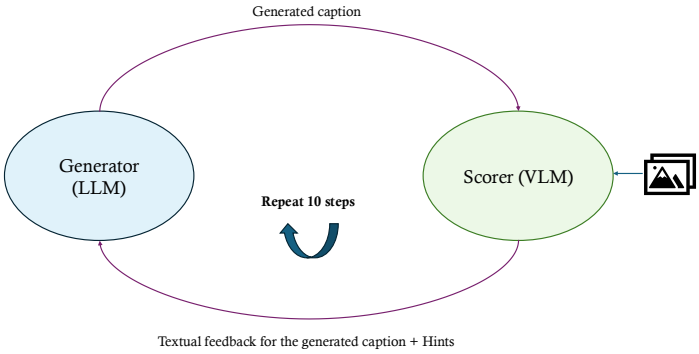

Figure 1: Hint-Only Iterative Caption Refinement Framework. An image is first processed by a vision–language model that extracts a sparse object inventory. A text-only large language model generates an initial caption based solely on this object list. At each iteration, the VLM acts as a critic and provides minimal, non-spoiler textual hints about missing, hallucinated, or overly generic elements without revealing the correct caption. The LLM then refines its caption using these hints. This generator–critic loop is repeated for a fixed number of steps, producing a sequence of progressively refined captions.

feedback, we aim to evaluate whether sparse corrective signals are sufficient to improve caption quality over multiple iterations.

In addition to this baseline setup, we investigate whether incorporating *reasoning* into the caption generator improves refinement behavior. Importantly, across all experiments, the visual critic, object inventory, refinement loop, and constraints remain fixed. The only modification lies in how the text-only generator processes the candidate captions and associated feedback signals.

2.2 Non-revealing structured feedback

The critic is instructed to follow two hard constraints: it must never quote the candidate caption, and it must never produce a full correct caption or a complete description of the scene. Instead, it outputs JSON fields that capture the kinds of errors most relevant to caption refinement. Concretely, it identifies objects that appear in the image but are missing from the caption, objects mentioned by the caption that are not supported by the image, and a short genericity hint indicating whether the caption is overly vague. It can also provide brief edit instructions phrased as actions (e.g., "remove an unsupported object" or "mention the main subject") and returns a coarse faithfulness score on a 0–10 scale.

This representation is intentionally minimal rather than exhaustive. It exposes a small set of corrective levers that the generator can act upon. The faithfulness score is used solely for analysis and is never optimized against during inference.

2.3 Caption generation and refinement constraints

The text-only generator is required to output exactly one short sentence in a COCO-like style. To reduce uncontrolled drift across refinement steps, we constrain captions to a narrow length range and discourage enumerations. The generator must mention only a small number of salient objects and is forbidden from introducing entities that are not present in the extracted object inventory. During refinement, it is instructed to remove hallucinated objects and optionally add missing objects when they are important, while preserving fluency and brevity.

These constraints are important for evaluation. Short captions can yield lower BLEU-4 and CIDEr scores than longer descriptions even when they are faithful, because higher-order n-gram overlap is harder to achieve with fewer tokens. This makes the setting challenging and highlights the importance of comparing methods under consistent generation regimes.

2.4 Model configuration and refinement procedure

Our refinement framework instantiates a two-model setup consisting of a vision–language model used as a visual critic and a text-only large language model used as the caption generator. The critic is `Qwen2-VL-7B-Instruct`, which processes the image together with textual prompts and produces structured feedback. The caption generator is `Qwen2.5-7B-Instruct`, which operates purely on text. Both models have approximately seven billion parameters and are used without any additional fine-tuning.

For each image, the VLM is first prompted to produce a compact object inventory containing the most salient visible entities. This object list serves as a grounding constraint for the entire refinement loop and remains fixed across steps. The LLM then generates an initial caption conditioned only on this object inventory. At each subsequent step, the VLM acts as a non-spoiling critic: given the image, the object list, and the current caption, it outputs structured JSON feedback containing missing objects, hallucinated objects, stylistic genericity hints, high-level edit instructions, and a scalar faithfulness score. The critic is explicitly instructed not to reveal or restate the correct caption.

The LLM consumes this feedback together with the object inventory and the previous caption to generate a revised caption. Prompts enforce COCO-style constraints, including a single-sentence format, a fixed word-length range, and prohibitions against enumerations or introducing unseen objects. This interaction is repeated for a fixed number of iterations, producing a sequence of candidate captions per image. Caption generation uses nucleus sampling with a low temperature to balance diversity and stability, while the VLM feedback is generated deterministically.

2.5 Reasoning-based caption refinement

Beyond the baseline generator configuration, we explore whether explicitly encouraging *reasoning* in the text-only generator improves caption refinement. These experiments modify only the internal processing of captions and feedback within the generator; the critic, object inventory, and refinement loop remain unchanged.

We consider two reasoning variants. In the first, we use the same base model, `Llama-3.1-7B-Instruct`, but introduce a chain-of-thought prompting strategy. Instead of merely instructing the model to "think step by step," the prompt explicitly encourages comparative and contrastive reasoning. The model is asked to analyze multiple candidate captions and their associated scores, identify common patterns among higher-scoring captions, contrast them with lower-scoring ones, reason about what information is missing or overly generic, and determine which details make a caption more specific and less likely to apply to other images. This structured reasoning prompt guides the model to internalize why certain captions receive higher scores and to synthesize improved variants accordingly, while still outputting only the final short caption.

In the second variant, we replace the caption generator with `Llama-3.1-7B-DeepSeek-R1-Distill`, a model explicitly aligned for reasoning and thinking. We enable its reasoning mode and prompt it to reflect on the relationship between the candidate caption, the structured feedback, and the implied visual content. As with the chain-of-thought setup, the model is encouraged to reason about omissions, hallucinations, and specificity, but this reasoning capability is embedded in the model itself rather than induced purely through prompting. Across both reasoning settings, no additional supervision or fine-tuning is applied. The generator is not shown the image, does not observe the critic's faithfulness score, and is constrained to the same output format and length as the baseline. The only difference is how the generator internally processes the available signals. This allows a controlled comparison between standard instruction-following generation and reasoning-enhanced generation within the same iterative refinement framework.

To support large-scale experiments and recovery from interruptions, the system writes a complete per-image trace to disk after processing each image. Each trace records the object inventory, all intermediate captions, and all critic outputs. If a trace already exists, the image is skipped, enabling exact resumption without recomputation. This design allows generation and evaluation to be fully decoupled and ensures that downstream metric computation and analysis operate on a fixed, reproducible set of caption trajectories. Running these experiments took approximately 36 hours on a single NVIDIA RTX A6000 with 48GB of memory, with roughly 6 hours spent on caption generation

and refinement and approximately 30 hours devoted to metric computation, dominated by SPICE evaluation.

3 Results

Iterative Refinement with Non-Revealing VLM Hints Figures 2 through 4 summarize the main experimental trends of the iterative refinement framework using non-revealing feedback from a vision–language model (VLM). The mean per-image metric curves in Figure 2 show that the refinement loop consistently improves caption quality relative to the initial caption across BLEU, ROUGE-L, METEOR, and SPICE. The largest gains occur immediately after the first refinement step, indicating that even minimal structured feedback is sufficient to correct the most salient errors, such as missing key objects or hallucinated entities. Subsequent steps yield diminishing returns and the curves gradually saturate, suggesting that most easily correctable issues are resolved early in the process.

In contrast, the VLM faithfulness score does not exhibit a strictly monotonic trend across steps. We attribute this behavior in part to the relatively high sampling temperature used for caption generation, which promotes diversity and can introduce variability across iterations. A systematic study of the interaction between decoding temperature and faithfulness is left for future work; therefore, we do not further analyze this metric in the present study.

Selection, Win Rate, and Best-So-Far Analysis The best-so-far analysis highlights the benefit of selection across refinement steps. Because individual steps can occasionally degrade a metric for a given image, selecting the best candidate among all steps yields a monotonic improvement curve, as shown in Figure 4. This suggests that a practical deployment could benefit from retaining multiple intermediate captions and selecting among them using a lightweight reranker or the critic score.

Complementarily, the win-rate analysis in Figure 3 measures the fraction of images whose score at step s exceeds their own initial caption at step 0. Across most metrics, a majority of images improve after only one or two refinement steps, demonstrating that the observed gains are not driven by a small subset of examples. Together, these analyses indicate that iterative refinement implicitly explores a set of candidate captions, and that simple selection strategies can capture most of the benefit without committing to later-step regressions.

Reasoning-Based Caption Generation Beyond the standard MILS-style refinement baseline, we evaluated two reasoning-oriented variants that modify only how the text generator processes existing information, without changing the feedback signal or the overall refinement loop. In the first variant, we apply chain-of-thought (CoT) prompting to the generator, explicitly encouraging it to analyze higher- and lower-scoring captions, reason about what distinguishes them, and identify which details are likely to improve scores. In the second variant, we replace the generator with a reasoning-aligned model distilled from DeepSeek-R1, using the same base architecture but enabling its internal thinking mode. As shown in Figure 2 and Table 1, both reasoning-based approaches reach their peak performance within the first few refinement steps. In some cases, their early-step performance is comparable to or slightly higher than the non-reasoning MILS baseline. However, unlike the baseline, performance under reasoning tends to plateau quickly or even decline with additional iterations. This behavior suggests that explicit reasoning helps the model rapidly exploit the available information, but also reduces exploration of alternative caption variants. Once the generator converges to a locally optimal caption according to its internal reasoning, subsequent refinements tend to make fewer meaningful mutations.

Interestingly, chain-of-thought prompting consistently outperforms the distilled reasoning model. We hypothesize that while both approaches encourage structured reasoning, CoT prompting preserves more flexibility in generation, allowing the model to continue exploring the caption space. In contrast, the reasoning-aligned model appears to converge more aggressively, limiting diversity and reducing the chance of discovering unexpected but high-quality captions in later steps.

Memory-Based Retrieval and Concept Extraction We explored a simple alternative to memory-based retrieval by extracting high-level visual concepts (such as objects, actions, relations, and scene attributes) from intermediate captions and adding them as structured feedback within the

Method	$BLEU_4$	CIDEr	METEOR	SPICE
ZeroCap [3]	2.6	14.6	11.5	5.5
ConZIC ([4])	1.3	13.3	11.2	5.0
CLIPRe ([2])	4.6	25.6	13.3	9.2
$\text{MeaCap}_{\text{TF}}$ ([5])	7.1	42.5	16.6	11.8
$\text{MeaCap}^{*}_{\text{TF}}$ ([5])	4.5	26.0	14.1	9.4
MILS ([1])	8.0	33.3	15.0	9.6
MILS (LLaMA-3.1-7B)	**8.83**	**36.89**	15.30	9.53
VLM Hint (Qwen2-VL + Qwen2.5-7B)	0.43	20.98	**15.77**	**13.00**
CoT Prompting (LLaMA-3.1-7B)	6.48	29.42	15.37	9.66
Reasoning Model (R1-Distill-LLaMA-3.1-7B)	4.46	24.07	13.26	7.84

Table 1: Comparison with prior methods. Results above the midrule are taken from prior work. Results below the midrule are from our experiments on 1000 MS COCO test examples. * denotes results obtained by running the authors' released code.

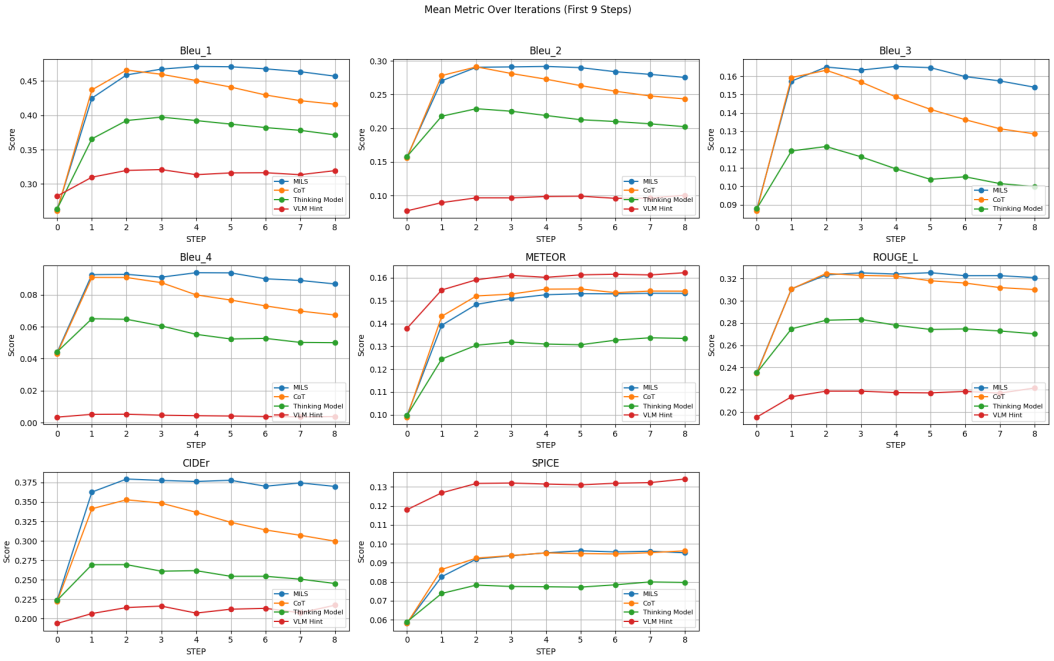

Figure 2: Mean per-image metrics as a function of refinement steps. VLM Hint exhibits substantial gains in the early iterations, followed by gradual saturation. In contrast, both CoT prompting and the reasoning model reach their peak performance within the first few steps and subsequently plateau or decline.

MILS optimization loop. The motivation was that a useful memory mechanism should provide the model with the key visual concepts it has already identified, so that future generations can build on this information. However, our experiments showed limited improvement over the original MILS framework. We believe this is because the Generator LLM already captures many of these concepts implicitly when generating and refining captions, making the additional concept-based feedback largely redundant. In contrast, we observed that the quality and diversity of the initial prompt set had a much stronger effect on convergence speed and final performance. This suggests that, in training-free iterative frameworks like MILS, better initialization may be more impactful than adding an explicit memory component that largely repeats information the Generator already encodes.

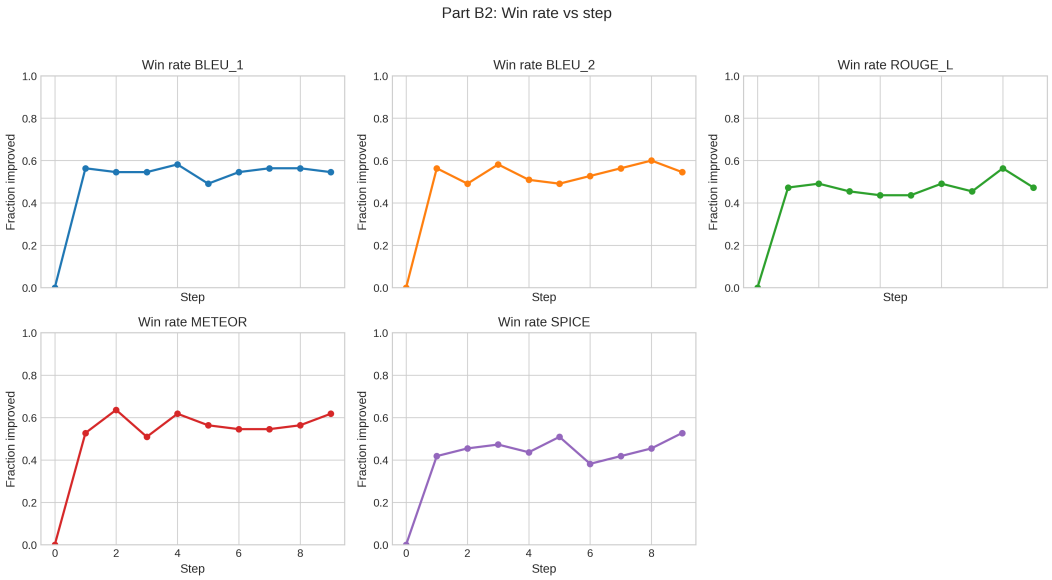

Figure 3: Win rate relative to step 0. For several metrics, a majority of images show improvement over their own initial caption.

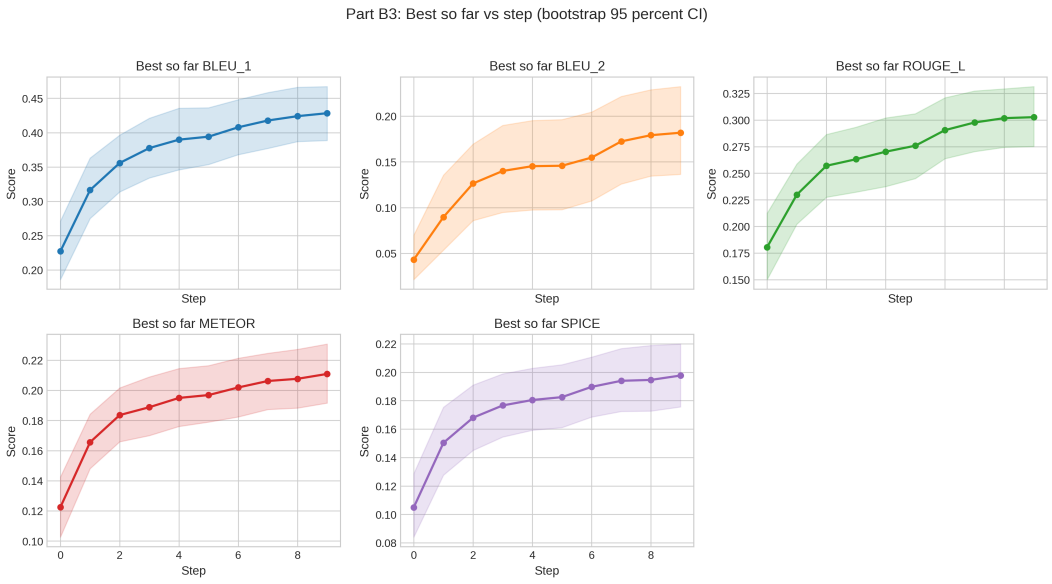

Figure 4: Best so far trajectories. Selecting the best caption across steps yields steadily increasing performance, suggesting that selection can further improve results beyond taking the final step.

4 Discussion

Our experiments reveal a clear trade-off between reasoning-guided refinement and exploratory refinement driven by structured but minimal feedback. Explicit reasoning—whether induced through CoT prompting or through a reasoning-aligned model—helps the generator rapidly identify high-scoring caption patterns. This leads to fast convergence and strong early performance. However, this same property appears to limit exploration. Once the generator identifies a locally optimal caption structure, subsequent refinements produce only small variations, reducing the diversity of candidate captions and increasing the risk of premature convergence.

In contrast, the baseline MILS-style refinement behaves more like a stochastic exploration process guided by score-frequency associations rather than explicit reasoning. Although this approach converges more slowly, it continues to explore alternative phrasing and object combinations that may initially appear suboptimal. Over multiple steps, this broader exploration allows the system to discover unexpectedly strong captions that are not immediately favored by reasoning heuristics.

The superior performance of CoT prompting relative to the reasoning-distilled model suggests that how reasoning is integrated matters. CoT prompting encourages the generator to analyze score patterns and caption differences while still relying on the base model's generative flexibility. The reasoning-distilled model, by contrast, appears more strongly biased toward optimization under its internal reasoning objective, leading to faster saturation and reduced adaptability across iterations.

These observations suggest that reasoning is beneficial for identifying promising directions in the caption space, but excessive or overly rigid reasoning may suppress beneficial mutations during iterative refinement. In training-free refinement frameworks like MILS, this balance between reasoning and exploration plays a critical role in determining long-term performance.

5 Conclusion

We presented an iterative caption refinement framework designed for constrained deployment settings in which the text generator never observes image pixels. A vision–language critic provides non-revealing structured feedback, while a text-only generator refines captions over multiple steps under strict COCO-style constraints. Beyond demonstrating consistent improvements with VLM-based hints, we systematically investigated the role of reasoning in the text generator.

Our results show that reasoning-based generation—via Chain-of-Thought prompting or reasoning-aligned models—can rapidly achieve strong early performance, sometimes matching or exceeding the baseline. However, these methods converge quickly and often plateau or degrade in later steps. In contrast, the baseline MILS refinement explores the caption space more broadly and continues to uncover improved captions over additional iterations. Among reasoning approaches, CoT prompting offers a more favorable balance than the reasoning-distilled model, combining early gains with greater robustness.

Overall, our findings highlight that while reasoning can guide refinement efficiently, exploration remains essential for discovering high-quality captions in iterative, training-free settings. Future work may investigate hybrid strategies that adaptively modulate reasoning strength across refinement steps to combine fast convergence with sustained exploration.

Acknowledgments

This work was conducted as part of a course project. We thank the course instructor and teaching assistants for their guidance and helpful discussions throughout the project.

References

[1] Kumar Ashutosh, Yossi Gandelsman, Xinlei Chen, Ishan Misra, and Rohit Girdhar. Llms can see and hear without any training. In *ICML*, 2025.

[2] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning via text-only training. In *The Eleventh International Conference on Learning Representations*.

[3] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17918–17928, 2022.

[4] Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, and Zhengjue Wang. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23465–23476, 2023.

[5] Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, and Zhengjue Wang. Meacap: Memory-augmented zero-shot image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14100–14110, 2024.