

MM-SeR: Multimodal Self-Refinement for Lightweight Image Captioning

Junha Song¹, Yongsik Jo², So Yeon Min³, Quanting Xie³,
Taehwan Kim², Yonatan Bisk³, Jaegul Choo¹

¹KAIST, ²UNIST, ³Carnegie Mellon University

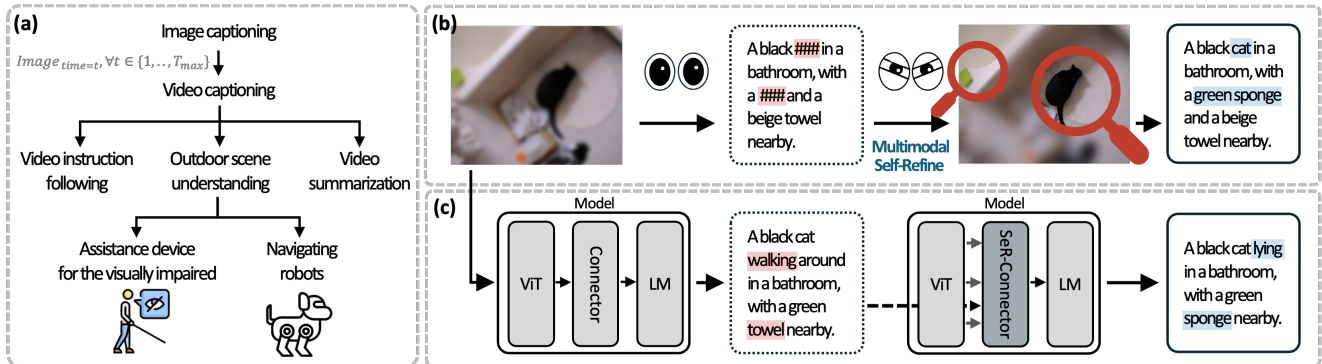


Figure 1. (a) Image captioning is a critical component for numerous assistive applications. However, current models often struggle to balance computational efficiency with performance, facing either deployment constraints or limited capability. We introduce a framework inspired by (b) the human visual process, which typically involves perceiving the global scene context before attending to local regions for specific details. (c) This observation highlights the necessity for multimodal self-refinement, a process our framework is designed to perform.

Abstract

Systems such as video chatbots and navigation robots often depend on streaming image captioning to interpret visual inputs. Existing approaches typically employ large multimodal language models (MLLMs) for this purpose, but their substantial computational cost hinders practical application. This limitation motivates our development of a lightweight captioning model. Our investigation begins by replacing the large-scale language component in MLLMs with a compact 125M-parameter model. Surprisingly, this compact model, despite a 93x reduction in size, achieves comparable performance to MLLMs, suggesting that **factual image captioning does not significantly require the complex reasoning abilities of LLMs**. Despite this promising result, our lightweight model still lacks reliability. To address this, we draw inspiration from the human visual process: perceiving a global and coarse understanding of the scene before attending to finer details. Accordingly, **we propose a multimodal self-refinement framework that guides the model to utilize features from salient regions, identified by referencing the previous coarse caption, and to produce a refined description**. Experimental results demonstrate the superiority of our model in both single-sentence and detailed captioning, extending even to long-range video QA tasks.

1. Introduction

Recent progress in image captioning has been driven by the remarkable capability of Multimodal Large Language models (MLLMs) [37, 41]. Building on these advances, image captioning has become a crucial component in various applications. For example, video-based chatbot systems utilize frame-wise *caption* generation for temporal understanding [73, 75, 81], while navigation robots construct graph-structured scene *descriptions* to operate in complex environments [28, 80]. Despite this progress, the substantial computational demands of MLLMs [29] pose a significant barrier to their practical deployment.

In many industrial systems, detection and segmentation models with fewer than 500M parameters are commonly deployed. This gap led us to ask whether captioning is truly so difficult that it must rely on MLLMs. To address this question, we design a lightweight captioning model that follows the architectural pattern of recent MLLMs and evaluate it on several captioning tasks. Specifically, we implement our model within the LLaVA framework [41] by replacing LLaMA-7B [64] with OPT-125M [90], a language model that is 56x smaller in parameter count. Surprisingly, the resulting model not only competes with MLLMs on the standard MS COCO [13] benchmark but also performs strongly

on more challenging detailed captioning tasks [67], while outperforming existing small captioning models [25, 57]. These results reveal the insight that the complex capabilities of LLMs are less critical for tasks that focus on enumerating *factual* visual details. These results also suggest that applying modern multimodal architectural designs to small captioning models is required to unlock their potential for deployment in real applications.

Despite the promising performance, the resulting model still exhibits a reliability gap compared to MLLMs. We attribute this gap primarily to visual blindness; prior work [62, 63] has highlighted that MLLMs often suffer from ambiguous visual features, which limits their ability to capture fine-grained details. This finding necessitates a method to supply the model with clearer and more informative visual inputs. We address this limitation by adopting a process similar to human visual perception, which begins with a global understanding of the scene before attending to local details. This multi-stage human approach contrasts with conventional captioning models, which typically operate in a single pass, processing the image only once to produce a description. To this end, we design a new framework, Multimodal Self-Refinement (MM-SeR), which enables the lightweight model to emulate this multi-stage human process. Specifically, MM-SeR enables the model to first generate an initial caption. The model then leverages this caption to guide its attention toward salient visual regions and extract richer information from the multi-layer features of the vision encoder. This refinement process allows the model to produce a more accurate final description.

In our experiments, we evaluate the MM-SeR framework on diverse benchmarks, extending beyond standard MS COCO [13] to include detailed captioning and long-form VideoQA tasks. We compare our model with MM-SeR against both existing small captioners and MLLMs, demonstrating comparable or even superior performance. Particularly in the challenging long-form VideoQA setting, our lightweight model not only outperforms other small specialists but also approaches the accuracy of MLLM generalists. This is accomplished while utilizing **93%** fewer parameters and achieving **82%** shorter inference time compared to the MLLMs. These results indicate that the proposed baseline and refinement framework offer a practical route toward lightweight captioning suitable for resource-constrained and on-device applications.

2. Motivation & Scope

Image Captioning as Foundational Technology. Image captioning converts visual content into natural language descriptions [24, 50]. Beyond being a standalone task, it serves as a core component in various vision–language applications. In video-grounded chatbot systems [60, 74, 81, 92], cap-

tioning generates textual representations of multiple frames, which are then integrated into LLMs as prompts to guide instruction following. Similarly, exploration robots [28, 80] operating in disaster environments rely on captioning to encode observed scenes, enabling navigation and human interaction. In this work, we study image captioning as a key enabler for these applications.

Real-World Deployment Challenges. Recent studies on the above applications often employ open-source MLLMs [73] or cloud-based APIs (*e.g.*, OpenAI API [80]) as image captioners. In practice, these approaches face two major limitations: (1) open-source MLLMs demand computational resources beyond the capacity of edge devices [7, 61] as shown in Table 1, and (2) cloud-based APIs rely on stable network connectivity, which may be unavailable in disaster environments. Moreover, repeated captioning across multiple scenes [89] further increases the computational burden, making real-world deployment more difficult.

Table 1. Available memory on edge devices and **GPU usage of recent MLLMs** by parameter size under FP16 precision.

Edge devices		LLaVA-1.5 7B,	LLaVA-NeXT 34B,	LLaVA-OA 72B,
iPhone 16	Galaxy S25	mPLUG-Owl3 8B	InternVL 40B	Qwen2-VL 72B
8GB	12GB	16GB +	68GB +	140GB +

3. Exploring Lightweight Captioning

Motivated by the challenges discussed above, we explore a lightweight captioning model and examine its performance through extensive evaluation.

Model construction. To construct a lightweight captioning model, we aim to reduce the dependence on LLMs, which account for most of the computational cost in MLLMs (*e.g.*, 96% in LLaVA-7B [41] arises from LLaMA [64]). Accordingly, we replace the LLaMA-7B in LLaVA-1.5 with OPT-125M [90], a 56× smaller language model.

Experimental details. We adopt the publicly available LLaVA-1.5 codebase [42]. Except for replacing the language model, all training configurations remain consistent with the original setup, including batch size and learning rate. More implementation details are provided in Section 1.1. To train our model, we first pretrain the multimodal connector on the Caption Concept-balanced 558K dataset [42], followed by fine-tuning on task-specific datasets such as MS COCO [13], DCI [67], and ShareGPT4V [11]. We evaluate our model using standard metrics like CIDEr [68] as well as MLLM-as-a-Judge [8, 9] with GPT-4o-mini [1].

Generalist vs. Specialist. Following prior studies [78, 84], we define generalists as MLLMs trained on diverse datasets for broad objectives and evaluated in a zero-shot manner. In contrast, specialists refer to compact captioning models such as SmallCap [57] and Tag2Text [25], which are trained only on captioning datasets.

Table 2. **Comparison of captioning performance.** We evaluate our model and existing captioning models. Despite not introducing any newly proposed methods, our model achieves strong performance. Here, ‘G’ represents a MLLM generalist, while ‘S’ denotes the small captioning model. ‘*’ indicates models that are fine-tuned in this study, as they were not trained for detailed captioning tasks.

MS COCO [13]										
	model	venue	# data	# params	B@4 [54]	MET [16]	CIDEr [68]	BERT [91]	CLAIR [8]	GPT [9]
G	InstructBLIP [15]	NeurIPS23	>130M	8.2B	38.0	29.4	127.8	69.1	-	-
	Unified-IOXL [45]	ICLR23	130M	7.3B	37.0	29.5	123.6	68.2	-	-
	Shikra [10]	arXiv23	-	7.2B	-	-	117.5	-	-	-
	Qwen-VL [2]	arXiv23	>1.5B	9.6B	39.1	30.1	131.9	69.8	77.8±3.4	2.89±0.11
	LLaVA-1.5 [42]	CVPR24	1M	7.3B	39.4	29.5	133.7	69.4	78.1±3.8	2.93±0.10
	Cambrian [62]	NeurIPS24	2M	10.5B	40.1	30.9	137.5	-	78.2±3.2	3.02±0.13
S	I-Tuning [47]	ICASSP23	0.5M	250M	35.5	28.8	120.0	-	-	2.50±0.10
	CapPa [65]	NeurIPS23	1B	650M	-	-	125.8	-	-	2.67±0.08
	LocCa [69]	NeurIPS24	1B	600M	-	-	127.1	-	-	2.66±0.11
	SmallCap [57]	CVPR23	0.5M	450M	37.6	28.7	122.7	67.2	73.7±3.9	2.46±0.10
	Tag2Text [25]	ICLR24	4M	900M	38.4	30.0	128.7	69.3	76.1±3.1	2.78±0.08
	ViPCap [31]	AAAI25	4M	225M	37.7	28.6	122.9	-	-	-
	Ours		1M	450M	39.4	30.3	129.6	69.4	76.3±3.0	2.74±0.06

ShareGPT4V [11] & DCI [67]										
	model	venue	# data	# params	B@4 [54]	CIDEr [68]	BERT [91]	CAPT [18]	CLAIR [8]	GPT [9]
G	Qwen-VL [2]	arXiv23	1.5B	9.6B	10.8	35.6	37.2	48.4	57.5±3.2	3.05±0.08
	Qwen3-VL [3]	arXiv23	>1.5B	2B	-	-	46.8	50.4	-	3.14
	SmallVLM2 [49]	arXiv23	-	2.2B	-	-	47.2	47.8	-	3.05
	LLaVA-1.5 [42]	CVPR24	1M	7.3B	10.6	36.1	36.6	48.6	±	±
	EyesWideShut [63]	CVPR24	1M	7.6B	11.6	36.5	37.2	49.0	-	-
	Cambrian [62]	NeurIPS24	2M	10.5B	13.3	38.7	38.1	50.1	58.2±3.1	3.03±0.09
S	SmallCap* [57]	CVPR23	0.5M	450M	14.5	20.1	28.9	23.3	-	-
	Tag2Text* [25]	ICLR24	4M	900M	17.8	32.5	40.7	40.1	54.2±3.1	2.72±0.14
	Ours		1M	450M	18.0	40.5	43.1	45.9	54.6±3.4	2.74±0.12

GLaMM [59]										
	model	venue	# data	# params	B@4 [54]	CIDEr [68]	BERT [91]	CAPT [18]	CLAIR [8]	GPT [9]
G	LLaVA-1.5 [42]	CVPR24	1M	7.3B	8.8	23.4	35.1	40.0	53.8±4.0	3.02±0.10
S	Ours		1M	450M	16.5	29.1	38.7	42.0	51.8±4.1	2.64±0.09

Input	Response
What are we having for dinner?	A nice dinner with a friend. I'm going to have to go to bed. I'm going to <EOS>
What should we eat for a snack?	We eat for snack. We eat for snack. <EOS>
A Italian street scene features	a man and a woman in a car, with a man in a car, with a woman in a car<EOS>
A garden scene centers on	a garden of a garden of a garden <EOS>

Figure 2. **Examples generated by OPT-125M** show its limited capacity. However, we found that with slight fine-tuning, the model exhibits a surprisingly strong captioning ability.

3.1. Results on Single Sentence Captioning

We use the MS COCO Captions dataset [13], the most widely adopted benchmark for image captioning, where each caption consists of a single sentence, with an average length of about ten words. As shown in Table 2, our model outperforms previous small-scale captioning methods [25, 69] that contain fewer than 1B parameters. In particular, it achieves a CIDEr score 6.9 points higher than SmallCap [57], which also employs OPT-125M as its language backbone. Despite the absence of any newly introduced techniques, the model demonstrates unexpectedly strong performance, which we

further analyze in Section E.1. Moreover, when compared with generalist MLLMs [2, 14], our model attains comparable results while requiring far less computational resources.

3.2. Results on Detailed Captioning

For this experiment, we fine-tune our model on the DCI [67] and ShareGPT4V [11] datasets, and additionally include GLaMM [59]. We initially expected the lightweight model to underperform on this task, as OPT-125M offers limited language capacity due to its small parameter size, as shown in Table 2. However, the results in Table 2 and Figure 3 show that our model achieves unexpectedly strong performance on detailed captioning, contrary to our assumption.

3.3. Key Insight

Although MLLMs depend on the advanced reasoning capabilities of large language models for tasks such as visual question answering and instruction following, our experiments reveal a contrasting trend in factual image captioning. We find that accurate caption generation can be achieved without such reasoning-intensive processes, as a lightweight 125M-parameter language model attains comparable per-

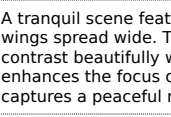
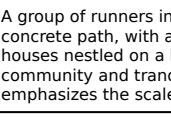
LLaVA-OneVision (7.3B)		Our lightweight model (450M)	
	A woman is holding up her cell phone to take a photo of a Christmas tree. The tree is decorated with bright, colorful lights that create a festive atmosphere. Other people can be seen in the background, possibly enjoying the holiday season.		A person's hand is seen holding a white smartphone, displaying a photo of a Christmas tree decorated with red and gold ornaments. The blurred background of red and gold lights creates a festive atmosphere, emphasizing the holiday spirit. The person's face is not visible.
	The image shows a blue butterfly resting on the top of a small green plant. Its wings are spread open, making the butterfly clearly visible near the center of the scene. The plant is located on the left side of the image, and the butterfly appears to be resting on it.		A tranquil scene features a blue butterfly perched on a green stem, its wings spread wide. The butterfly's striking blue and orange wings contrast beautifully with the stem, while the blurred background enhances the focus on the butterfly and its natural habitat. The image captures a peaceful moment in nature.
	A large group of people is running along a street. The runners are spread out along the road, with some closer to the foreground and others farther back. The scene appears to be taking place on a cloudy day.		A group of runners in vibrant athletic wear is captured mid-stride on a concrete path, with a clear blue sky overhead. The scene features houses nestled on a hill in the background, enhancing the sense of community and tranquility in the setting. The low-angle perspective emphasizes the scale of the race.

Figure 3. **Qualitative results on detailed captioning.** Despite its small size, our lightweight specialist produces captions that are competitive with those generated by large-scale MLLMs [34]. These results highlight that effective detailed captioning are achievable even with a lightweight language model. Qualitative results across various MLLM models are also presented in Figure 14, Figure 10, and Figure 11.

formance to full-scale MLLMs. This result implies that image captioning primarily relies on perceptual grounding rather than abstract reasoning, suggesting that a compact model can serve as an efficient and practical alternative for captioning-based applications.

4. MM-SeR: Multimodal Self-Refinement

Despite the impressive performance of our lightweight captioner, we still observe a capacity gap relative to large-scale MLLMs in certain evaluations. To address this limitation, we introduce a new framework, *Multimodal Self-Refinement* (MM-SeR), illustrated in Figure 1, which improves caption quality through self-guided refinement. First, we draw from the human description process: forming an initial global understanding before attending to salient details. Following this principle, MM-SeR adopts a multi-stage generation procedure. The model first produces an initial caption capturing the overall scene, and then uses this possibly coarse output to guide the extraction of clearer and more informative visual features, which support subsequent refinement.

4.1. Proposed Framework

As illustrated in Figure 4, our framework extends the conventional single-pass captioning approach by incorporating a self-refinement stage. Although the same language model is used for both the initial caption and the refinement step, the refinement process operates with a dedicated connector (*i.e.*, SeR-Connector). This connector is specifically tasked with integrating the following inputs: the previously generated caption and multi-layer features from the vision encoder. These inputs each serve a distinct purpose, as detailed below.

Looking at what matters. When asked to refine a caption such as “A cat relaxing on a brown chair,” humans naturally attend to the key elements referenced in the text, like the cat and the chair. Following this intuition, we feed the initial caption into the SeR-Connector and Language models, allowing the modules to identify visually relevant regions and direct their attention toward them during refinement.

Looking in detail. Coarse visual features often limit fine-grained description quality. While some prior studies [27, 62, 63] address this by adding auxiliary vision encoders, this strategy increases model size; for instance, Interleaved MoF [63] introduces DINOv2 [52], adding 300M parameters—a 66.7% increase relative to our 450M-parameter model. Instead of expanding the architecture, we aim to maximize the utility of the existing encoder by leveraging multi-layer features, which provide richer and more detailed representations. Although earlier works [39] have explored multi-layer features, our approach uses them specifically to supply finer-grained visual cues for the refinement stage.

Novelty of MM-SeR. This refinement paradigm parallels the concept of *self-refinement* explored in LLM research [33, 48, 53]. By extending this idea to the multimodal domain, our framework integrates visual evidence directly into the refinement. To the best of our knowledge, this is the first attempt to realize self-refinement within multimodal models.

4.2. Training Strategy

For the proposed framework to operate effectively, the model needs to be trained to produce reliable initial captions and to extract the refined visual features described in Figure 4. To achieve this, we adopt a two-stage training strategy.

In the first stage, the model is trained to generate the initial caption following the standard LLaVA procedure [42]. In the second stage, the model learns how to perform refinement. Let the training set be $X = \{x_1, x_2, \dots, x_N\}$, where each $x_k = (i_k, c_k)$ contains an image and its ground-truth caption. A straightforward strategy might treat the model’s momentary first-pass caption as the refinement input and the corresponding ground-truth caption c_k as the target. However, this setup often produces pairs with little semantic alignment. For example, if the initial caption is “a table in front of a window” while the c_k is “a cat sitting on a table,” the two descriptions offer no meaningful basis for progressive refinement. Training on such mismatched pairs would likely lead the model to disregard the initial caption

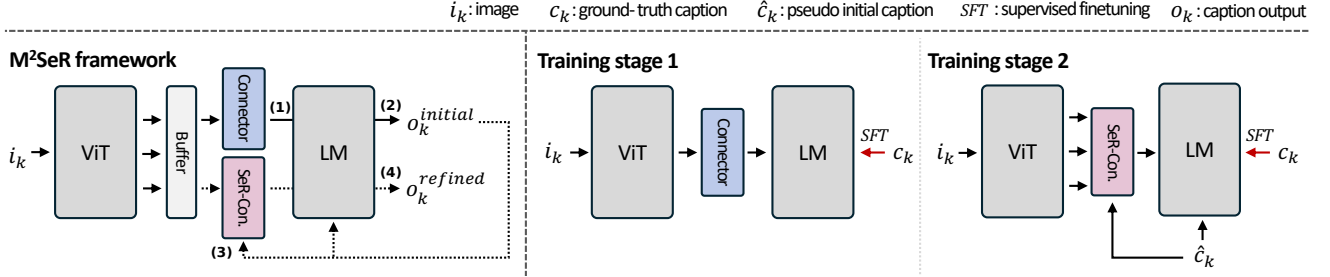


Figure 4. **Overview of the Multimodal Self-Refinement (MM-SeR) framework.** (left) The model first generates an initial caption, which guides the extraction of refined visual features for a second-stage generation. Fine-tuning is performed in two stages (right): (1) supervised training using ground-truth captions, and (2) refinement training using pseudo-initial captions that slightly deviate from the ground-truth to encourage self-correction. Here, ‘SeR-Con.’ denotes the SeR-Connector, which differs from the connector used for generating the initial caption. It processes additional inputs and is utilized in the refinement process.

and simply regenerate a new one, rather than learn how to refine it. Further discussion is in Section E.5.

To address this issue, we generate pseudo-initial captions $\hat{c}_k$ by prompting GPT-4o-mini [1] to introduce small perturbations to entities, attributes, or relations in the c_k . For example, given the c_k “a cat sitting on a chair,” the pseudo-initial $\hat{c}_k$ version may become “a dog sitting on a chair.” During training, the SeR-Connector receives multi-layer visual features together with the $\hat{c}_k$ and learns to extract features that are more informative for caption refinement. The language model takes the visual embeddings and $\hat{c}_k$ as an additional textual prompt and predicts a refined caption, which is supervised to match c_k .

Rationale for refinement training. Let π_θ denote the language model. Each pseudo-initial caption $\hat{c}_k$ deviates from the ground truth c_k at only a few token positions $E_k = \{t \mid \hat{c}_{k,t} \neq c_{k,t}\}$. Under the sequence-level objective

$$\mathcal{L}(\theta) = -\mathbb{E} \left[\sum_j \log \pi_\theta(c_{k,j} \mid i_k, \hat{c}_k, c_{k,<j}) \right], \quad (1)$$

gradients are likely to be primarily concentrated on tokens in E_k , leading to a form of *targeted optimization*, in which the model retains the correct parts of $\hat{c}_k$ while rewriting only the erroneous ones. Writing $\Delta_k(\theta) = \log \pi_\theta(c_k \mid i_k, \hat{c}_k) - \log \pi_\theta(\hat{c}_k \mid i_k, \hat{c}_k)$, we have $\mathcal{L}(\theta) \propto -\mathbb{E}[\Delta_k]$; hence, minimizing $\mathcal{L}$ is equivalent to maximizing the expected margin Δ_k , thereby directly increasing the likelihood of the refined caption relative to its *flawed* precursor. In effect, each gradient step encourages SeR-Connector to focus on visual regions likely responsible for errors in the initial caption and guides the language model to better interpret these refined features, resulting in more accurate captions. This targeted optimization resembles the philosophy of Direct Preference Optimization (DPO) [56], if we consider c_k and $\hat{c}_k$ as preferred and less-preferred responses. In contrast to DPO, which treats both responses symmetrically during optimization, our method offers a new perspective by assigning them distinct roles as input and target.

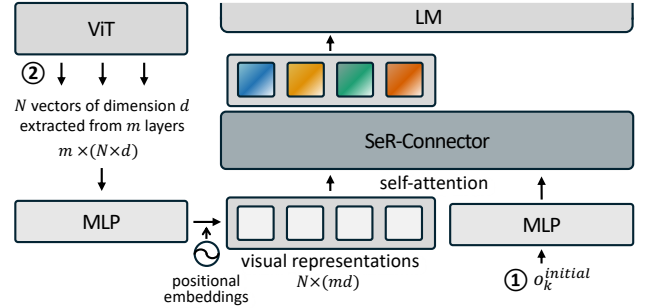


Figure 5. **Details of the SeR-Connector.** Unlike the standard two-layer MLP connectors used in typical MLLMs, the SeR-Connector is designed to support the refinement process and to effectively incorporate the inputs described in Section 4.1.

5. Experiments

5.1. Implementation Details

We train the framework in two stages. The first stage runs for 10 epochs to learn initial caption generation, and the second stage runs for 2 epochs to train the refinement process. Both stages use a batch size of 256 and a learning rate of 2×10^{-5} . While our primary experiments use OPT-125M as the language model, we additionally validate generalizability by replacing it with Qwen2.5-500M [82]. All experiments are conducted on two NVIDIA A6000 GPUs. Additional training settings are described in Section I.1.

SeR-Connector. The SeR-Connector is simply implemented with a set of Transformer blocks [17]. As shown in Figure 9, it receives two inputs: ① the previously generated caption, encoded as token embeddings, and ② multi-layer features from the vision encoder. From the vision encoder, we collect N visual token vectors of dimension d from m selected layers. The features are concatenated along the channel dimension, yielding a representation of size $N \times (md)$ that preserves hierarchical visual information. The resulting output is then forwarded to the language model for refinement. The ablation study is discussed in Section D.1.

Table 3. **Quantitative results of MM-SeR.** The results demonstrate the effectiveness of extending single-pass captioning with a self-refinement stage. Refinement relies on two key inputs: ① the initially generated caption and ② multi-layer features from the vision encoder. MM-SeR yields consistent performance gains on single-sentence and detailed captioning tasks.

MS COCO [13]									
model	#params	B@4 [54]	gain	CIDEr [68]	gain	CLAIR [8]	gain	GPT [9]	gain
LLaVA-1.5 [42]	7.3B	39.4	-	133.7	-	78.1±3.8	-	2.93±0.10	-
Our model	450M	39.4	-	129.6	-	76.3±3.0	-	2.74±0.06	-
+ MM-SeR with ①+②	500M	39.9	+0.5	133.5	+3.9	77.6±2.9	+1.3	2.82±0.09	+0.08
+ MM-SeR with ①	500M	39.6	+0.2	131.9	+2.3	-	-	-	-
+ MM-SeR with ②	500M	39.6	+0.2	132.3	+2.7	-	-	-	-
Single pass with ②	450M	39.7	+0.3	130.6	+1.0	-	-	-	-

ShareGPT4V [11] & DCI [67]									
model	#params	CIDEr [68]	gain	CAPT [18]	gain	CLAIR [8]	gain	GPT [9]	gain
Cambrian [62]	10.5B	38.7	-	50.1	-	58.2±3.1	-	3.00±0.10	-
Our model	450M	40.5	-	45.9	-	54.6±3.4	-	2.74±0.12	-
+ MM-SeR with ①+②	500M	43.6	+3.1	48.4	+2.5	57.7±3.0	+3.1	3.02±0.12	+0.28
+ MM-SeR with ①	500M	42.8	+2.3	47.1	+1.2	55.8±3.1	+1.2	2.78±0.11	+0.04
+ MM-SeR with ②	500M	43.1	+2.6	47.6	+1.7	56.8±3.4	+2.2	2.88±0.12	+0.14
Single pass with ②	450M	42.5	+2.0	46.5	+0.6	56.3±2.9	+1.7	2.90±0.11	+0.16
Ours (Qwen2.5-500M)	820M	44.2	-	49.0	-	58.3±3.0	-	3.10±0.10	-
+ MM-SeR with ①+②	870M	46.4	+2.2	50.8	+1.8	60.5±2.8	+2.2	3.17±0.09	+0.07

GLaMM [59]									
model	#params	CIDEr [68]	gain	CAPT [18]	gain	CLAIR [8]	gain	GPT [9]	gain
LLaVA-1.5 [42]	7.3B	23.4	-	40.0	-	53.8±4.0	-	3.02±0.10	-
Our model	450M	29.1	-	42.0	-	51.8±4.1	-	2.64±0.09	-
+ MM-SeR with ①+②	500M	30.4	+1.3	42.8	+0.8	53.4±3.8	+1.6	2.88±0.11	+0.24

Initially generated captions		After multimodal refinement	
	A striking blue and yellow train engine is stationed on a railway track, ready for its next journey. A red and white train car is visible in the background, set against a clear blue sky with a few clouds. The scene captures the essence of railway travel.		A striking blue and yellow train engine is stationed on a railway track, ready for its next journey. A black and grey car is visible in the background, set against a clear blue sky with no visible clouds. The scene captures the essence of railway travel.
	A woman in a white t-shirt and blue jeans is feeding a light brown sheep in a rustic barn. The sheep, with white coats and brown spots, stand on straw, while a black bucket and a wooden fence frame the scene. The image captures a peaceful moment in rural life.		A woman in a white t-shirt and blue jeans is petting a cream-colored sheep in a rustic barn. The sheep, with thick woolly coats, stand or lie on straw, while a black bucket and a wooden fence frame the scene. The image captures a peaceful moment in rural life.
	A man in a white lab coat and black pants is standing in front of a line of orange cheese blocks, with a metal fence and people in the background. The cheese blocks have different shapes and sizes, and the man's face is blurred out.		A man in a blue coat and black pants is standing in front of a line of round orange cheese wheels, with a red rope barrier and people in the background. The cheese blocks have different shape and size, and the man's face is blurred out.

Figure 6. **Qualitative comparison of initial and refined captions.** The examples demonstrate how the proposed MM-SeR improves the descriptive quality of image captions. Through the refinement process, some entity- and attribute-level errors are corrected, and vague expressions are replaced with more specific and visually grounded descriptions. More results can be found in Section F.

5.2. Results

Effect of MM-SeR. The results in Table 3 and Figure 6 demonstrate the effectiveness of our framework in improving caption quality. We present ablation studies on the two key inputs of our framework, ① the initial captions and ② the multi-layer visual features, to examine their individual contributions. The refinement stage improves the initial captions by +3.9 and +2.5 CIDEr points for the single-sentence and detailed captioning tasks, respectively. We further evaluate the use of ② without applying the refinement stage, observing that single-pass captioning provides only limited gains, indicating the necessity of the step. In our framework, adding the refinement stage inevitably introduces additional

overhead, including roughly 50M parameters for the SeR-Connector and one extra inference step with the language model. Nevertheless, as shown in Table 5, the computational overhead remains minimal compared with MLLMs.

Evaluation on long-range video question answering. Since captioning models serve as core components in downstream applications, as discussed in Section 2, we assess their utility in a practical setting by using our captioner as the captioning baseline for the long-range VideoQA task introduced in LLoVi [89]. (Setup.) For fair comparison, all small specialists [25, 57], including ours, are trained on ShareGPT4V [11] and DCI [67]. Following [89], we adopt Qwen2.5-14B as the backbone LLM for answer generation across all captioners to isolate the effect of caption quality.

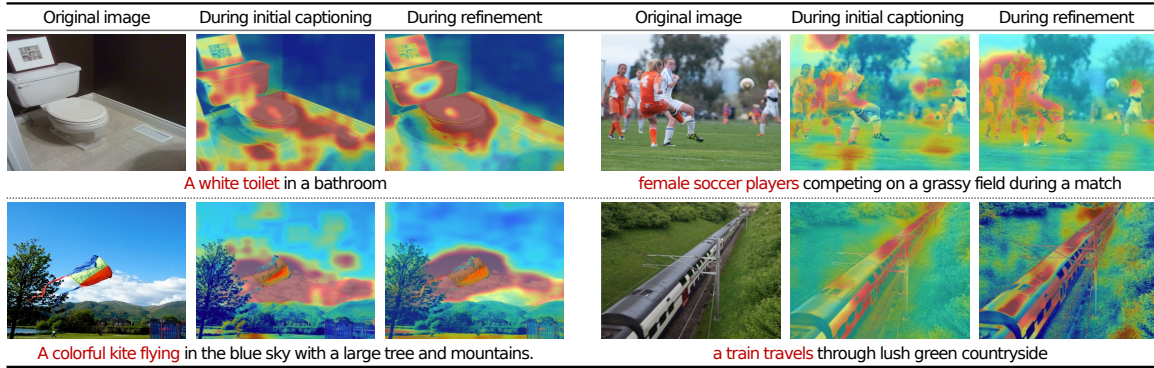


Figure 7. **Attention patterns during caption generation.** The left panel shows the model’s attention when producing captions from a *first-pass* view, where attention is broadly distributed. The right panel visualizes the refinement stage, in which the initial caption guides the model to focus on more relevant regions associated with the highlighted words.

Table 4. **Evaluation on Long-Range VideoQA.** We follow the baseline setup of LLoVi [89], replacing the captioner with our model or MLLMs. Frame-level captions are aggregated and provided to Qwen2.5-14B for answering video-related questions. Our specialist delivers competitive accuracy with significantly fewer parameters, demonstrating its applicability even in long-range video understanding.

LLM	Captioner	LLaVA-1.0 [41]	BLIP-2 [37]	LLaVA-1.5 [42]	SmallCap* [57]	Tag2text* [25]	Our specialist	+ MM-SeR
	#params	7.3B	7.4B	7.3B	450M	900M	450M	500M
Qwen2.5 14B [82]	accuracy	47.6	50.6	51.1	41.8	47.1	49.3	50.8
	time	29m 20s	29m 44s	29m 20s	4m 45s	7m 14s	4m 53s	5m 10s

Table 5. Inference time required to generate captions for 100 streaming images. Given the captioning performance in Table 3, our model demonstrates remarkable efficiency.

	LLaVA-1.5 [42]	Ours	+ MM-SeR
time	274.49 s	5.55 s (97.97%↓)	7.44 s (97.28%↓)

Inference time is measured end-to-end, accounting for both caption generation and LLM reasoning under a 10-minute video scenario. (*Results:*) As shown in Table 4, our specialist achieves 49.3 accuracy, outperforming prior small captioners such as SmallCap (41.8) and Tag2Text (47.1). When equipped with MM-SeR, performance further increases to 50.8, approaching the best generalist pipeline LLaVA-1.5 (51.1) despite using over 14× fewer parameters. In terms of efficiency, our specialist requires only 4m 53s, and 5m 10s with MM-SeR, which is substantially faster than generalist MLLMs (≈ 29 m). These results indicate that our lightweight captioner, combined with MM-SeR, offers strong suitability for real-world applications.

Visual analysis of initial caption utilization in MM-SeR.

A key input to our MM-SeR framework is the initial caption, which may reflect a rough and first-pass view of the image. To understand how this caption guides refinement, we examine the model’s visual attention during the generation process. (*Setup:*) We analyze the regions the model attends to when generating specific words and visualize attention maps using code adapted from API [86]. Further implementation details are provided in Section I.3. (*Results:*) As shown in Figure 7, the single-pass captioner often distributes its attention broadly across the image, struggling to localize

fine-grained regions. This behavior reflects the limitation of describing an image in a single glance, where the model attempts to process all information at once without focusing on details. In contrast, when the model performs the refinement step using the previously generated caption, the attention patterns become more concentrated on the relevant regions associated with each word. This suggests that the model leverages the initial caption as a guide, enabling it to “look at what matters” during refinement.

MM-SeR with larger language models. We extend MM-SeR to larger language models (LMs), including OPT-1.3B and LLaMA-2-7B, to examine whether the framework generalizes beyond lightweight models. To build captioning specialists, we first trained these LMs on the ShareGPT and DCI datasets. We then integrated MM-SeR and applied an additional refinement stage using the procedure described in Section 4.2. Our results in Table 6 show that MM-SeR provides consistent gains over these stronger baselines. Specifically, the framework improves CAPT [18] scores by 1.2 points for OPT-1.3B and 0.9 points for LLaMA-2-7B, indicating that MM-SeR remains effective when applied to larger LMs. Although larger LMs such as OPT-1.3B and LLaMA-2-7B offer stronger captioning capability as observed in prior scaling studies [24, 30, 76], their scale, approximately 10× and 56× larger than OPT-125M, can limit practical deployment. Developing captioning models that balance accuracy and efficiency, therefore, remains an important direction.

Iterative self-refinement. We also examine whether MM-SeR benefits from iterative refinement rather than a single refinement pass. As shown in Table 7, applying multiple

Table 6. **MM-SeR on larger language models.** We evaluate detailed captioning performance on ShareGPT4V [11] and DCI [67] using captioning specialists built from OPT-1.3B and LLaMA-2-7B. Across both models, MM-SeR consistently improves performance over their respective baselines, demonstrating that the refinement framework generalizes beyond lightweight LMs.

Language model	vision encoder	LoRA [23]	total #params	CIDEr [68]		CAPT [18]	
				initial gen.	+ MM-SeR	initial gen.	+ MM-SeR
OPT-1.3B [90]	CLIP ViT-L	×	1.6B	50.2	53.1 (+2.9)	49.0	50.2 (+1.2)
LLaMA-2-7B [64]	CLIP ViT-L	✓	7.3B	57.3	61.7 (+4.4)	52.7	53.6 (+0.9)

Table 7. **Effect of iterative refinement.** We compare initial captions and one to three refinement steps on the detailed captioning benchmarks ShareGPT4V [11] and DCI [67].

Language model	Stage	CAPT [18]	GPT [9]
OPT-125M [90]	initial caption	45.9	2.74±0.12
	refinement × 1	48.4	3.02±0.12
	refinement × 2	47.8	3.01±0.10
	refinement × 3	48.1	3.03±0.10
OPT-1.3B [90]	initial caption	49.0	3.04±0.10
	refinement × 1	50.2	3.14±0.11
	refinement × 2	50.5	3.16±0.10
	refinement × 3	50.3	3.20±0.10

refinement steps to the smallest LM, OPT-125M, provides no measurable improvement. In contrast, two or three refinement iterations yield meaningful gains for OPT-1.3B. This gap suggests that smaller LMs lack the capacity required to leverage multi-step refinement, whereas larger LMs are able to utilize the additional refinement signal effectively. The trend aligns with recent findings in LLM-based self-refinement [32, 48, 58, 79], which report increasing benefits as model capacity grows. Future work could explore scalable strategies, such as dynamically adjusting the iteration count, which remains a challenge in LLM refinement. We note that our core contribution is the demonstration that self-refinement can be an effective paradigm for MLLMs.

Remarks. We also evaluate our model on a resource-constrained device, Jetson Nano, in Section B.2. The strategy for generating pseudo-initial captions is detailed in Section G.1, and we discuss the limitations of our specialist captioner and existing MLLMs in Section E.

6. Related Work

Multimodal Large Language Models (MLLMs) have attracted considerable research attention due to their versatile applications, such as chat-bots [77]. Early approaches integrated contrastive image-language pretrained models [71] with powerful LLMs, enabling complex reasoning. The development of instruction-based datasets [41] and innovative training strategies [35, 84] has further accelerated progress, substantially improving MLLM performance and broadening their capabilities. Despite these achievements, most MLLMs heavily depend on large-parameter LLMs, making deployment on memory-constrained devices unfeasible.

Image Captioning Models. Recent advances in image captioning have improved training efficiency and descriptive

fidelity. Approaches such as CaMEL [6] and SmallCap [57] emphasize minimizing *trainable* parameters by leveraging mean-teacher distillation and employing retrieval augmentation, while Tag2Text [25] and LoCCa [69] introduce novel mechanisms such as dedicated tagging and location-aware refinement to improve caption quality. Unlike existing approaches that focus on reducing *trainable* parameters, or rely on single-pass inference, our method prioritizes *inference* efficiency, considering on-device operation and systematically addressing the limitations of single-pass generation.

Self-Refinement in LLMs. Humans often refine their writing through iterative review to enhance clarity and precision [19, 48]. Recent research has applied this refinement concept to LLMs, introducing techniques. For instance, Self-Refine [33, 53, 58] enabled models to autonomously critique and iteratively enhance their outputs. Unlike such approaches limited to the LLM domain, our method introduces refinement in a multimodal context, guided by both language and vision.

7. Conclusion & Broader Impact

We introduced a lightweight captioning approach with multimodal self-refinement, motivated by the need for efficient visual understanding systems on edge devices. Our study began with an OPT-125M-based captioner, which surprisingly achieved performance comparable to large MLLMs despite a substantial reduction in parameters (93%↓) and inference time (97%↓). To further enhance this lightweight model, we proposed a multimodal self-refinement framework, the first refinement-based approach explored in the MLLM community. Through extensive experiments, we demonstrated that the proposed architecture and framework provide more accurate and informative captions. These improvements hold consistently across single-sentence captioning, detailed captioning, and practical downstream tasks such as long-range video question answering. We hope that the model explored in this work serves as a practical solution for on-device applications and that our findings on visual self-refinement inspire deeper investigation in future multimodal research.

Broader Impact. Future research may explore integrating external tools (e.g., zoom or crop, as in GPT-o3 [51]), and designing a unified multimodal connector for both the initial and secondary glance. More research questions that aim to enhance both captioning performance and efficiency in real-world applications are provided in Section E.6.

Acknowledgments

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2025-00555621), and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2025-02653113, High-Performance Research AI Computing Infrastructure Support at the 2 PFLOPS Scale). This research was also supported by the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (Ministry of Science and ICT).

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmerschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 5, 14
- [2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 3, 13, 21
- [3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 3
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 22
- [5] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 14, 17
- [6] Manuele Barraco, Matteo Stefanini, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. Camel: mean teacher learning for image captioning. In *ICPR*, 2022. 8, 14, 16
- [7] Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. Tinytl: Reduce memory, not parameters for efficient on-device learning. In *NeurIPS*, 2020. 2
- [8] David Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. Clair: Evaluating image captions with large language models. In *EMNLP*, 2023. 2, 3, 6, 16, 17, 21, 27
- [9] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *ICML*, 2024. 2, 3, 6, 8, 16, 17, 19, 21, 27
- [10] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 3, 21
- [11] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*, 2024. 2, 3, 6, 8, 14, 16, 17, 19, 20, 21, 22, 23, 24
- [12] Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *ICLR*, 2023. 13
- [13] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1, 2, 3, 6, 16, 20, 21, 22, 23, 24
- [14] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023. 3
- [15] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 3, 21
- [16] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 2014. 3, 17
- [17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 5
- [18] Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024. 3, 6, 7, 8, 16, 17, 19, 20, 22
- [19] Shiha Dou, Yan Liu, Haoxiang Jia, Limao Xiong, Enyu Zhou, Wei Shen, Junjie Shan, Caishuang Huang, Xiao Wang, Xiaoran Fan, et al. Stepcode: Improve code generation with reinforcement learning from compiler feedback. *arXiv preprint arXiv:2402.01391*, 2024. 8
- [20] Zhengcong Fei. Efficient modeling of future context for image captioning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 13
- [21] Aviv Gabbay, Niv Cohen, and Yedid Hoshen. An image is worth more than a thousand words: Towards disentanglement in the wild. In *NeurIPS*, 2021. 20
- [22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 19
- [23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 8
- [24] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *CVPR*, 2022. 2, 7

- [25] Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. Tag2text: Guiding vision-language model via image tagging. In *ICLR*, 2024. 2, 3, 6, 7, 8, 13, 16
- [26] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaying Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *CVPR*, 2024. 22
- [27] Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin’e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*, 2023. 4
- [28] Hanxiao Jiang, Binghao Huang, Ruihai Wu, Zhuoran Li, Shubham Garg, Hooshang Nayyeri, Shenlong Wang, and Yunzhu Li. Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation. In *CoRL*, 2024. 1, 2
- [29] Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muiyang He, Bo Zhao, Xin Tan, Zhenye Gan, et al. Efficient multimodal large language models: A survey. *arXiv preprint arXiv:2405.10739*, 2024. 1
- [30] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 7, 17
- [31] Taewhan Kim, Soeun Lee, Si-Woo Kim, and Dong-Jin Kim. Vipcap: Retrieval text-based visual prompts for lightweight image captioning. In *AAAI*, 2025. 3
- [32] Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. In *ICLR*, 2025. 8, 13
- [33] Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. In *NeurIPS*, 2022. 4, 8
- [34] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *TMLR*, 2025. 4, 16, 22
- [35] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In *EMNLP*, 2022. 8
- [36] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 16
- [37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1, 7, 14, 16
- [38] Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. Factual: A benchmark for faithful and consistent textual scene graph parsing. *arXiv preprint arXiv:2305.17497*, 2023. 20
- [39] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 4
- [40] Fengyuan Liu, Zhongjian Hu, Peng Yang, and Xingyu Liu. Iterative caption generation with heuristic guidance for enhancing knowledge-based visual question answering. *Computer Vision and Image Understanding*, 2025. 13
- [41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2, 7, 8, 13, 16
- [42] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 2, 3, 4, 6, 7, 14, 21, 22
- [43] Imms-lab. Llava-recap-558k. <https://huggingface.co/datasets/imms-lab/LLaVA-ReCap-558K>. Accessed: 2025-05-20. 14
- [44] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *CVPR*, 2018. 13
- [45] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Motlaghi, and Aniruddha Kembhavi. UNIFIED-IO: A unified model for vision, language, and multi-modal tasks. In *ICLR*, 2023. 3, 21
- [46] Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Jianlin Feng, Hongyang Chao, and Tao Mei. Semantic-conditional diffusion networks for image captioning. In *CVPR*, 2023. 13
- [47] Ziyang Luo, Zhipeng Hu, Yadong Xi, Rongsheng Zhang, and Jing Ma. I-tuning: Tuning frozen language models with image for lightweight image captioning. In *ICASSP*, 2023. 3
- [48] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*, 2023. 4, 8, 13
- [49] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 3
- [50] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2, 16
- [51] OpenAI. Introducing O3 and O4 Mini. <https://openai.com/index/introducing-o3-and-o4-mini/>, 2025. Accessed: 2025-05-15. 8, 14
- [52] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 4, 14, 16, 17
- [53] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*, 2023. 4, 8

- [54] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. [3](#), [6](#), [16](#), [17](#)
- [55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. [14](#), [16](#), [17](#)
- [56] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023. [5](#)
- [57] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjieva. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *CVPR*, 2023. [2](#), [3](#), [6](#), [7](#), [8](#), [13](#), [14](#), [16](#)
- [58] Leonardo Ranaldi and André Freitas. Self-refine instruction-tuning for aligning reasoning in language models. *arXiv preprint arXiv:2405.00402*, 2024. [8](#), [13](#)
- [59] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, 2024. [3](#), [6](#), [20](#), [21](#)
- [60] Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-augmented contrastive learning for image and video captioning evaluation. In *CVPR*, 2023. [2](#)
- [61] Junha Song, Jungsoo Lee, In So Kweon, and Sungha Choi. Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In *CVPR*, 2023. [2](#)
- [62] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024. [2](#), [3](#), [4](#), [6](#), [13](#), [14](#), [16](#)
- [63] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024. [2](#), [3](#), [4](#), [13](#), [14](#), [16](#), [17](#)
- [64] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. [1](#), [2](#), [8](#), [18](#)
- [65] Michael Tschanen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. In *NeurIPS*, 2023. [3](#)
- [66] Michael Tschanen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. [14](#), [16](#), [17](#)
- [67] Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *CVPR*, 2024. [2](#), [3](#), [6](#), [8](#), [14](#), [16](#), [19](#), [20](#), [21](#), [22](#), [23](#), [24](#)
- [68] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. [2](#), [3](#), [6](#), [8](#), [16](#), [17](#), [19](#), [22](#)
- [69] Bo Wan, Michael Tschanen, Yongqin Xian, Filip Pavetic, Ibrahim M Alabdulmohsin, Xiao Wang, André Susano Pinto, Andreas Steiner, Lucas Beyer, and Xiaohua Zhai. Locca: Visual pretraining with location-aware captioners. In *NeurIPS*, 2024. [3](#), [8](#), [16](#)
- [70] Depeng Wang, Zhenzhen Hu, Yuanen Zhou, Richang Hong, and Meng Wang. A text-guided generation and refinement model for image captioning. *IEEE Transactions on Multimedia*, 2022. [13](#)
- [71] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. [8](#)
- [72] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *ECCV*, 2024. [22](#)
- [73] Yi Wang, Yanan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*, 2024. [1](#), [2](#)
- [74] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yanan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, 2024. [2](#)
- [75] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *CVPR*, 2025. [1](#)
- [76] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *TMLR*, 2022. [7](#)
- [77] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, 2023. [8](#)
- [78] Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhui Wang, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. In *NeurIPS*, 2024. [2](#)
- [79] Zhenyu Wu, Qingkai Zeng, Zhihan Zhang, Zhaoxuan Tan, Chao Shen, and Meng Jiang. Large language models can self-correct with key condition verification. In *EMNLP*, 2024. [8](#), [13](#)
- [80] Quanting Xie, So Yeon Min, Tianyi Zhang, Kedi Xu, Aarav Bajaj, Ruslan Salakhutdinov, Matthew Johnson-Roberson, and Yonatan Bisk. Embodied-rag: General non-parametric embodied memory for retrieval and generation. *arXiv preprint arXiv:2409.18313*, 2024. [1](#), [2](#)

- [81] Yifang Xu, Yunzhuo Sun, Benxiang Zhai, Ming Li, Wenxin Liang, Yang Li, and Sidan Du. Zero-shot video moment retrieval via off-the-shelf multimodal large language models. In *AAAI*, 2025. [1](#), [2](#)
- [82] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, others, and Jingren Zhou. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. [5](#), [7](#)
- [83] Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Rethinking temporal search for long-form video understanding. In *CVPR*, 2025. [22](#)
- [84] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*, 2024. [2](#), [8](#)
- [85] Qinghao Ye, Xianhan Zeng, Fu Li, Chunyuan Li, and Haoqi Fan. Painting with words: Elevating detailed image captioning with benchmark and alignment learning. In *ICLR*, 2025. [20](#)
- [86] Runpeng Yu, Weihao Yu, and Xinchao Wang. Attention prompting on image for large vision-language models. In *ECCV*, 2024. [7](#), [22](#)
- [87] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. [18](#)
- [88] Chaoning Zhang, Chenshuang Zhang, Junha Song, John Seon Keun Yi, Kang Zhang, and In So Kweon. A survey on masked autoencoder for self-supervised learning in vision and beyond. *arXiv preprint arXiv:2208.00173*, 2022. [18](#)
- [89] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. In *EMNLP*, 2024. [2](#), [6](#), [7](#), [22](#)
- [90] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. [1](#), [2](#), [8](#), [21](#)
- [91] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *ICLR*, 2020. [3](#), [17](#), [22](#)
- [92] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *CVPR*, 2023. [2](#)
- [93] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. [13](#), [22](#)