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

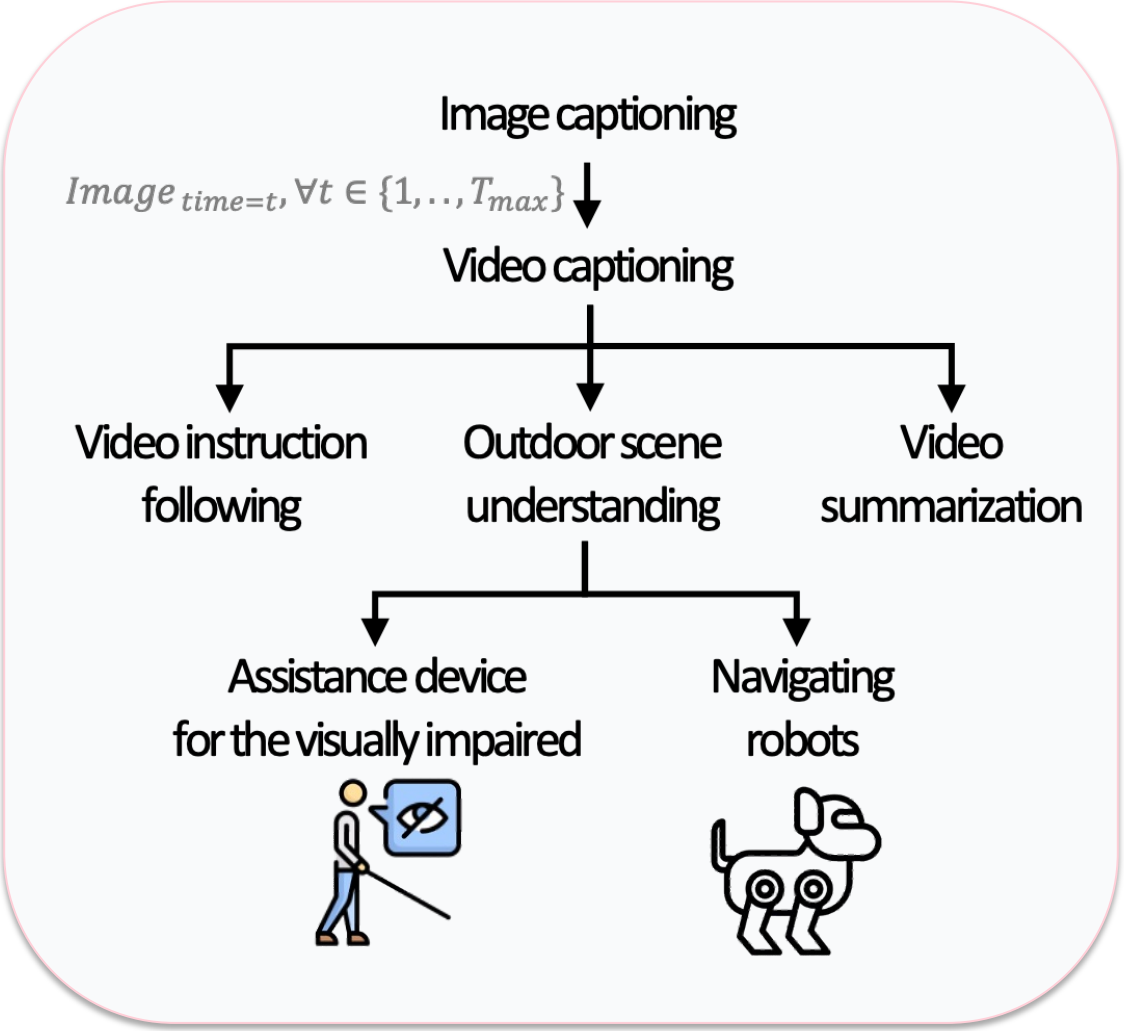
CVPR 2026 Main

KAIST UNIST CMU

MOTIVATION

Why Image Captioning Matters

Image captioning converts visual scenes into language for assistive tools, video understanding, and vision-language systems. These applications often need repeated captioning under practical compute limits.

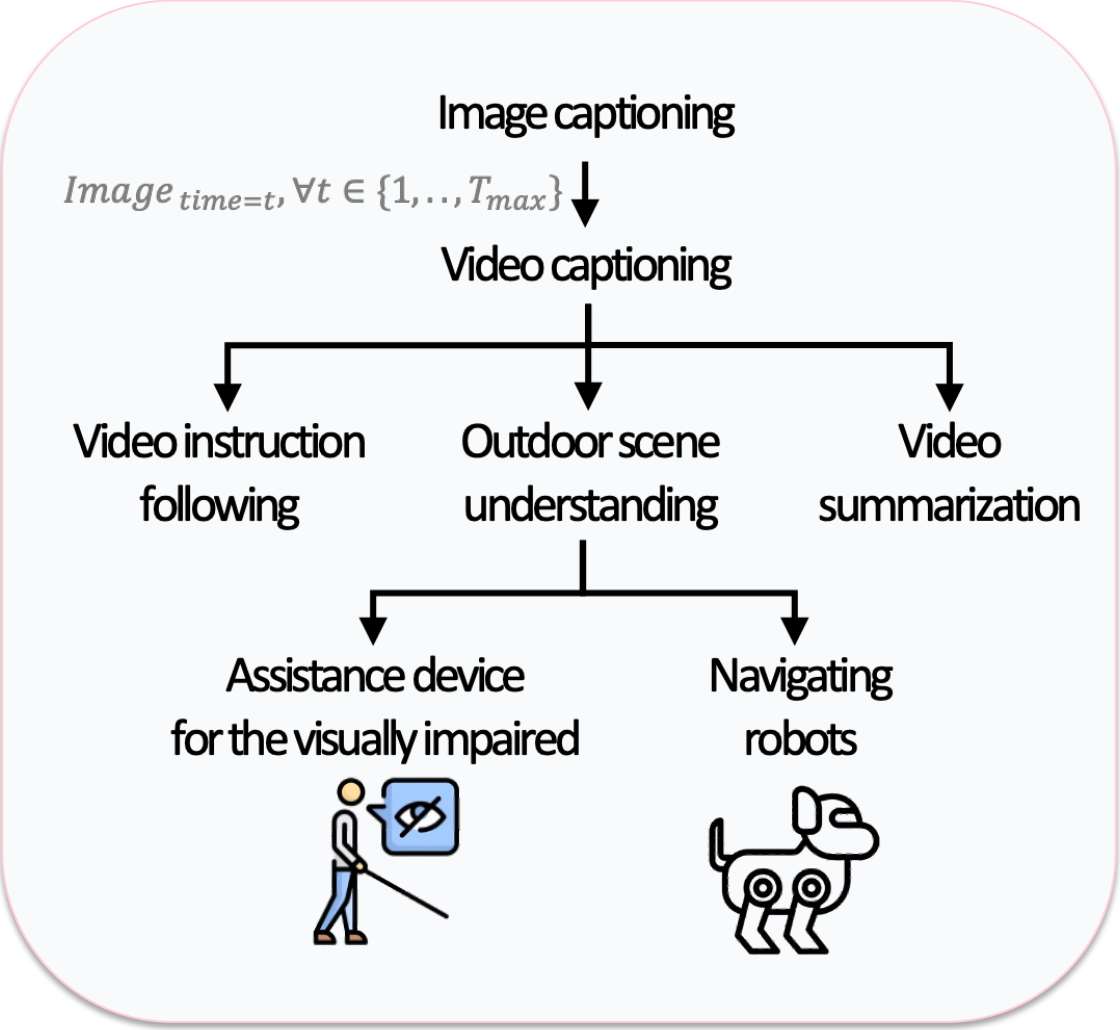


SCOPE

A Specialist Instead of a General MLLM

The goal is not to build another general-purpose multimodal LLM. MM-SeR targets a compact captioning specialist that can support deployment with lower cost.

Practical captioning needs reliable visual detail, not necessarily broad conversational reasoning.



QUESTION

Do Captions Need Large-Scale Reasoning?

The central question is whether factual image captioning truly requires the full reasoning capacity of large MLLMs. We test whether a compact language model can still act as a strong captioning specialist.

7B+

large captioning backbones

125M

compact LM tested

If compact models are competitive, captioning can become cheaper and easier to deploy.

EXPERIMENT

Lightweight Captioning Setup

We replace LLaVA-1.5's LLaMA-7B backbone with OPT-125M and evaluate detailed captioning on ShareGPT4V and DCI. This isolates whether the language model scale is the main driver of caption quality.

Backbone swap: LLaMA-7B -> OPT-125M

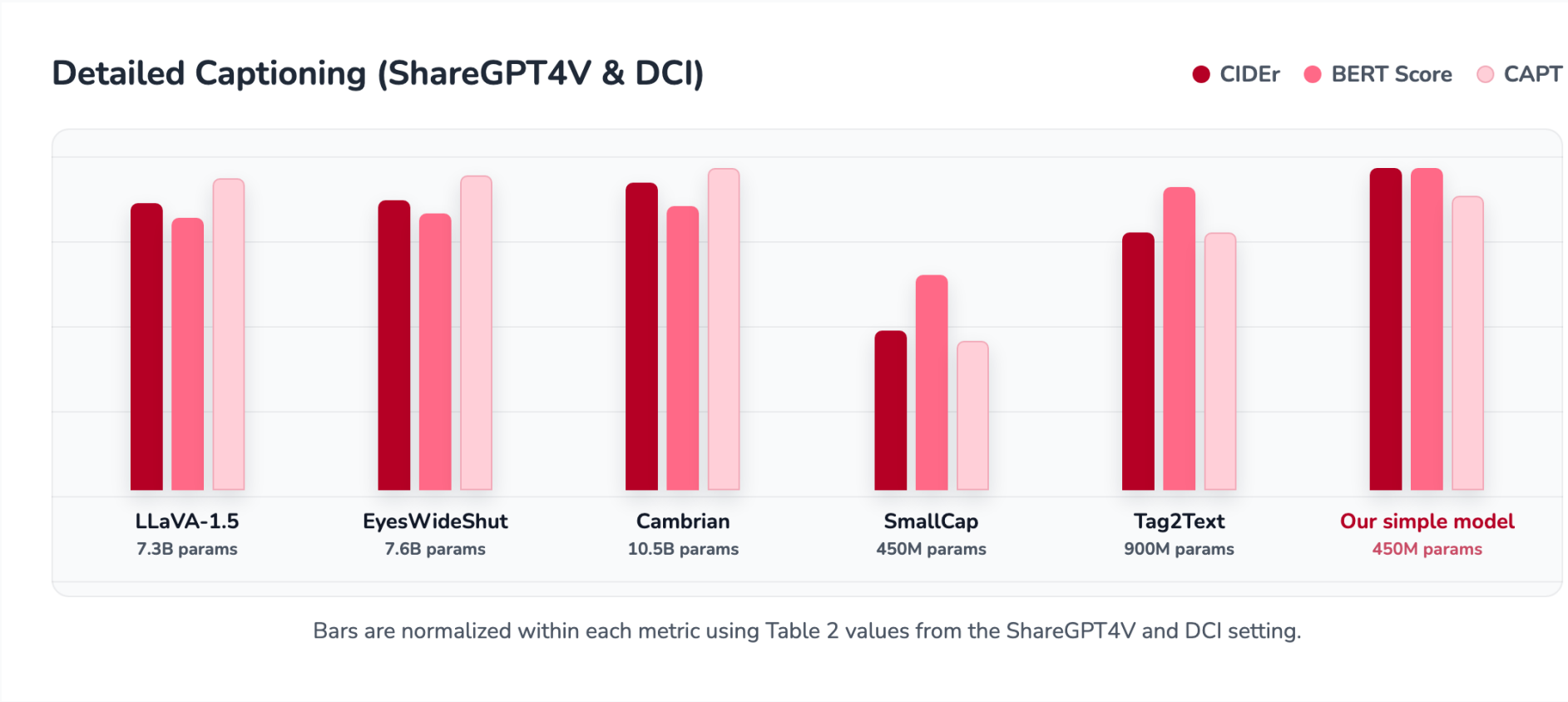
Benchmarks: ShareGPT4V and DCI

Metrics: CIDEr, BERT Score, CAPT

RESULT

A Compact Model Remains Competitive

The lightweight captioning model is highly competitive against much larger MLLM captioners. The result suggests that detailed captioning can be handled by a focused specialist.



INTERPRETATION

Key Finding

A 450M-parameter captioning specialist can match or exceed much larger captioning MLLMs on several detailed captioning metrics. Model specialization matters as much as raw language model scale for this task.

450M
simple model size

7.3B
LLaVA-1.5 size

10.5B
Cambrian size

900M
Tag2Text size

PROBLEM

The Single-Pass Limitation

A compact captioner can produce strong captions, but a single pass can still miss visual details or keep small errors. The model needs a way to revisit the image with guidance from its own first attempt.

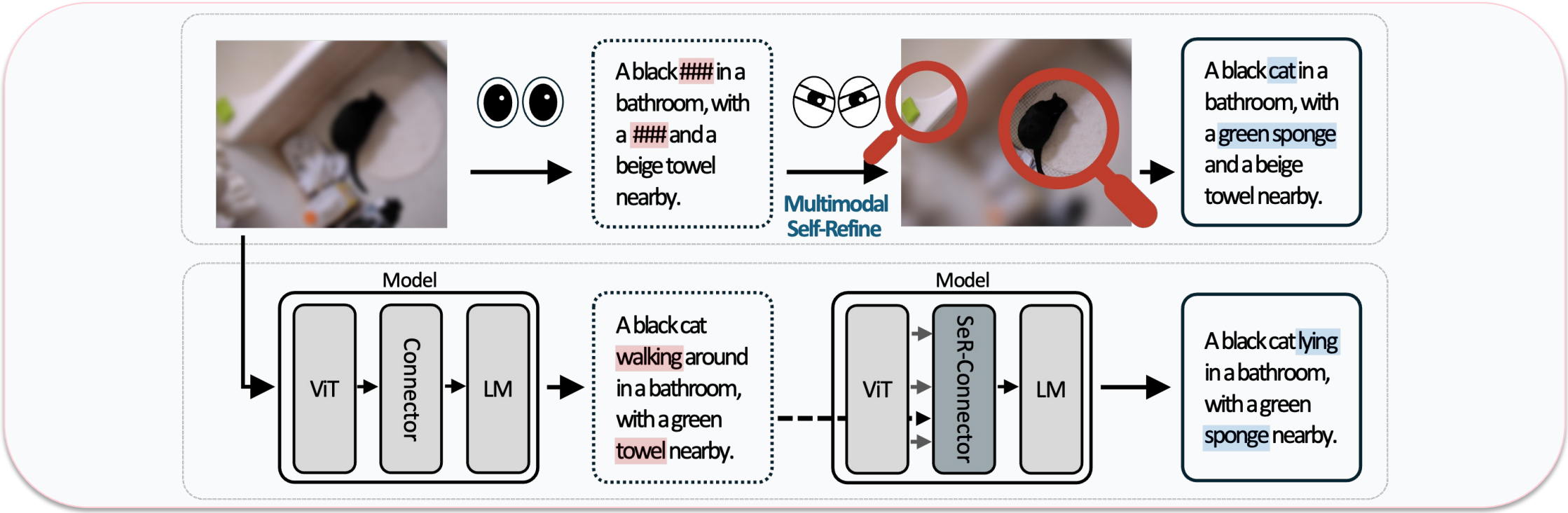
Initial captions are useful signals, but they are not always final answers.

Refinement should correct localized visual mistakes instead of regenerating blindly.

METHOD

MM-SeR in One View

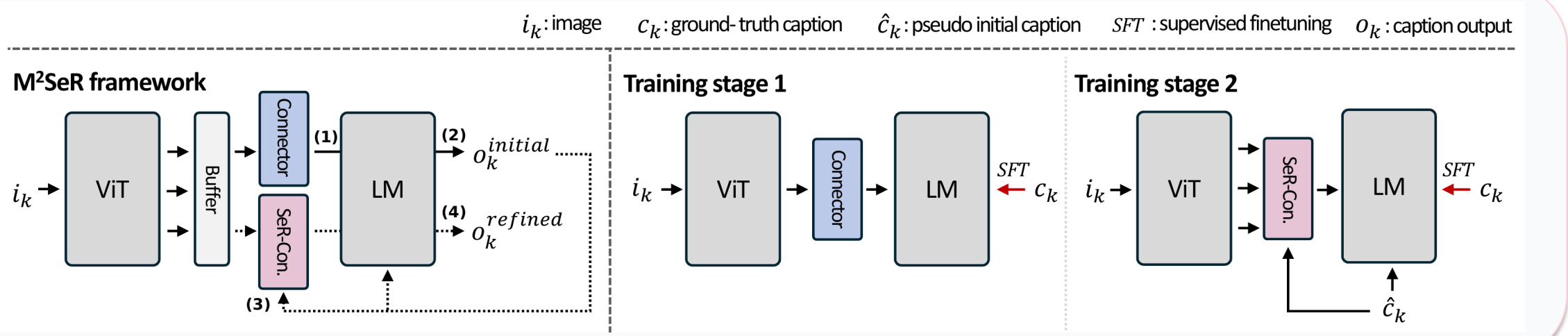
MM-SeR generates a coarse caption, uses it to focus on relevant visual evidence, and then produces a refined caption. This turns self-refinement into a multimodal process grounded in the image.



METHOD

Multimodal Self-Refinement Workflow

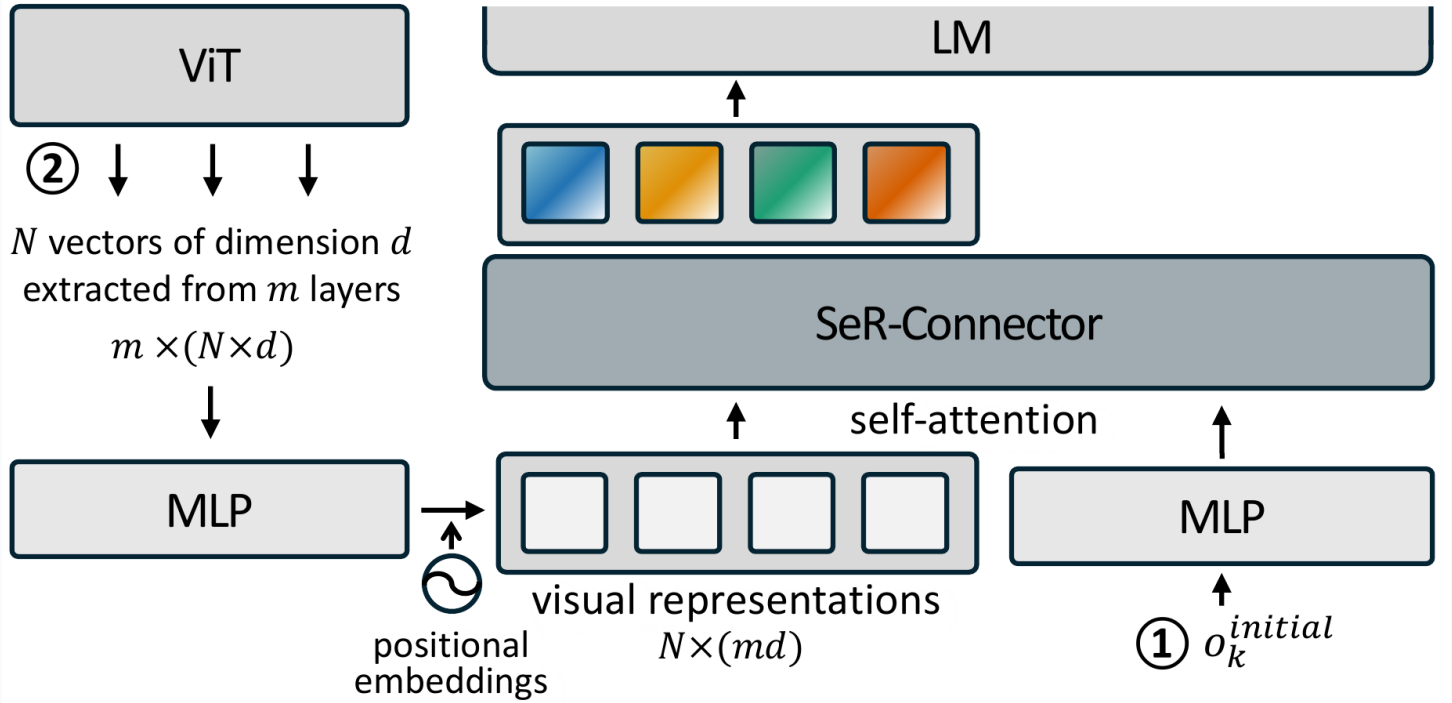
The initial caption guides the model toward salient visual regions. The second pass uses that guidance to produce a more reliable and visually grounded description.



Coarse caption -> guided visual revisit -> refined caption

The SeR-Connector

The SeR-Connector combines caption guidance with multi-layer vision features. This lets the model look at what matters and inspect those regions in more detail during refinement.



TRAINING

Training for Localized Correction

For refinement training, an LLM creates pseudo-initial captions by perturbing entity, attribute, or relation details. The model learns to correct localized mistakes rather than rewrite the whole caption from scratch.

Entity perturbation

Attribute perturbation

Relation perturbation

Target behavior: correction guided by visual evidence

EVALUATION

Experiment Suite

We evaluate MM-SeR across quantitative captioning, qualitative examples, long-range VideoQA, and iterative refinement. The experiments test both caption quality and downstream usefulness.

Detailed Captioning

Qualitative Results

Long-Range VideoQA

Iterative Refinement

RESULT

Detailed Captioning Improves

MM-SeR consistently improves CIDEr, CAPT, CLAIR, and GPT evaluation scores. The gains show that visually grounded refinement improves caption quality beyond the initial pass.

MM-SeR consistently improves CIDEr, CAPT, CLAIR, and GPT evals, demonstrating the effectiveness of our framework in improving caption quality.

RESULT

Qualitative Refinement

The refinement step corrects some entity- and attribute-level errors. It also replaces vague phrases with more specific descriptions grounded in the image.

	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.

DOWNSTREAM

Long-Range VideoQA Setup

Long-range VideoQA needs captions that preserve relevant details across many frames. We replace only the captioner in an LLoVi-style pipeline and pass aggregated captions to Qwen2.5-14B.

Captioner	Params	Acc.	Time
LLaVA-1.5	7.3B	51.1	29m 20s
Tag2Text	900M	47.1	7m 14s
Our simple model	450M	49.3	4m 53s
+ MM-SeR	500M	50.8	5m 10s

RESULT

VideoQA Accuracy with Lower Cost

With MM-SeR, the lightweight specialist reaches accuracy comparable to LLaVA-1.5 while using far less captioning time. This points to a useful cost-quality tradeoff for video pipelines.

Captioner	Params	Acc.	Time
LLaVA-1.5	7.3B	51.1	29m 20s
Tag2Text	900M	47.1	7m 14s
Our simple model	450M	49.3	4m 53s
+ MM-SeR	500M	50.8	5m 10s

50.8

MM-SeR accuracy

5m 10s

MM-SeR time

ANALYSIS

Iterative Refinement

The OPT-125M specialist gains little beyond the first refinement pass. The OPT-1.3B specialist benefits more from repeated refinement, matching trends seen in LLM self-refinement.

TAKEAWAY

Deployment Implication

MM-SeR suggests that captioning systems can be specialized, compact, and still reliable. This is especially relevant when captions must be generated repeatedly for frames, streams, or assistive applications.

Smaller captioner

Visual self-refinement

Better downstream efficiency

Practical multimodal systems

CLOSING

Main Takeaways

Lightweight captioning can be competitive when the model is specialized for the task. MM-SeR improves reliability by grounding self-refinement in visual evidence.

Compact specialists can reduce deployment cost.

Caption-guided revisiting improves detail and correctness.

Future work can scale adaptive iterative refinement.

Citation: Song et al., MM-SeR: Multimodal Self-Refinement for Lightweight Image Captioning, CVPR 2026.