WORLD LAUREATES FORUM
世界顶尖科学家论坛
国际工程智能大会
2026 International Conference on AI for Engineering
World Laureates Foundation
上海世界顶尖科学家发展基金会
2026
工程智能研究院
INSTITUTE OF AI FOR ENGINEERING

Unveiling Effective In-Context Configurations for Image Captioning: An External & Internal Analysis

Li Li[1], Yongliang Wu[1], Jingze Zhu[1], Jiawei Peng[1], Jianfei Cai[2], Xu Yang[1]*

[1] Southeast University, China [2] Monash University, Australia * Corresponding author

01 ABSTRACT

Multimodal in-context learning for image captioning depends on how demonstration images and captions are selected. We vary shot count, image retrieval, and caption assignment across OpenFlamingo, IDEFICS, Qwen2.5-VL, and Qwen3-VL. We then use attention-based diagnostics to explain performance changes. The results expose trade-offs between linguistic scores and visual grounding, alongside recurring anchor-token and within-example attention patterns.

02 EXTERNAL & INTERNAL ANALYSIS

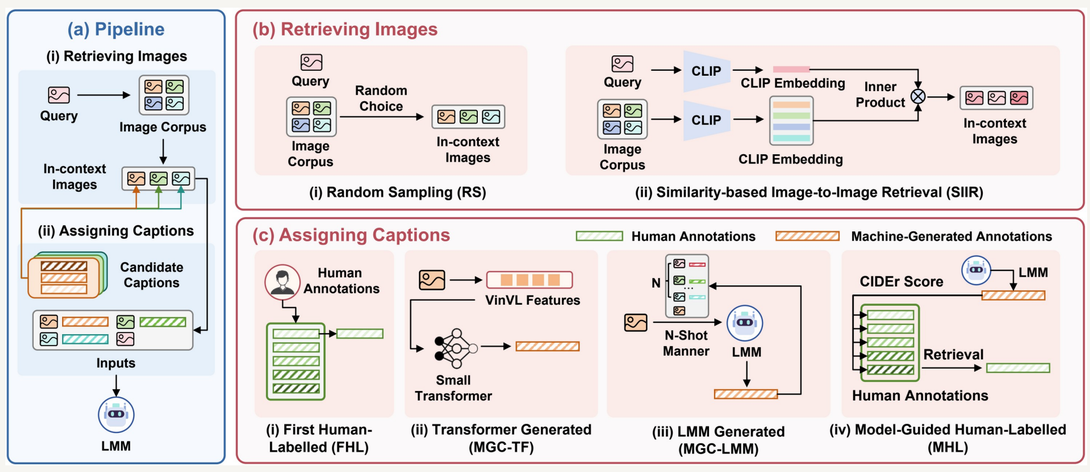


Figure 1. Image retrieval and caption assignment for in-context examples.

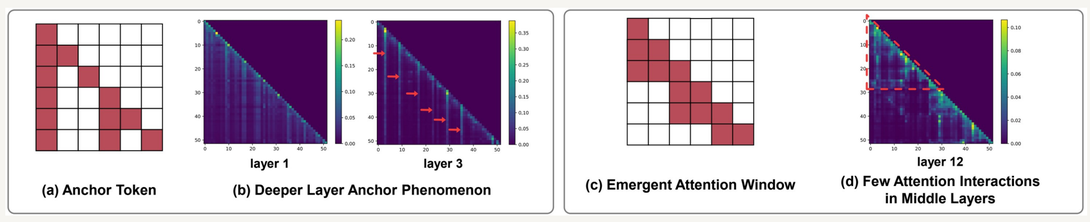


Figure 2. Anchor tokens and emergent attention windows in IDEFICS.

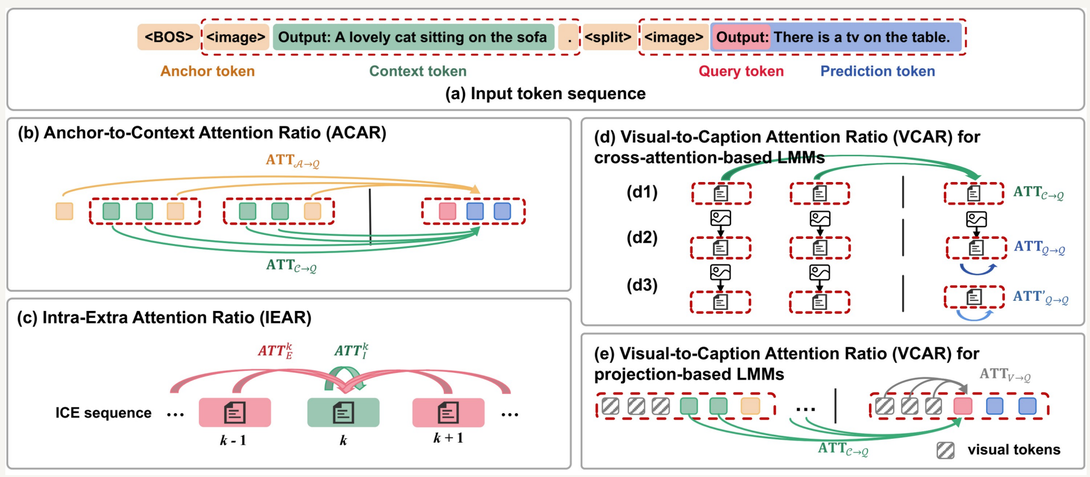


Figure 3. ACAR, IEAR, and VCAR quantify attention concentration, locality, and visual reliance.

Experimental design and metrics

External comparisons vary shot count (4–32 for Flamingo-style models; 1–8 for Qwen-VL), random sampling versus similar-image retrieval, and human versus generated demonstration captions. CIDEr, CLIPScore, hallucination, and caption copying assess output quality. Internally, ACAR measures anchor-token attention, IEAR measures within-example attention, and VCAR compares query-image use with demonstration-caption reliance.

03 HOLISTIC CONCLUSIONS FROM EXTERNAL EXPERIMENTS

1 More shots can separate language and vision

CIDEr and fluency may improve as context grows, while visual alignment weakens and OpenFlamingo hallucinates more. More text context does not ensure stronger visual reasoning.

2 Pretraining corpus affects this disparity

Related model designs behave differently when trained on corpora with different sequence lengths and image-text quality. Weaker data impair long-context grounding and stability.

3 Caption quality creates metric conflicts

Low-quality in-context captions compound errors at higher shot counts. A gain in linguistic coherence can coexist with lower visual alignment or more hallucination.

4 Caption style can displace visual evidence

Models imitate the wording of demonstration captions, sometimes at the expense of query-image accuracy. Diverse, model-consistent captions help stabilize results.

5 Similar-image retrieval can inflate CIDEr

Visually similar examples encourage copying of their captions. Higher CIDEr can coincide with weaker use of the query image and more hallucination in Flamingo-style models.

6 Similarity amplifies caption influence

Cross-attention magnifies both benefits and errors in captions paired with similar images. IDEFICS, with stronger visual understanding, is particularly sensitive.

04 SIGNIFICANT RESULTS IN INTERNAL ANALYSES

Anchor Token

ACAR confirms strong attention aggregation on a few image, punctuation, and boundary tokens, especially in middle and deep layers. IDEFICS shows stronger aggregation than OpenFlamingo, and Qwen-VL models exhibit the same pattern. With IDEFICS at 32 shots, masking non-anchor context tokens in layers 15–30 changes CIDEr from 104.8 to 102.7 (−2.0%).

Emergent Attention Window

IEAR shows that tokens interact more strongly within each in-context example than across examples, reaching about 20× in IDEFICS in middle and later layers. Context-centric masking at 32 shots leaves CIDEr essentially unchanged (104.8 versus 105.1). Qwen2.5-VL and Qwen3-VL show the same qualitative pattern.

These patterns support selective token and local-window computation, although the paper reports limited wall-clock acceleration in its tested implementation.

CONTACT

Xu Yang · xuyang_palm@seu.edu.cn

REFERENCE

Li et al., IEEE TPAMI (2026) · DOI: 10.1109/TPAMI.2026.3719106