AGIC: Attention-Guided Image Captioning to Improve Caption Relevance

L. D. M. S. Sai Teja[1] Ashok Urlana[2] Pruthwik Mishra[3]
[1]NIT Silchar [2]TCS Research, Hyderabad [3]SVNIT Surat, India
lekkalad_ug_22@cse.nits.ac.in, ashok.urlana@tcs.com,

https://image-caption-relevance.github.io/AGIC/

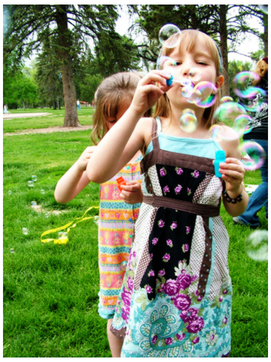
Abstract

Despite significant progress in image captioning, generating accurate and descriptive captions remains a long-standing challenge. In this study, we propose Attention-Guided Image Captioning (AGIC), which amplifies salient visual regions directly in the feature space to guide caption generation. We further introduce a hybrid decoding strategy that combines deterministic and probabilistic sampling to balance fluency and diversity. To evaluate AGIC, we conduct extensive experiments on the Flickr8k, Flickr30k and MSCOCO datasets. The results show that AGIC matches or surpasses several state-of-the-art models while achieving faster inference. Moreover, AGIC demonstrates strong performance across multiple evaluation metrics, offering a scalable and interpretable solution for image captioning.

1 Introduction

Image captioning, a prominent task in computer vision, aims to generate a visually grounded description of an image (Cornia et al., 2020, 2019; Chen et al., 2015). While significant performance improvements have been achieved, current methods frequently generate generic captions, limiting their utility in capturing nuanced visual details (Al Badarneh et al., 2025; Ma et al., 2023; Guo et al., 2020; Lu et al., 2018). The core challenge in producing relevant and descriptive image captions stems from the inherent difficulty in comprehensively capturing every visual aspect within an image.

To address limitations in caption relevancy and descriptiveness, existing approaches are broadly categorized into three main types: supervised, unsupervised, and semi-supervised methods. Supervised methods rely on large-scale, manually annotated image-caption pairs for training (Xu et al., 2015; Cornia et al., 2019; Anderson et al., 2018). While they often yield highly accurate captions, they are resource-intensive, requiring significant time and costly "gold-standard" data. In contrast, unsupervised methods utilize unpaired image sets and independent text corpora to learn visual concept detection and cross-modal alignment without explicit image-caption pairings (Feng et al., 2019; Guo et al., 2020; Laina et al., 2019). However, these models typically generate less precise or coherent captions due to the absence of direct supervision. Finally, semi-supervised methods (Liang et al., 2024) attempt to bridge this gap by generating pseudo-labels for unlabeled data, which are then used to guide the learning process. Although less data-hungry, these methods can suffer from error propagation if the pseudo-labels are inaccurate.



LlaVA: Two little girls are playing with bubbles in the park.
Qwen: Two young friends share a joyful moment, creating a world of bubbles in the park.
Fuyu: Two little girls blowing bubbles in the park.
BRNN: Two girls playing in the park.
LSTNet: A young girl is blowing bubbles, holding them in her hands.
R^2M: Two young girls playing with bubbles in grass.
AGIC: Two young girls wearing floral dress blowing bubbles in a park covered with grass.

Figure 1: Comparison of various image caption generation models. red: zero-shot, cyan: supervised, violet: unsupervised approaches and blue: our approach.

To address these challenges, we propose a training-free attention-guided approach that amplifies the pretrained model's attention weights to better capture salient image regions, combined with a hybrid decoding strategy for fluent and diverse caption generation. Our method enhances salient regions directly in the feature space of pretrained vision transformers without modifying or retraining the attention mechanism, enabling data-efficient and supervision-free captioning. In addition, we integrate a decoding strategy that combines beam search with Top-k and Top-p sampling to achieve

Findings of the Association for Computational Linguistics: EACL 2026, pages 6517–6528
March 24-29, 2026 ©2026 Association for Computational Linguistics

an effective balance between fluency and diversity—an aspect rarely explored in attention-based captioning. As illustrated in Figure 1, existing zero-shot and unsupervised methods often yield generic captions, while supervised ones fail to capture all relevant objects. In contrast, our AGIC approach successfully covers all key objects and produces grammatically fluent, descriptive captions, offering a practical, low-cost solution suitable for zero-shot and unsupervised settings.

Our key contributions are: 1) A novel attention-guided image caption approach designed to generate relevant and descriptive captions. 2) A novel hybrid decoding strategy for enhanced caption generation. 3) Extensive experiments conducted on three popular image caption datasets using several Vision-Language Models (VLMs), along with rigorous ablation studies.

2 Related Work

Image captioning is a challenging task, and several efforts have been made to solve this problem. A family of attention-based approaches (Anderson et al., 2018; Huang et al., 2019; Cornia et al., 2020; Sharma et al., 2018) have been incorporated for the object detection by capturing attributes, and other characteristics in images. Visual attention-based approaches focus on salient regions in an image using an object relation transformer that captures spatial dependencies between the detected objects (Herdade et al., 2019). Similarly, Anderson et al. (2018) propose bottom-up and top-down attention to combine object-level visual features with textual words. Huang et al. (2019) extends this using Attention-on-Attention (AoA) with an additional gating mechanism to determine the relevance of the final attention results and the incoming queries.

More recently, to infuse relevancy in image captioning, Honda et al. (2021) tackle spurious word-level alignments in pseudo-captioning by implementing a gating mechanism to filter out irrelevant visual words. Shi et al. (2021) develops a descriptive image captioning model using Natural Language Inference (NLI) between all pairs of reference captions. To better capture semantic information, Shi et al. (2020) constructs visual relationship graphs guided by captions. They also develop a multitask and weakly multi-instance learning framework for accurate explicit object and predicate detection. Another notable contribution by Pan et al. (2020) involves the modification of the conventional attention block into X-Linear attention networks to learn second-order feature interactions using bilinear pooling.

Although supervised approaches have achieved state-of-the-art performance, they rely on large paired datasets, whereas unsupervised efforts (Feng et al., 2019; Gu et al., 2019; Honda et al., 2021) suffer from semantic drift and coherence. In contrast, we propose a training-free approach by amplifying the pretrained model's attention weights to capture the relevant regions of the image and generate descriptive captions.

3 Model Description

We propose a framework to improve image caption relevance using a contextual relevance amplification mechanism, implemented through an attention-guided process. Our approach is inspired by Liu et al. (2025), who used attention patterns and self-reflection to detect hallucinations in large language models without relying on labeled data.

3.1 Attention Weights Extraction

In the AGIC framework, input images are passed to a pre-trained vision transformer model to obtain the corresponding attention weights of all image features. These attentive weights help to identify the most relevant regions of an image.

Let $X^{l-1} \in \mathbb{R}^{N \times d}$ represent the patch embeddings at layer $l-1$, where N is the number of image patches and d is the embedding dimension. The attention matrix $A^{l,h} \in \mathbb{R}^{N \times N}$ at layer l and head h is computed as:

$$A^{l,h} = \text{softmax}\left(\frac{(X^{l-1}W_Q^{l,h})(X^{l-1}W_K^{l,h})^T}{\sqrt{d_h}}\right) \quad (1)$$

Where $W_Q^{l,h}$ and $W_K^{l,h}$ are the query and key projection matrices at layer l, head h, and d_h is the head dimension. Then, to aggregate attention weights across all heads for a specific layer l, the attention weights received by the visual patch i in layer l are given as follows: $a_i^l = \frac{1}{H}\sum_{h=1}^{H} A_i^{l,h}$, where H is the total number of attention heads.

3.2 Image Amplification

To make all the relevant features in the image more prominent, we perform the image amplification step using the attention weights. For amplification, we multiply attention weights of all the image features with the original image representation with an

Dataset	Approach	B1	B2	B3	B4	R-L	METEOR	CIDEr	SPICE	CLIPScore
Flickr8k	BLIP2-opt-2.7b (Li et al., 2023)	0.391	0.255	0.163	0.102	0.325	0.259	0.077	0.041	0.685
	LLaVA-1.5B-7B (Liu et al., 2023)	0.440	0.293	0.193	0.128	0.357	0.049	0.198	0.080	0.721
	Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	0.441	0.276	0.170	0.107	0.311	0.034	0.280	0.072	0.702
	Fuyu-8B (Bavishi et al., 2023)	0.630	0.448	0.302	0.201	0.147	0.441	0.414	0.12	0.756
	BRNN (Karpathy and Fei-Fei, 2015)	0.563	0.362	0.219	0.131	0.350	0.171	0.512	0.112	0.744
	R^2M (Guo et al., 2020)	0.496	0.302	0.177	0.081	0.320	0.132	0.284	0.030	0.689
	LSTNet (Ma et al., 2023)	0.669	0.448	0.304	0.241	0.417	0.210	0.623	0.137	0.781
	AoANet (Huang et al., 2019)	0.638	0.437	0.302	0.207	0.398	0.229	0.549	0.172	0.768
	CapMAS (Lee et al., 2024)	0.151	0.086	0.046	0.023	0.147	0.173	0.011	0.129	0.565
	SPARC (Jung et al., 2025)	0.424	0.241	0.128	0.066	0.412	0.195	0.033	0.161	0.690
	AGIC + LLaVA-1.5B-7B	0.651	0.478	0.342	0.239	0.442	0.242	0.697	0.213	0.812
	AGIC + BLIP2-opt-2.7b	0.665	0.499	0.355	0.251	0.248	0.445	0.734	0.195	0.823
	AGIC + CLIP + BLIP2	0.696	**0.524**	**0.392**	**0.291**	**0.486**	**0.444**	**0.786**	**0.239**	**0.851**
MSCOCO	BLIP2-opt-2.7b (Li et al., 2023)	0.692	0.540	0.405	0.299	0.476	0.265	0.926	0.095	0.755
	LLaVA-1.5B-7B (Liu et al., 2023)	0.370	0.236	0.149	0.093	0.352	0.075	0.248	0.100	0.644
	Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	0.461	0.289	0.179	0.111	0.346	0.065	0.305	0.094	0.672
	Fuyu-8B (Bavishi et al., 2023)	0.691	0.539	0.411	0.305	0.473	0.278	0.958	0.110	0.783
	BRNN (Karpathy and Fei-Fei, 2015)	0.590	0.401	0.287	0.165	0.133	0.163	0.606	0.111	0.726
	R^2M (Guo et al., 2020)	0.467	0.281	0.102	0.099	0.351	0.144	0.255	0.125	0.701
	LSTNet (Ma et al., 2023)	**0.822**	**0.605**	0.375	0.216	0.514	0.233	0.991	**0.241**	0.834
	AoANet (Huang et al., 2019)	0.539	0.313	0.299	0.180	0.324	0.255	0.924	0.095	0.748
	CapMAS (Lee et al., 2024)	0.165	0.091	0.049	0.025	0.158	0.180	0.012	0.124	0.562
	SPARC (Jung et al., 2025)	0.415	0.232	0.121	0.062	0.401	0.191	0.035	0.155	0.701
	AGIC + LLaVA-1.5B-7B	0.411	0.263	0.164	0.100	0.382	0.270	0.649	0.115	0.768
	AGIC + BLIP2-opt-2.7b	0.698	0.547	0.414	0.310	0.475	0.288	1.026	0.130	0.789
	AGIC + CLIP + BLIP2	0.723	0.551	**0.414**	**0.312**	**0.518**	**0.310**	**1.092**	0.235	**0.853**

Table 1: Image captioning results for Flickr8k and MSCOCO; B1 to B4 refer to BLEU and R-L: ROUGE-L scores.

amplification factor *k* as shown in the equation 2.

$$I_a(i,j) = I_o(i,j) \cdot (1 + k \cdot a(i,j)) \tag{2}$$

Where $I_a(i,j)$ and $I_o(i,j)$ denote the amplified and original values at the spatial location (i,j), respectively. The term $a(i,j)$ represents the attention weights obtained during the first pass, and k is a hyperparameter that controls the strength of amplification. We obtain the attention-guided image representation I_a and pass it into the image captioning model to generate attention-guided captions.

3.3 Caption Generation

To enhance both diversity and fluency in the caption generation for the amplified image, we adopt a *hybrid decoding strategy* that performs stochastic beam search by combining beam search with Top-k, Top-p (nucleus) sampling, and temperature scaling. At each decoding step, candidate tokens are sampled from a probability distribution obtained by temperature-scaled softmax, followed by Top-k and Top-p filtering. Sampling is performed independently within each beam, and the final caption is selected from the best completed beam.

Formally, for the token logit distribution $\mathbf{z}_t$ at decoding step t, we sample the next token x_t as:

$$x_t \sim \text{Top-}p\left(\text{Top-}k\left(\text{Softmax}\left(\frac{\mathbf{z}_t}{T}\right)\right)\right) \tag{3}$$

where T is the temperature and the sampling is carried out within each of the B beams, which represents the number of parallel decoding paths to enhance contextual relevance in the generated captions. This enables our pipeline to focus on generating contextually relevant regions that lead to more detailed and meaningful image captions.

4 Experiments and Results Analysis

4.1 Setup

In this study, we use three popular datasets, Flickr8k (Hodosh et al., 2013), Flickr30k (You et al., 2016), and MSCOCO (Lin et al., 2014) to conduct the image captioning experiments. To perform the evaluation of the models, we utilize the BLEU ($n = 1, 2, 3, 4$) (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), Consensus-based Image Description Evaluation (*CIDEr*) (Vedantam et al., 2015), Semantic Propositional Image Caption Evaluation (*SPICE*) (Anderson et al., 2016), and CLIPScore (*CLP*) (Hessel et al., 2021) metrics.

4.2 Baselines

We compare our approach with comprehensive collection of baselines categorized as follows: (1) *zero-shot prompting-based* approaches by utilizing the popular VLMs such as BLIP2-opt-2.7b (Li et al., 2023), LLaVA-1.5-7B (Liu et al., 2023), Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and Fuyu-8B (Bavishi et al., 2023), (2) *unsupervised method* R^2M (Guo et al., 2020), (3) *supervised methods* such as BRNN (Karpathy and Fei-Fei, 2015), LSTNet (Ma et al., 2023), AoANet (Huang et al., 2019), CapMAS (Lee et al., 2024), and SPARC (Jung et al., 2025). The results can be seen in the Tables 1 and 10. More details on the experimental setup can be found in Appendix B.

4.3 Results Analysis

In our AGIC framework, we perform experiments using both the same and different models for attention extraction and caption generation, employing BLIP2-OPT-2.7B, LLaVA-1.5-7B and CLIP-ViT-Large-Patch14 (Radford et al., 2021). The combination (AGIC+CLIP+BLIP2), which utilizes CLIP for attention extraction and BLIP2 for caption generation, outperforms state-of-the-art models across multiple evaluation metrics. Results on the Flickr30k dataset are provided in Table 10 in the Appendix. As per Figure 2, our approach achieves significantly lower inference time compared to existing methods, offering a more cost-effective and time-efficient solution.

5 Details on Inference time comparison

When comparing the inference time per sample with other state-of-the-art models, as shown in Figure 2, our AGIC approach demonstrates significantly lower latency. Specifically, AGIC takes only 0.297 seconds (sec) per sample, whereas all zero-shot-based approaches require more than one second. The amplification step takes only 0.00028 sec, which is insignificant relative to the overall inference time. The extraction step takes 0.032 sec and caption generation takes 0.264 sec, resulting in a total of 0.297 sec per sample. Additionally, we observe that the unsupervised method (R^{2}M) is faster than the supervised methods (BRNN and LSTNet). Overall, AGIC achieves the lowest inference time among all compared methods.

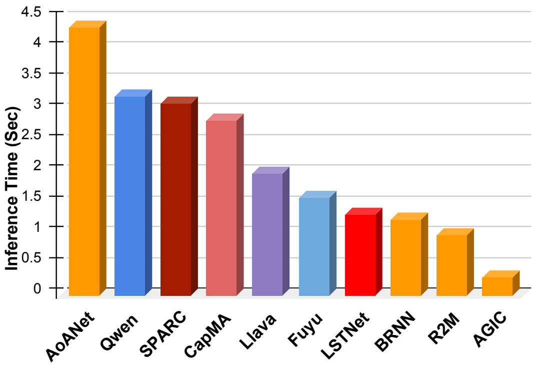


Figure 2: Inference time comparison.

		Config.	BLEU	R-L	CIDEr	SPICE	CLP
Decoding strategy	Flickr8k	CLIP-BLIP2	0.10	0.32	0.35	0.05	0.68
		CLIP-BLIP2+Top-k	0.07	0.29	0.31	0.11	0.68
		CLIP-BLIP2+Top-p	0.11	0.32	0.38	0.13	0.72
		CLIP-BLIP2+Beam	0.23	0.43	0.69	0.19	0.81
		CLIP-BLIP2+Top-p+Top-k	0.17	0.37	0.53	0.15	0.75
		CLIP-BLIP2+Top-p+Beam	0.22	0.43	0.68	0.20	0.81
		CLIP-BLIP2+Top-k+Beam	0.21	0.42	0.66	0.19	0.80
		CLIP-BLIP2+Top-p+Top-k+Beam	**0.29**	**0.48**	**0.78**	**0.23**	**0.85**
	MSCOCO	CLIP-BLIP2	0.16	0.33	0.45	0.09	0.71
		CLIP-BLIP2+Top-k	0.10	0.31	0.42	0.13	0.70
		CLIP-BLIP2+Top-p	0.14	0.35	0.52	0.15	0.75
		CLIP-BLIP2+Beam	0.31	0.47	0.95	0.21	0.85
		CLIP-BLIP2+Top-p+Top-k	0.21	0.41	0.76	0.18	0.81
		CLIP-BLIP2+Top-p+Beam	0.30	0.47	0.94	0.20	0.85
		CLIP-BLIP2+Top-k+Beam	0.30	0.46	0.94	0.20	0.85
		CLIP-BLIP2+Top-p+Top-k+Beam	**0.31**	**0.51**	**1.09**	**0.23**	**0.85**
Amplification factor	Flickr8k	AGIC (k=-5)	0.14	0.39	0.35	0.13	0.75
		AGIC (k=-3)	0.15	0.40	0.40	0.13	0.76
		AGIC (k=-1)	0.17	0.41	0.45	0.14	0.79
		AGIC (k=0)	0.23	0.43	0.70	0.18	0.84
		AGIC **(k=1)**	**0.29**	**0.48**	**0.78**	**0.23**	**0.85**
		AGIC (k=3)	0.23	0.44	0.71	0.18	0.84
		AGIC (k=5)	0.22	0.43	0.67	0.18	0.84
	MSCOCO	AGIC (k=-5)	0.18	0.32	0.40	0.15	0.77
		AGIC (k=-3)	0.18	0.32	0.40	0.15	0.77
		AGIC (k=-1)	0.22	0.35	0.45	0.19	0.78
		AGIC (k=0)	0.26	0.37	0.52	0.22	0.85
		AGIC **(k=1)**	**0.31**	**0.47**	**1.02**	**0.23**	**0.85**
		AGIC (k=3)	0.26	0.37	0.52	0.20	0.85
		AGIC (k=5)	0.25	0.37	0.55	0.20	0.85

Table 2: **Top:** Performance variation across decoding strategies. ‘All’ represents the combined strategy (Beam Search + Top-k + Top-p). **Bottom:** Performance comparison with varying amplification factors.

	Correctness (Corr)	**Completeness (Comp)**	**Relevance (Rele)**
Mean	4.04	3.64	4.26
	Corr–Comp	**Comp–Rele**	**Rele–Corr**
Correlation (r)	0.72	0.81	0.79

Table 3: Human evaluation scores and Correlations.

6 Ablation study

Varying the layers: To validate the contribution of the attention patterns of various layers, we generate the captions with the first, middle, last, max, and mean attention layers. We find that mean-layer attention scores outperform other variants, as shown in Appendix C, Table 9.

Varying the amplification factor: As depicted in Tables 2 and 11, we vary the amplification factor (k) and for both datasets, k=1 resulted in better performance.

Varying the decoding strategy: We experiment with different decoding strategies such as Top-k sampling, Top-p (nucleus) sampling, and beam search, combining all of these resulted in better performance compared to individual decoding strategies as shown in Table 2. A detailed justification for using hybrid decoding is given in Appendix I.

7 Qualitative Assessment

To assess the quality of image captions generated by our AGIC approach, we perform human evaluations and error analysis on 500 random samples.

7.1 Human Evaluation and Error Analysis

Our human evaluation focuses on ***Correctness, Completeness, Relevancy*** metrics, and each is scored on a scale of 1 to 5.

Table 3 presents the average human evaluation scores and inter-metric correlations. Captions are rated highly overall, with Relevancy achieving the highest average (4.26). Correlation analysis shows strong alignment among metrics, particularly between Completeness and Relevancy (r $\approx$ 0.81), suggesting consistency in human assessment of caption quality. Captions with the highest human scores often demonstrate precise object identification, contextually appropriate action descriptions, and inclusion of secondary and background scene details. As detailed in Table 4, we compute the inter-rater reliability using Intra-class correlation coefficient as per ICC (Koo and Li, 2016) and observe that the evaluator exhibits moderate to good agreement for Flickr8k, Flickr30k, and MSCOCO data samples.

	Relevancy	**Correctness**	**Completion**
Flickr8k	0.84	0.87	0.82
Flickr30k	0.82	0.85	0.87
MSCOCO	0.86	0.90	0.84

Table 4: Inter rater reliability (ICC).

We performe a focused error analysis on 500 random images from Flickr8k, Flickr30k, and MSCOCO. We consider four categories: ***hallucination***, ***omission***, ***irrelevance***, and ***ambiguity***. More details on the definitions of categories considered above are given in the Appendix E. Table 5 summarizes the counts. Although AGIC improves relevance overall, we still observe occasional omissions across datasets and a higher rate of hallucination on Flickr30k, indicating room for strengthening grounding and coverage of secondary details. Furthermore, illustrative examples for each of the above error categories are provided in Figure 7.

	Hallucination	Omission	Irrelevance	Ambiguity
Flickr8k	38	82	18	36
Flickr30k	55	94	20	31
MSCOCO	40	73	17	33

Table 5: Error analysis of AGIC model.

Method	**Y/N**	**Num**	**Oth**	**All**
BLIP-2-opt-2.7B	0.743	0.344	0.509	0.532
AGIC (CLIP+BLIP2)	**0.807**	**0.435**	**0.548**	**0.596**

Table 6: VQA v2 results on 500-image subset.

8 Generalization beyond Captioning

To assess whether attention-guided amplification generalizes beyond captioning, we apply AGIC to the Visual Question Answering task on VQA v2 (Antol et al., 2015). We evaluate on a randomly sampled set of 500 validation images, covering the standard answer categories: *Yes/No* (203 samples), *Number* (73), and *Other* (224). We use the official VQA accuracy, $\mathrm{Acc} = \min\left(\frac{\text{\# humans who gave the predicted answer}}{3}, 1\right)$, computed against the 10 human references per question. Table 6 compares the baseline BLIP-2-opt-2.7B with our method (AGIC with CLIP attention and BLIP2 decoding). AGIC improves all categories and yields a **+6.4** absolute point gain in overall accuracy (0.596 vs. 0.532), indicating that the proposed attention-guided preprocessing is beneficial for downstream reasoning tasks that require strong visual grounding.

9 Conclusion

We propose *AGIC*, an unsupervised image captioning framework that uses attention weights from different layers of a Vision Transformer to amplify contextually relevant regions for enhanced caption generation. *AGIC* integrates an attention-guided amplification process with a hybrid decoding strategy to balance grounding and diversity with relevance to the image. Experimental results on standard benchmarks show that AGIC performs better than supervised and unsupervised baselines on several metrics, providing the most relevant captions.

10 Limitations

While AGIC exhibits impressive performance when compared to several state-of-the-art models, it has the following limitations. 1) Due to reliance on attention allocation patterns, the applicability of the approach is restricted to only open-source models. 2) AGIC is highly sensitive to the amplification factor; this suggests that even slight over-amplification can be detrimental, diluting focus rather than enhancing it.

References

Israa Al Badarneh, Bassam H Hammo, and Omar Al-Kadi. 2025. An ensemble model with attention based mechanism for image captioning. *Computers and Electrical Engineering*, 123:110077.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Introducing our multimodal models.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8307–8316.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587.

Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4125–4134.

Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. 2019. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10323–10332.

Dan Guo, Yang Wang, Peipei Song, and Meng Wang. 2020. Recurrent relational memory network for unsupervised image captioning. *arXiv preprint arXiv:2006.13611*.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. *Advances in neural information processing systems*, 32.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Ukyo Honda, Yoshitaka Ushiku, Atsushi Hashimoto, Taro Watanabe, and Yuji Matsumoto. 2021. Removing word-level spurious alignment between images and pseudo-captions in unsupervised image captioning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3692–3702, Online. Association for Computational Linguistics.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643.

Mingi Jung, Saehyung Lee, Eunji Kim, and Sungroh Yoon. 2025. Visual attention never fades: Selective progressive attention recalibration for detailed image captioning in multimodal large language models. *arXiv preprint arXiv:2502.01419*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Iro Laina, Christian Rupprecht, and Nassir Navab. 2019. Towards unsupervised image captioning with shared multimodal embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7414–7424.

Saehyung Lee, Seunghyun Yoon, Trung Bui, Jing Shi, and Sungroh Yoon. 2024. Toward robust hyper-detailed image captioning: A multiagent approach and dual evaluation metrics for factuality and coverage. *arXiv preprint arXiv:2412.15484*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Xu Liang, Chen Li, and Lihua Tian. 2024. Generative adversarial network for semi-supervised image captioning. *Computer Vision and Image Understanding*, 249:104199.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Qiang Liu, Xinlong Chen, Yue Ding, Shizhen Xu, Shu Wu, and Liang Wang. 2025. Attention-guided self-reflection for zero-shot hallucination detection in large language models. *arXiv preprint arXiv:2501.09997*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228.

Yiwei Ma, Jiayi Ji, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. 2023. Towards local visual modeling for image captioning. *Pattern Recognition*, 138:109420.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10971–10980.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Zhan Shi, Hui Liu, and Xiaodan Zhu. 2021. Enhancing descriptive image captioning with natural language inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 269–277, Online. Association for Computational Linguistics.

Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. 2020. Improving image captioning with better use of caption. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7454–7464, Online. Association for Computational Linguistics.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.

A AGIC Pipeline Overview with an Illustrative Example

The AGIC framework operates in three key stages:

- **Attention Weights Extraction:** In the first stage, the model identifies and extracts attention weights corresponding to semantically relevant objects within the image. As illustrated in Figure 3, the mean attention is concentrated around the object of interest, in this case, the ‘dog’, indicating strong model focus.
- **Image Amplification:** Using the extracted attention weights, the image is selectively amplified by applying an amplification factor to regions associated with high attention. This

step enhances salient features, making them more prominent for the captioning model.

- **Caption Generation:** Finally, the amplified image is input into a Vision Transformer (ViT) to produce a detailed and contextually accurate caption that aligns closely with the amplified regions of interest.

B Experimental Setup

We conduct all our experiments on Amazon Web Services (*AWS*) cloud server, Amazon elastic compute cloud (*EC2*) instance. In EC2, we initiate an instance for accelerated Computing. The specifications are *g6e.xlarge* instance, which provides 3rd generation AMD EPYC processors (*AMD EPYC 7R13*), with a NVIDIA L40S Tensor Core GPU with 48 GB GPU memory, and 4x vCPU with 150 GB memory. All the hyperparameter details are listed in Table 7. The model sources are detailed in Table 8.

Hyperparameter	**Value**
`top_p`	0.9
`num_return_sequences`	1
`num_beams`	5
`max_new_tokens`	30
`top_k`	50
`early_stoping`	True
`do_sample`	True
`temperature`	1.0
`K (amplification factor)`	-5, -3, -1, 0, 1, 3, 5

Table 7: Hyperparameters details.

Model	**Source**
fuyu-8b	`https://huggingface.co/adept/fuyu-8b`
blip2-opt-2.7b	`https://huggingface.co/Salesforce/blip2-opt-2.7b`
llava-1.5-7b-hf	`https://huggingface.co/llava-hf/llava-1.5-7b-hf`
clip-vit-large-patch14	`https://huggingface.co/openai/clip-vit-large-patch14`
Qwen2.5-VL-7B-Instruct	`https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct`

Table 8: Source of Huggingface models.

C More details on ablation study

C.1 Effect of various layers attention visualization

We perform various experiments to find out the optimal layer to extract the attention visualizations and experiment with First, Middle, Last, Max, and Mean of attention layers. If there are L attention layers, then the first attention layer can be represented as a^0, and middle layer will be represented as a^M where $M = L//2$, the last layer will be a^{L-1}, the layer with maximum attention can simply be determined by the following $a^{max} = \max_l a^l$, and finally the mean attention of all the layers can be computed as $a^{mean} = \frac{1}{L}\sum_{l=0}^{L-1} a^l$. We obtain the caption for the image in all the layers mentioned earlier. We have evaluated our method AGIC on benchmark datasets: Flickr8k and Flickr30k, using standard image captioning metrics: BLEU-1 to BLEU-4 (B1–B4), METEOR, ROUGE-L, CIDEr, and SPICE. In Table 9, we report performance under different attention layers, decoding strategies. On both datasets, the BLIP2, a 2.7 billion-parameter model, performs well. Still, its performance is significantly improved with the help of hybrid decoding strategies and the attention-guided amplification technique.

C.2 Effect of Decoding Strategies

On both Flickr8k and Flickr30k datasets, combining the decoding strategies: 1) Top-k sampling, 2) Top-p (nucleus) sampling, and 3) Beam Search yields substantial performance gains across all evaluation metrics compared to the base model. As shown in Table 2, the CIDEr score on Flickr8k improves from 0.19 (base) to 0.73 (All), and on Flickr30k from 0.43 to 0.60.

Among these, beam search plays a pivotal role in enhancing caption quality, contributing the most significant improvement, particularly in BLEU and CIDEr scores. This can be attributed to its ability to maintain the most probable caption candidates across decoding steps, ensuring better fluency and grammatical correctness. However, beam search alone tends to produce less diverse captions. The inclusion of Top-p and Top-k sampling addresses this limitation by introducing diversity, which improves recall-based metrics such as ROUGE-L and METEOR when used in conjunction with beam search.

C.3 Effect of Amplification Factor k

The amplification factor k controls how many times the resulting attention maps from the first pass are integrated into the image tensor for the amplification of important and relevant features in the image. As seen in Table 2, $k = 1$ yields the best results. This shows that using only one prominent attention map effectively enhances key visual regions without introducing noise. As k increases, performance begins to drop as observed on Flickr8k, CIDEr drops from 0.734 ($k = 1$) to 0.565 ($k = 5$). This is

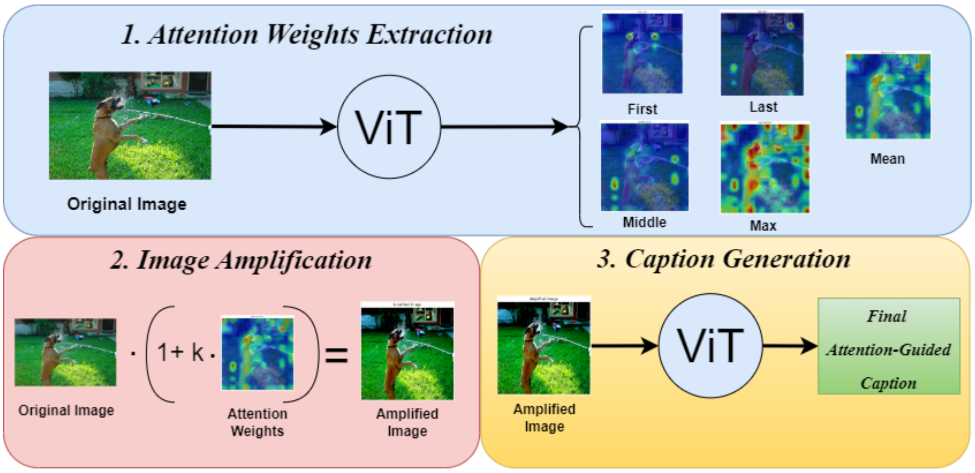


Figure 3: Attention guided image captioning (AGIC) pipeline.

Dataset ↓	Model ↓	Layer ↓	B1	B2	B3	B4	METEOR	R-L	CIDEr	SPICE	CLIPScore
Flickr8k	CLIP + BLIP-2-OPT	First	0.65	0.48	0.35	0.25	0.44	0.43	0.71	0.20	0.82
		Mid	0.64	0.47	0.34	0.24	0.43	0.43	0.70	0.19	0.81
		Last	0.65	0.48	0.35	0.25	0.44	0.43	0.70	0.20	0.82
		Max	0.63	0.46	0.33	0.23	0.42	0.42	0.68	0.19	0.80
		Mean	**0.69**	**0.52**	**0.39**	**0.29**	**0.48**	**0.44**	**0.78**	**0.23**	**0.85**
Flickr30k	CLIP + BLIP-2-OPT	First	0.63	0.46	0.33	0.23	0.42	0.39	0.58	0.17	0.81
		Mid	0.62	0.45	0.32	0.22	0.41	0.38	0.57	0.16	0.80
		Last	0.63	0.46	0.33	0.23	0.42	0.39	0.58	0.17	0.81
		Max	0.61	0.44	0.31	0.21	0.40	0.37	0.55	0.15	0.79
		Mean	**0.67**	**0.51**	**0.37**	**0.27**	**0.45**	**0.41**	**0.65**	**0.20**	**0.84**
MSCOCO	CLIP + BLIP-2-OPT	First	0.68	0.51	0.38	0.28	0.48	0.29	1.02	0.21	0.82
		Mid	0.67	0.50	0.37	0.27	0.47	0.28	1.00	0.20	0.81
		Last	0.68	0.51	0.38	0.28	0.48	0.29	1.02	0.21	0.82
		Max	0.66	0.49	0.36	0.26	0.46	0.27	0.98	0.19	0.80
		Mean	**0.72**	**0.55**	**0.41**	**0.31**	**0.51**	**0.31**	**1.09**	**0.23**	**0.85**

Table 9: Evaluation metrics BLEU (1-4), METEOR, ROUGE-L, CIDEr, SPICE, and CLIPScore on three datasets Flickr8k, Flickr30k and MSCOCO, for the captions from different layer strategies.

because including more attention weights tends to dilute the focus and overemphasize not only on the relevant regions but also the less relevant or irrelevant regions, which leads to a decline in caption generation quality. The decrease in performance as the amplification factor increases is consistent across metrics, confirming that excessive amplification causes distraction in the model's attention.

C.4 Effect of caption length

To explore how caption length correlates with evaluation metrics, we conduct a case study. Our observations show that longer captions generally yield slightly higher scores in CIDEr, METEOR, and SPICE, likely due to their ability to capture more contextual details. By grouping captions into length buckets (1–5, 6–10, 11–15), we find that:
a) Captions with 6–10 words strike a good balance between precision (BLEU) and semantic adequacy (SPICE, METEOR).
b) Very short captions (<5 words) often omit key details, resulting in lower completeness and SPICE scores.
c) Very long captions (>15 words) do not consistently improve performance and may introduce noise or redundancy, affecting correctness.
Overall, captions of moderate length (around 7–12 words) tend to achieve the best scores across both human and automatic evaluations.

D More detail on Human Evaluation

To perform the human evaluation, we have given 50 samples each to evaluators and instruct them to assess each of the parameter based on the following definition.
Correctness: which assesses the accurate identification of all prominent objects;
Completeness: which evaluates whether all rele-

Dataset	Approach	B1	B2	B3	B4	R-L	METEOR	CIDEr	SPICE	CLIPScore
Flickr30k	BLIP2-opt-2.7b (Li et al., 2023)	0.406	0.260	0.164	0.103	0.306	0.245	0.083	0.083	0.691
	LLaVA-1.5-7B (Liu et al., 2023)	0.471	0.316	0.213	0.145	0.348	0.070	0.234	0.096	0.739
	Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	0.454	0.277	0.166	0.100	0.285	0.055	0.254	0.088	0.715
	Fuyu-8B (Bavishi et al., 2023)	0.613	0.427	0.288	0.193	0.387	0.166	0.396	0.134	0.751
	BRNN (Karpathy and Fei-Fei, 2015)	0.559	0.340	0.194	0.113	0.317	0.153	0.499	0.109	0.740
	R^2M (Guo et al., 2020)	0.531	0.328	0.192	0.117	0.359	0.137	0.181	0.083	0.687
	LSTNet (Ma et al., 2023)	0.654	0.493	0.350	0.224	0.410	0.197	0.633	0.310	0.782
	AoANet (Huang et al., 2019)	0.613	0.421	0.287	0.193	0.389	0.227	0.612	0.143	0.765
	CapMAS (Lee et al., 2024)	0.168	0.089	0.046	0.024	0.160	0.182	0.009	0.121	0.565
	SPARC (Jung et al., 2025)	0.403	0.225	0.117	0.060	0.398	0.188	0.030	0.150	0.698
	AGIC + LLaVA-1.5-7B	0.629	0.449	0.315	0.216	0.382	0.215	0.553	0.153	0.805
	AGIC + BLIP2-opt-2.7b	0.650	0.469	0.336	0.235	0.232	0.399	0.601	0.164	0.819
	AGIC + CLIP + BLIP2	0.672	0.517	0.375	**0.274**	**0.459**	**0.414**	**0.653**	0.209	**0.846**

Table 10: Image captioning results for various models on Flickr30k; B1 to B4 refer to BLEU scores and R-L: ROUGE-L score.

		Config.	BLEU	R-L	CIDEr	SPICE	CLP
Decoding strategy	Flickr30k	CLIP-BLIP2	0.10	0.28	0.08	0.08	0.67
		CLIP-BLIP2+Top-k	0.07	0.25	0.24	0.09	0.65
		CLIP-BLIP2+Top-p	0.10	0.28	0.30	0.11	0.71
		CLIP-BLIP2+Beam	0.22	0.37	0.55	0.15	0.75
		CLIP-BLIP2+Top-p+Top-k	0.16	0.32	0.44	0.13	0.72
		CLIP-BLIP2+Top-p+Beam	0.26	0.39	0.60	0.17	0.76
		CLIP-BLIP2+Top-k+Beam	0.26	0.38	0.60	0.17	0.75
		CLIP-BLIP2+Top-p+Top-k+Beam	**0.27**	**0.45**	**0.65**	**0.20**	**0.84**
Amplification factor	Flickr30k	AGIC (k=-5)	0.15	0.32	0.33	0.11	0.74
		AGIC (k=-3)	0.16	0.32	0.35	0.12	0.74
		AGIC (k=-1)	0.16	0.33	0.38	0.13	0.74
		AGIC (k=0)	0.21	0.37	0.52	0.16	0.80
		AGIC **(k=1)**	**0.27**	**0.45**	**0.65**	**0.20**	**0.84**
		AGIC (k=3)	0.21	0.37	0.52	0.15	0.80
		AGIC (k=5)	0.21	0.37	0.55	0.15	0.80

Table 11: **Top:** Performance variation across decoding strategies. ‘All’ represents the combined strategy (Beam Search + Top-k + Top-p). **Bottom:** Performance comparison with varying amplification factors.

vant image content, including objects, attributes, and actions, is comprehensively covered;
Relevancy: which measures the pertinence and salience of the information presented in relation to the image’s primary content.

E More Details on Error Analysis

As mentioned in Subsection 7.1, to evaluate the AGIC approach, we analyze 500 image samples from the Flickr8k, Flickr30k, and MSCOCO datasets by considering four error categories: Hallucination, Omission, Irrelevance, and Ambiguity.

1. **Hallucination:** Hallucination refers to the inclusion of objects, actions, or details in the caption that are not actually present in the image. This typically occurs when models rely too heavily on prior knowledge or context rather than visual evidence, leading to inaccurate or fabricated descriptions.

2. **Omission:** Omission occurs when the caption fails to mention relevant visual elements that are clearly present in the image. While it may describe the main subject, it might overlook important background details or secondary objects, resulting in an incomplete representation of the scene.

3. **Irrelevance:** Irrelevance indicates a mismatch between the generated caption and the visual content of the image. It occurs when the caption includes information that is off-topic or not grounded in the image, reflecting poor alignment between vision and language.

4. **Ambiguity:** Ambiguity arises when the caption uses vague or generic terms to describe entities in the image, for example, using “someone” or “a person” without specifying characteristics such as gender, role, or activity. This lack of clarity can make the caption less informative or confusing.

F Mapping 1D Attention to 2D Spatial Layout

The attention scores obtained during the *Attention Weights Extraction* step (as mentioned in Section 3.1) are one-dimensional vectors, representing the mean attention across all heads in the model. Since these scores are in 1D, we must transform them into a 2D spatial layout to apply amplification to the image. To achieve this, we utilize attention weights associated with the [CLS] token, which serves as a global aggregator of information across the image. The resulting 1D attention vector is then reshaped into a 2D attention map corresponding to the spatial dimensions of the preprocessed image (typically forming a square grid).

G Prompt for zero-shot captioning

We perform image captioning using a zero-shot prompting strategy with vision-language models (VLLMs). The prompt template shown in Figure 4

Zero-shot Prompt

```
 {"role": "user", "content": [
{"type":      "image",    "image":
<image>},
{"type": "text", "text": "Generate
 a COCO-style caption focused on the
 main objects and their interactions.
 Avoid names; keep it concise and
 grammatically correct."}
 ]}
```

Figure 4: Zero-shot prompt used for caption generation.

is used for zero-shot caption generation with models such as LLaVA, Qwen, and Fuyu.

H Algorithm

Algorithm 1 The AGIC Approach

```
Input: IC model $\mathcal{M}$, image $I$, ref cap $C_{\text{ref}}$
Output: Amplified cap $C_{\text{attn}}$, eval scores $E$
1. Preprocess $I$ to obtain tensor $I_{\text{tensor}}$.
2. Pass $I_{\text{tensor}}$ through $\mathcal{M}$'s encoder to extract attention maps $A^{l,h}$ for all layers $l$ and heads $h$.
3. Compute patch attention scores $a_i^l$ and reshape into $a^l(x,y)$.
4. For each strat s $\in$ {first, mid, last, max, mean}:
    • Determine $a^{\text{s}}(x,y)$ as acc to 3.2:
        first: $a^0(x,y)$,
        mid: $a^M(x,y)$, where $M = L//2$
        last: $a^L(x,y)$,
        max: $\max_l a^l(x,y)$,
        mean: $\frac{1}{L}\sum_l a^l(x,y)$
    • Amp: $I_a{}^{\text{s}}(x,y) = I_o(x,y)\cdot(1 + k\cdot a^{\text{s}}(x,y))$
    • Gen caption $C_{\text{attn}}[\text{s}] = \mathcal{M}(\hat{I}^{\text{s}})$ acc to 3.3.
5. Evaluate $C_{\text{attn}}$ using BLEU-1 to BLEU-4, ROUGE-L, METEOR, CIDEr, SPICE, and CLIP-Score w.r.t. $C_{\text{ref}}$ acc to 4.2.
6. Return: $C_{\text{attn}}, E$
```

I Justification for Hybrid Decoding

We justify hybrid decoding using token-level decoding dynamics and a grounding check. Figure 5 plots the top-K next-token probabilities at the steps where greedy decoding and hybrid decoding differ, together with the next-token entropy. In our example, both methods agree on the high-confidence prefix ("a man sleeping on a bench"), while the divergence occurs at higher-entropy steps where multiple continuations are plausible; greedy decoding terminates early, whereas hybrid decoding continues and adds the scene context ("in a train station"). This shows that hybrid decoding is most beneficial precisely at uncertain decision points, where it preserves plausible alternatives and selects the continuation that better matches the image. Figure 6 then provides qualitative evidence that the added phrase is visually supported: CLIP patch-to-text similarity maps localize "man" and "bench" to the corresponding foreground regions and activate "train station/station" primarily on background structures consistent with a station environment. Together, these results support that hybrid decoding improves image-to-text alignment by making better choices at uncertain decoding steps and that the resulting extra details are grounded in the image.

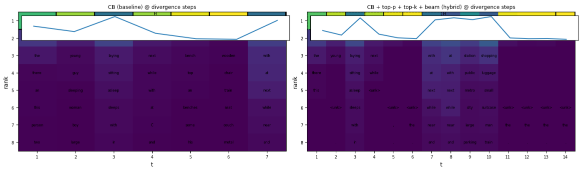


Figure 5: Hybrid decoding diverges at uncertain token decisions.

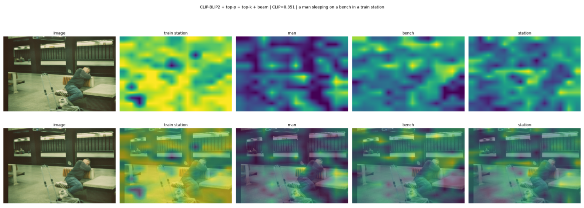
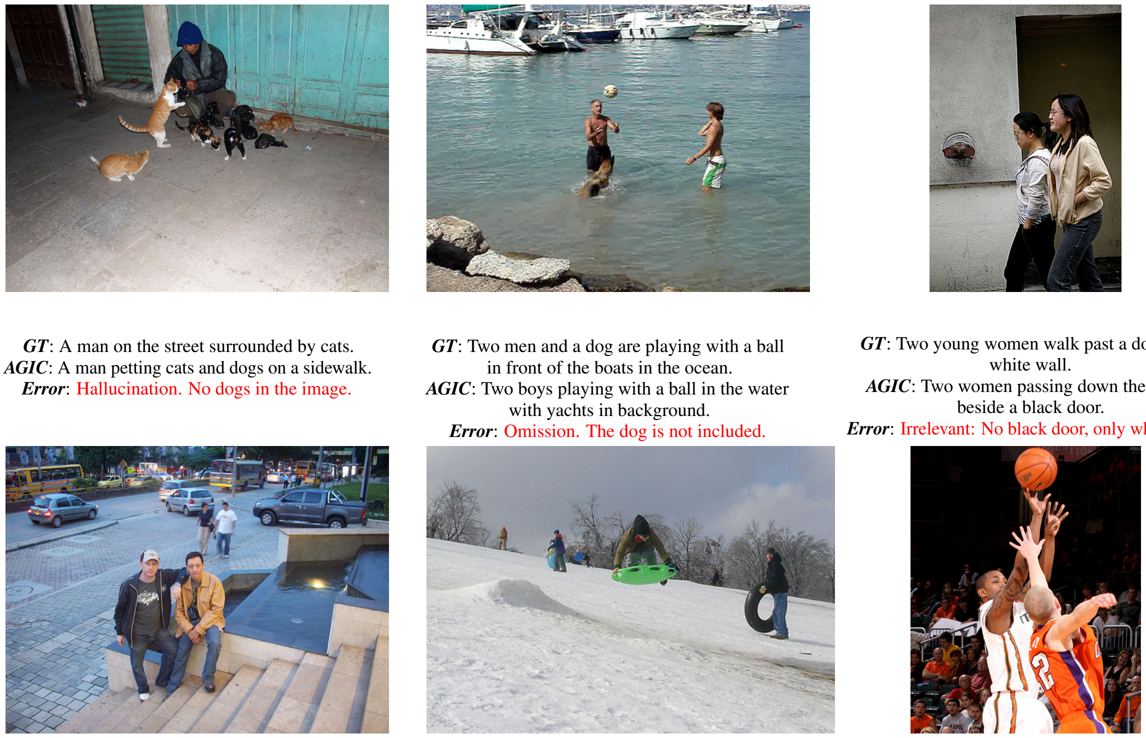


Figure 6: Heatmaps showing that the caption words are visually grounded.



GT: A man on the street surrounded by cats.
AGIC: A man petting cats and dogs on a sidewalk.
Error: Hallucination. No dogs in the image.

GT: Two men and a dog are playing with a ball in front of the boats in the ocean.
AGIC: Two boys playing with a ball in the water with yachts in background.
Error: Omission. The dog is not included.

GT: Two young women walk past a door in a white wall.
AGIC: Two women passing down the street beside a black door.
Error: Irrelevant: No black door, only white wall.

GT: Two guys are posing next to each other on the steps of a building with traffic behind them.
AGIC: Two men sitting on steps.
Error: Vague: The caption can be more descriptive besides sitting on steps.

GT: A child sleds over a mound of snow as others watch him.
AGIC: A person is watching a man snowboard down a hill.
Error: Ambiguity: The caption mentioned it as just the person and a man, but couldn't identify the gender and age.

GT: A basketball player shooting while another player is trying to block his shot.
AGIC: a basketball player trying to block the ball from going into the basket thrown by another player.
Error: All Correct with no errors.

Figure 7: Qualitative error analysis of image captions generated by the AGIC model compared to ground truth (GT) descriptions. Each example highlights a specific type of error: hallucination, omission, irrelevance, vagueness, and ambiguity. One example demonstrates a correct caption with no notable issues.