Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Retrieval Free Context Augmentation for Image Captioning: Learning from Self-Generated Beam Hypotheses

Ansar HANI
University of Sfax
Najiba TAGOUGUI
University of Sfax
Mouhammed Laid ABIMOULOUD
mohamed.abimouloud@enis.tn
University of Sfax
Abdelmalik Taleb-Ahmed
Univ. Polytechnique Hauts-de-France
Monji KHERALLAH
University of Sfax

Research Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11184261/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Retrieval Free Context Augmentation for Image Captioning: Learning from Self-Generated Beam Hypotheses

Ansar HANI[1], Najiba TAGOUGUI[2†], Mouhammed Laid ABIMOULOUD[3*†], Abdelmalik Taleb-Ahmed[4†], Monji KHERALLAH[5†]

[1*]National Engineering School of Sfax, University of Sfax, Sfax, Tunisia.
[2]Higher Institute of Computer Science and Multimedia, University of Sfax, Sfax, Tunisia.
[2]Toxicology, Environmental Microbiology and Health,Faculty of Science, University of Sfax, Sfax, Tunisia.
[3]National Engineering School of Sfax, University of Sfax, Sfax, Tunisia.
[4]Univ. Polytechnique Hauts-de-France, Valenciennes, France.
[5]Faculty of Sciences, University of Sfax, Sfax, Tunisia.
[1 2 3 5] Advanced Technologies for Environment and Smart Cities (ATES Unit), Sfax University, Sfax, Tunisia.

*Corresponding author(s). E-mail(s): mohamed.abimouloud@enis.tn;
Contributing authors: ansar.hani@stud.enis.tn; najiba.tagougui@isims.usf.tn; abdelmalik.taleb-ahmed@uphf.fr; monji.kherallah@fss.usf.tn;
†These authors contributed equally to this work.

Abstract

Vision-language models (VLMs) generate image captions by grounding visual perception in semantically coherent natural language, a process whose quality is largely determined by the supervision signal used during training rather than by model capacity alone. Existing Retrieval-Augmented Generation (RAG), approaches rely exclusively on fixed human-annotated captions, leaving unexploited the rich, diverse hypotheses that the model itself generates during inference. To address this limitation, we introduce CAG-Cap, a retrieval-free caption-refinement framework inspired by the offline caching principle of

Cache-Augmented Generation (CAG). Unlike conventional CAG, where cached knowledge is injected into the inference context, CAG-Cap uses cached information exclusively as auxiliary sequence-level supervision during training, with no additional inference-time parameters or retrieval operations. CAG-Cap operates in two stages. In Stage A, a baseline captioner comprising a frozen SigLIP 2-B/16 vision encoder and a four-layer Transformer decoder is trained under next-token prediction to convergence, a caption cache is then constructed by running beam search over the training set and retaining the top-$\boldsymbol{k}$ hypotheses per image as self-generated pseudo-labels. In Stage B, the model is fine-tuned under a dual-loss objective that combines ground-truth negative log-likelihood (NLL) supervision with a CAG loss, defined as the mean NLL over all $\boldsymbol{k}$ cached hypotheses, weighted by a tunable coefficient $\boldsymbol{\lambda}_{\mathrm{CAG}}$. Experiments on MS COCO 2017 and Flickr30k demonstrate the effectiveness of CAG-Cap. The optimal configuration on MS COCO 2017 achieves a CIDEr of **2.2106**, a BLEU-4 of **0.3902**, and a SPICE of **0.4115**, on Flickr30k, it attains a BLEU-4 of **0.3218** and a METEOR of **0.4990**. CAG-Cap establishes a principled, retrieval-free alternative to RAG for training-time caption supervision, offering a scalable pathway toward self-improving VLMs. The source code is accessible: https://github.com/abimouloud/CAG-Cap

Keywords: Context-Augmented Generation (CAG), Image Captioning, Vision-Language Models (VLMs), Retrieval-Augmented Generation (RAG), MS COCO 2017, Flickr30k.

1 Introduction:

The automatic generation of natural language descriptions from visual inputs lies at the intersection of computer vision and natural language processing (NLP) and serves as the foundation for a wide range of image captioning applications, including assistive technology for the visually impaired, content-based image retrieval, visual question answering, and autonomous scene understanding [1]. The encoder-decoder architectures employed CNNs such as VGGNet and ResNet to extract spatial feature maps, which were then consumed by LSTM-based decoders to generate captions token by token, the integration of attention over convolutional feature grids, demonstrated that selectively weighting spatial regions at each decoding step could substantially improve lexical grounding and semantic coherence [2]. The subsequent adoption of the Transformer architecture and large-scale contrastive pre-training, exemplified by models such as CLIP [3], BLIP [4], and SigLIP [5], established VLMs as the dominant paradigm for image captioning, enabling zero-shot generalisation and instruction following capabilities that far exceed those of task specific predecessors [6]. More recently architecture, SigLIP 2 [7], a family of multilingual vision-language encoders released by Google DeepMind, further advanced visual representation quality by integrating captioning-based pre-training into a unified recipe, establishing a new standard for zero-shot classification and feature extraction in VLMs. Despite these advances, the training supervision for caption generation remains almost universally anchored to

static human-annotated reference captions, an assumption that constrains the diversity and self-improving capacity of the learned representations.

RAG has emerged as a compelling strategy to complement the parametric knowledge of language models with dynamically retrieved external evidence [8]. In the multimodal domain, RAG has been extended to VLMs through frameworks that retrieve visually similar images, related captions, or structured knowledge at inference time to condition the generation process [9]. SmallCap [10] demonstrated that even a lightweight captioner could achieve competitive performance by prompting a language model with a small set of retrieved captions, while more works such as DIR [11] and MIRA-CAP [12] pushed the retrieval-augmented paradigm further by incorporating comprehensive semantic understanding and episodic memory into the retrieval loop. Within multimodal large language models (MLLMs), RAG has been applied to visual question answering [13], hallucination mitigation [14], consistently demonstrating that external context enriches both factual accuracy and semantic alignment.

Despite its benefits, retrieval-augmented image captioning suffers from several practical drawbacks that limit its real-world applicability. The system is tightly coupled to an external retrieval index at every inference step, such that semantically misaligned candidates can actively mislead the model into generating hallucinated captions that reflect database artefacts rather than actual visual content [15]. Beyond correctness, querying a dense vector index and rencoding retrieved documents at run time adds measurable latency, making RAG-based systems ill-suited for applications with strict response time requirements, while growing retrieval corpora further compound this problem through unfavourable scaling of index construction, storage, and search costs, with retrieved passages frequently exceeding the context limits of current VLMs [16]. These limitations collectively motivate the need for a retrieval-free alternative that preserves the contextual richness of augmentation without incurring its operational costs, which is precisely the role filled by CAG in the proposed framework.

Context-Augmented Generation (CAG) [17], has recently been proposed as a principled retrieval-free paradigm for knowledge integration in language models. In its original formulation, CAG preloads a curated knowledge base into the model's extended context window and caches the resulting representations, eliminating real-time retrieval whilst maintaining context relevance. Empirical comparisons show that CAG can reduce generation latency by an order of magnitude relative to RAG [18, 19]. Critically, however, the application of CAG principles to training-time supervision in visual captioning rather than merely to inference time context conditioning remains unexplored.

We introduce CAG-Cap, a two-stage CAG framework for self-supervised caption refinement in VLMs.The framework builds upon a baseline captioner that pairs a frozen SigLIP 2-B/16 vision encoder with a four-layer Transformer decoder trained under next-token prediction. Following convergence of Stage A, a caption cache is constructed by executing length normalised beam search over the entire training set and retaining the top-k hypotheses per image. In Stage B, the model is fine-tuned under a dual-loss objective that combines standard ground-truth supervision with a complementary context-augmented loss computed as the mean negative log-likelihood over all k cached beam hypotheses per image, weighted by a tunable coefficient λ_{CAG}. The

same encoder-decoder architecture is retained in stage B, and the caption cache is generated entirely from the model's own beam search outputs, rendering the framework fully self contained and independent of any external retrieval system.
The proposed CAG-Cap framework offers several advantages over RAG techniques:

1. We improve caption quality, the dual-loss formulation exposes the decoder to diverse, internally generated hypotheses that collectively span a broader region of the caption space than any single ground-truth annotation, improving surface-level and semantic metrics.
2. We enhance contextual understanding by regularising training with beam hypotheses that reflect the model's own prior distribution, CAG-Cap encourages the decoder to internalise richer visual-linguistic correspondences.
3. We improve semantic consistency, unlike retrieval, which can introduce corpus-level noise when retrieved documents conflict with the query image, CAG pseudo-labels are generated conditioned on the actual training image, guaranteeing visual grounding by construction.
4. We reduce retrieval dependency and improve efficiency, the cached hypotheses are precomputed once at negligible cost and require no index at inference time and memory footprint at the level of the standalone baseline.

Building upon these advantages, the main contributions of this work are as follows:

1. We introduce Context-Augmented Generation (CAG) as a retrieval-free supervision strategy that repurposes the model's own beam search hypotheses as structured pseudo-label targets, providing diverse training-time regularisation without any external retrieval component, auxiliary database, or additional learnable parameters.
2. We propose a principled two-term training objective that combines ground-truth negative log likelihood (NLL) supervision with a context-augmented NLL loss computed over cached beam hypotheses, enabling the model to learn from its own diverse outputs alongside reference captions.
3. By confining all augmentation to the training phase and populating the caption cache exclusively from the model's own outputs, CAG-Cap achieves consistent caption quality improvements at inference-time and GPU memory identical to those of the standalone beam search baseline, establishing a scalable and deployment friendly pathway toward retrieval-free, self-improving VLMs.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents the proposed CAG-Cap framework and the experimental setup. Section 4 reports the evaluation metrics, implementation details, experimental results, and ablation studies. Section 5 discusses the results and provides a comparative analysis with existing methods. Finally, Section 6 concludes the paper, discusses its limitations, and outlines future research directions.

2 Related Work

Image captioning has evolved significantly from recurrent neural network (RNN)-based approaches to Transformer architectures and, more recently, to large-scale VLMs that integrate powerful visual encoders with large language models. The introduction of the Transformer architecture by Vaswani et al. [20] marked a turning point in image captioning by replacing recurrent structures with self-attention mechanisms capable of modelling long-range dependencies between image regions and textual tokens. The growing maturity of Transformer-based captioning has been summarized by recent survey papers. Al-Malla, et al. [2] reviewed more than 170 deep learning image captioning studies published between 2018 and 2025, concluding that supervised caption generation still suffers from overconfident predictions and limited caption diversity. Building on this foundation, several studies explored increasingly sophisticated Transformer-based captioning models. Lv et al. [21] proposed the Hierarchical and Sequential Transformer (HIST), which learns visual representations at multiple feature granularities while introducing a sequential enhancement module to improve sentence coherence. Experiments on MS COCO and Flickr30K demonstrated consistent improvements over earlier Transformer models, highlighting the importance of hierarchical visual representations. Similarly, Abdulgalil et al. [22] identified multimodal large language models as the dominant research direction while emphasizing their heavy dependence on massive pre-training datasets and expensive computational resources. These observations suggest that improving caption quality without increasing model complexity remains an important research challenge. Recent studies have also demonstrated that captioning objectives themselves produce highly transferable visual representations. Tschannen et al. [**?**] showed that vision encoders trained through caption generation perform competitively with contrastive learning approaches on multiple downstream vision tasks. This finding motivated the development of SigLIP 2 [7], which combines sigmoid-loss contrastive learning, caption generation, masked prediction and self-distillation within a unified training framework. To reduce the dependence on massive end-to-end training, another research direction incorporates external knowledge through Retrieval-Augmented Generation (RAG) [8]. Instead of relying solely on visual features, RAG-based methods retrieve semantically similar image-caption pairs from external databases and use the retrieved information to guide caption generation. Ramos et al. [10] proposed SmallCap, which combines a frozen CLIP encoder with GPT-2 while prepending retrieved captions to the decoder input. This lightweight framework significantly reduces the number of trainable parameters and generalizes well across datasets. However, later analysis by Li et al. [15] revealed that SmallCap often copies frequently occurring words from retrieved captions, making the model sensitive to retrieval quality and reducing robustness under noisy retrieval conditions. Other subsequent studies attempted to improve retrieval augmentation. EVCap [23] replaces full retrieved captions with retrieved object concepts stored in an external visual-name memory, enabling better captioning of unseen objects while reducing redundant textual information. Similarly, MeaCap [24] maintains a textual memory and retrieves only image-relevant concepts to reduce hallucinations in zero-shot caption generation. Although these approaches demonstrate notable improvements, they all require external retrieval databases during

inference, introducing additional storage requirements, indexing overhead and retrieval latency. In a related direction, Pourkeshavarz et al. [25] proposed a stacked cross-modal feature consolidation (SCFC) attention network that iteratively fuses region-level visual features with context-aware semantic attributes through a multi-step compounding function, coupled with a dedicated SCFC-LSTM decoder. Evaluated on MS-COCO and Flickr30K, SCFC showed that progressively consolidating visual and attribute information across multiple reasoning steps yields more discriminative caption representations than single-step attention fusion. A separate direction has pursued caption quality through scaling large language models rather than refining the training objective. Zhu et al. [26] introduced MiniGPT-4, which aligns a frozen ViT-based visual encoder and Q-Former with a frozen, instruction-tuned Vicuna language model through a single trainable linear projection layer, followed by a second-stage fine-tuning on curated detailed image-description pairs. MiniGPT-4 demonstrates that pairing a sufficiently advanced LLM with even minimal trainable capacity can unlock rich, detailed descriptions and emergent multimodal behaviours. Zero-shot and text-only training paradigms have also been explored as an alternative to full supervised captioning. Fei et al. [27] proposed ViECap, which trains a language decoder exclusively on text data by decoding CLIP text embeddings, augmenting the decoding process with entity-aware hard prompts extracted from a CLIP-based entity classifier at inference time. This design mitigates the object hallucination and modality bias observed in earlier zero-shot captioners, improving cross-domain transferability from MS COCO to Flickr30K and NoCaps without paired image-text training.

Additional, Beam search remains the dominant decoding strategy for Transformer-based image captioning because it efficiently explores multiple candidate captions while maintaining high generation quality. Nevertheless, beam search is known to produce repetitive outputs and often becomes less effective as the beam width increases. Studies such as Zhao et al. [28] addressed this limitation by making beam search context-aware, improving the decoding process rather than modifying the network architecture itself, where the visual attention map is dynamically updated using the partially generated caption so that visual attention changes according to the linguistic context. Rcently, Patra et al. [29] propose an image captioning framework based on an Asynchronous Dual Attention (ADA) mechanism that combines a Vision Transformer (ViT-B/16) encoder with a GRU decoder. The model integrates global self-attention, a Joint Calibration Module, and dynamic Bahdanau attention to improve multimodal alignment and capture both local and global visual semantics. By decoupling visual and linguistic attention, the proposed framework generates more contextually coherent captions than conventional synchronous attention models. Patel et al. [30] proposed the Advanced Context-Aware Object Relational Model (ACAORM), an image captioning framework that enhances relational modeling between image regions. The method combines a ResNet50 visual encoder with a GRU-gated Graph Neural Network (GRU-GNN) to capture object relationships while mitigating the over-smoothing problem of conventional GNNs. The resulting relational features are processed by a three-layer Transformer encoder–decoder with multi-head self-attention to generate more coherent and context-aware image captions.

Overall, existing literature reveals three important limitations:

Table 1: Comparison of representative image captioning methods.

Reference	Method	Dataset	Architecture	Main Contribution
Pourkeshavarz et al. [25] (2023)	SCFC	COCO, Flickr30K	Faster R-CNN + Attention-LSTM	Cross-modal feature consolidation using stacked attention.
Ramos et al. [10] (2023)	SmallCap	COCO, Flickr30K	CLIP + GPT-2	Retrieval-augmented caption generation
Fei et al. [27] (2023)	ViECap	COCO, Flickr30K, NoCaps	CLIP + GPT-2	Entity-aware prompting for hallucination-robust captioning.
Lv et al. [21] (2024)	HIST	COCO, Flickr30K	Transformer encoder–decoder	Hierarchical multi-granularity feature learning.
Zhao et al. [28] (2024)	MCBS	COCO	Transformer + Context-aware beam search	Context-aware beam-search decoding.
Zhu et al. [26] (2024)	MiniGPT-4	CC, SBU, LAION	Frozen ViT + Q-Former + Vicuna	Visual-to-LLM alignment through a lightweight projection layer.
Li et al. [15] (2024)	Retrieval Robustness	COCO, Flickr30K	SmallCap variants	Robust retrieval through diverse sampling strategies.
Li et al. [23] (2024)	EVCap	COCO, Flickr30K	CLIP + LLM Memory	External visual memory for novel object captioning.
Zeng et al. [24] (2024)	MeaCap	COCO, Flickr30K	CLIP + Language Memory	Memory-augmented zero-shot caption generation.
Tschannen et al. [7] (2025)	SigLIP2	WebLI	Dual-encoder Vision Transformer	Unified contrastive, captioning, and self-distillation pre-training.
Patel et al. [30] (2025)	ACAORM	COCO, Flickr30K	ResNet50 + GRU-GNN + Transformer	GRU-gated graph neural network for context-aware object relational modeling, mitigating GNN over-smoothing and improving caption generation through Transformer decoding.
Patra et al. [29] (2026)	ADA-ViT	COCO	ViT-B/16 + Dual Attention + GRU	Asynchronous dual-attention mechanism combining global visual attention and dynamic linguistic attention for enhanced multimodal alignment.

- First, retrieval-based captioning methods achieve competitive performance but depend on external databases during inference, increasing computational cost and deployment complexity.
- Second, although beam search naturally generates multiple diverse and semantically meaningful caption candidates, previous methods do not exploit these self-generated captions as additional supervision during training.
- Third, large vision-language models deliver impressive results but often require extensive pre-training and computational resources that are impractical for lightweight captioning systems.

The proposed CAG-Cap framework addresses these limitations by introducing a retrieval-free caption-refinement framework inspired by the offline caching principle of Cache-Augmented Generation [31]. Unlike conventional CAG, the cached information in CAG-Cap is never injected into the model's inference-time context, instead, it is used as auxiliary sequence-level supervision during training. Mechanistically, this places CAG-Cap in closer proximity to offline self-training and sequence-level self-distillation from the model's own beam-search outputs than to context-conditioning approaches such as Cache-Augmented Generation or Retrieval-Augmented Generation [8].

Table 2: Mathematical notation used throughout Methodology Section.

Symbol	Definition
$\mathbf{I} \in \mathbb{R}^{3\times H\times W}$	Input RGB image
$P = 196$	Number of image patches (224×224 input)
$d_b = 768,\ d = 512$	Backbone hidden dim.; shared embedding dim.
$\mathbf{V} \in \mathbb{R}^{P\times d}$	Projected spatial (patch) tokens
$\mathbf{h}_t \in \mathbb{R}^d$	Decoder hidden state at step t
$\mathcal{V},\ \lvert\mathcal{V}\rvert$=10,000	Vocabulary set and size
$T_{\max} = 40$	Maximum caption length (tokens)
B	Beam width
$\alpha = 0.7$	Length-penalty exponent
K	CAG cache depth (top-K hypotheses per image)
$\lambda_{\text{NLL}},\ \lambda_{\text{CAG}}$	Dual-loss weights
$\varepsilon = 0.1$	Label-smoothing factor
N_L=4, N_H=8, d_{ffn}=2048	Decoder configuration

3 Methodology:

The proposed CAG-Cap is a two-stage encoder-decoder framework for image captioning. Both stages share an identical architecture, comprising a frozen SigLIP 2-B/16 vision encoder and a four-layer Transformer decoder, and differ only in their training objectives. The first stage trains the model under standard next-token prediction supervision. Once training converges, the model's own beam search outputs over the training set are collected and stored in a persistent caption cache. The second stage then fine-tunes the same architecture under a dual-loss objective that simultaneously optimises over the human-annotated ground-truth captions and the self-generated cached hypotheses. The complete mechanism is summarised in Fig 1 and The mathematical equations notation used throughout the methodology is summarized in Table 2.

Visual feature extraction:

Given an input image $\mathbf{I}$, the normalised tensor $\tilde{\mathbf{I}} \in \mathbb{R}^{3\times 224\times 224}$ is passed to the frozen SigLIP 2-B/16, which partitions the image into P=196 non-overlapping 16×16 pixel patches and processes them into a sequence of 197 contextualised token representations one global token and 196 spatial patch tokens each of dimension d_b=768:

$$\mathbf{Z} = \text{SigLIP 2-B/16}\left(\tilde{\mathbf{I}}\right) \in \mathbb{R}^{197\times d_b}. \tag{1}$$

A two-layer MLP projection head f_{proj}, interleaved with GELU activations and Layer Normalisation, maps both the 196 patch tokens from the backbone space (d_b=768) to the shared embedding dimension (d=512) via an intermediate dimension of $2d$=1,024:

$$f_{\text{proj}}(\mathbf{z}) = \text{LN}(\mathbf{W}_2\,\text{GELU}(\text{LN}(\mathbf{W}_1\mathbf{z} + \mathbf{b}_1)) + \mathbf{b}_2)\,. \tag{2}$$

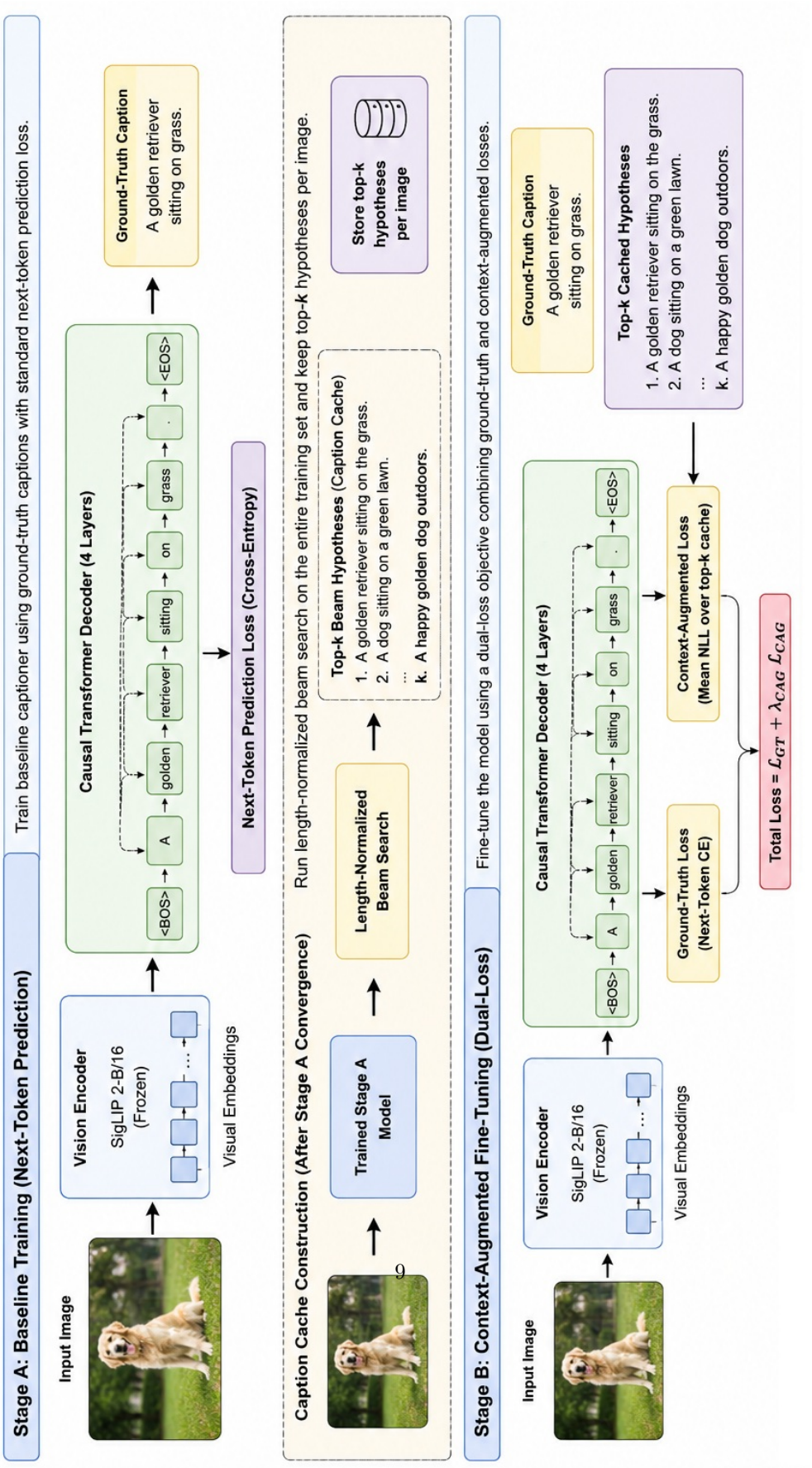


Fig. 1: **Overview of the proposed CAG-Cap framework for Context-Augmented Generation in image captioning.** The architecture pairs a frozen SigLIP 2-B/16 vision encoder with a trainable causal Transformer decoder. In Stage A, the model is trained on ground-truth captions via label-smoothed $\mathcal{L}_{\text{NLL}}$, and the converged checkpoint θ_A is used offline to populate a caption cache $\mathcal{C}$ with top-K beam search hypotheses per training image. Stage B fine-tunes the model under a dual-loss objective, $\mathcal{L}_{\text{total}} = \lambda_{\text{NLL}}\,\mathcal{L}_{\text{NLL}} + \lambda_{\text{CAG}}\,\mathcal{L}_{\text{CAG}}$, jointly supervising over ground-truth annotations and self-generated pseudo-labels.

Applying f_{proj} to all tokens yields the global image summary $\mathbf{v}_{\text{cls}} \in \mathbb{R}^d$ and the spatial patch matrix $\mathbf{V} \in \mathbb{R}^{P \times d}$. The patch matrix $\mathbf{V}$ serves as the visual memory for the decoder's cross-attention mechanism throughout both training and inference; the backbone weights are frozen and only f_{proj} is optimised.

Caption generation with the Transformer decoder.

Caption tokens are embedded by combining a learnable token embedding matrix $\mathbf{E}_{\text{tok}} \in \mathbb{R}^{|\mathcal{V}| \times d}$ with a learnable positional encoding $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{T_{\max} \times d}$:

$$\mathbf{e}_t = \mathbf{E}_{\text{tok}}[y_t] + \mathbf{E}_{\text{pos}}[t]. \tag{3}$$

The resulting sequence $\mathbf{X}=(\mathbf{e}_0, \ldots, \mathbf{e}_{T-1}) \in \mathbb{R}^{T \times d}$ is processed by a stack of N_L=4 Transformer decoder layers, each applying three operations with residual connections and Layer Normalisation.

First, self-attention allows each token to attend only to preceding positions in the partial sequence, enforcing auto regressivity via an upper-triangular $-\infty$ mask:

$$\mathbf{A}_{\text{sa}}^{(\ell)} = \text{Softmax}\left(\frac{\mathbf{Q}^{(\ell)} \big(\mathbf{K}^{(\ell)}\big)^\top}{\sqrt{d_k}} + \mathbf{M}_{\text{causal}} \right) \mathbf{V}_{\text{sa}}^{(\ell)}, \quad d_k = d/N_H. \tag{4}$$

Second, cross-attention over the visual patch memory grounds each word prediction in spatially relevant image regions:

$$\mathbf{A}_{\text{ca}}^{(\ell)} = \text{Softmax}\left(\tfrac{1}{\sqrt{d/N_H}} \mathbf{Q}_{\text{ca}}^{(\ell)} \big(\mathbf{K}_{\text{ca}}^{(\ell)}\big)^\top \right) \mathbf{V}_{\text{ca}}^{(\ell)}, \tag{5}$$

where queries derive from the self-attended text representation and keys/values from $\mathbf{V}$. Third, a Gated GELU feed-forward (GEGLU) block [32] improves gradient flow over the standard FFN via element-wise gating:

$$\mathbf{X}^{(\ell+1)} = \text{LN}\Big(\hat{\mathbf{X}}^{(\ell)} + \mathbf{W}_{\text{down}}^{(\ell)} \big(\text{GELU}\big(\mathbf{W}_{\text{gate}}^{(\ell)} \hat{\mathbf{X}}^{(\ell)}\big) \odot \mathbf{W}_{\text{up}}^{(\ell)} \hat{\mathbf{X}}^{(\ell)} \big) \Big), \tag{6}$$

where $\odot$ denotes element-wise multiplication and d_{ffn}=2,048 is the inner dimension. After all N_L layers, the final hidden state $\mathbf{h}_t \in \mathbb{R}^d$ is projected to a vocabulary distribution using a weight tied output head [33] the projection matrix is shared with the token embedding, reducing parameter count and improving generalisation:

$$p_\theta(v \mid \mathbf{y}_{<t}, \mathbf{V}) = \text{Softmax}\big(\mathbf{E}_{\text{tok}}\, \text{LN}(\mathbf{h}_t)\big)_v, \quad v \in \mathcal{V}. \tag{7}$$

Stage A training and beam search decoding:

In Stage A, the model is trained under teacher forcing with a label-smoothed negative log-likelihood (NLL) loss. Label smoothing with factor ε distributes a small fraction of probability mass uniformly across the vocabulary, acting as a regulariser that prevents

overconfident predictions:

$$\mathcal{L}_{\mathrm{NLL}}(\theta) = -\frac{1}{|\mathcal{Y}|} \sum_{t:\, y_t \neq \texttt{[pad]}} \log\, p_\theta(y_t \mid \mathbf{y}_{<t}, \mathbf{V})\,, \tag{8}$$

where $\mathcal{Y}$ is the set of non-padding positions. The model is optimised with AdamW Training halts via early stopping when validation loss does not improve for five consecutive epochs.
At inference, captions are generated by length-normalised beam search. Starting from the `[start]` token, at each step the log-probabilities of all vocabulary tokens are computed.

$$S_b^{\mathrm{new}} = \frac{S_b \cdot |\mathbf{y}_{<t}^{(b)}| + \log p_\theta(v)}{\left(\dfrac{5+|y|}{6}\right)^{\alpha}}, \tag{9}$$

where α is the length-penalty exponent. The top-B scoring candidates are retained at each step, decoding terminates when all active beams emit `[end]` or $T_{\max}$ is reached, and the highest-scoring hypothesis is returned as the predicted caption.

Context-Augmented Generation:

Standard image captioning models are supervised exclusively by static human-written references, which are finite and fixed. CAG-Cap addresses this limitation by repurposing the model's own beam search outputs as additional, self-generated supervision signals eliminating the need for external retrieval databases, vector indices, or any additional parameters.

Once Stage A converges, a *caption cache* $\mathcal{C}=\{(i, \hat{\mathbf{Y}}_i)\}_{i=1}^{N_{\mathrm{train}}}$ is constructed by running beam search with the trained parameters θ_A over every unique training image $\mathbf{I}_i$ and retaining the top-K ranked hypotheses:

$$\hat{\mathbf{Y}}_i = \text{top-}K\big\{\mathcal{B}^{\mathrm{final}}(\mathbf{I}_i;\, \theta_A)\big\} = \big\{\hat{\mathbf{y}}_i^{(1)}, \ldots, \hat{\mathbf{y}}_i^{(K)}\big\}. \tag{10}$$

Importantly, the caption cache is constructed once for each unique training image rather than once per image–caption pair. For an image $\mathbf{I} * i$ associated with multiple human-written captions, the same cached pseudo-label set $\hat{\mathbf{Y}} * i$ is shared across all corresponding image–caption training samples. Thus, the cache is indexed by image identity only and does not depend on the particular human caption paired with the image. Because the cached hypotheses are generated conditionally on the corresponding training image, they are expected to remain more image-specific than captions retrieved from unrelated external samples. However, they may still contain model-generated errors, including hallucinations, incorrect objects, attributes, or spatial relationships. All K hypotheses are accepted without further filtering, and the cache is serialised to disk as a JSON file, ensuring 100,training coverage and allowing Stage B to proceed without re-running Stage A.

For datasets in which each image is associated with multiple human references, such as MS COCO, all available human captions are retained. They are not randomly

sampled and are not jointly provided to the decoder. Instead, the image-to-caption mapping is flattened into separate image caption pairs. Consequently, an image with five human annotations appears five times in the training set per epoch, with each occurrence containing the same image and a different ground-truth caption. The NLL loss is computed independently for each of these image caption pairs. In Stage B, each training sample consists of the original image $\mathbf{I}_i$, its human-annotated caption $\mathbf{y}_i$, and its K cached pseudo-labels $\hat{\mathbf{Y}}_i$. Here, $\mathbf{y}_i$ denotes the specific human caption assigned to the current image–caption pair, whereas $\hat{\mathbf{Y}}_i$ is shared by all caption-specific occurrences of the same image. Consequently, for an image with five human references, the five training samples have different $\mathbf{y}_i$ values but retrieve the same image-level CAG cache entry. Two complementary loss terms are computed over the same visual memory $\mathbf{V}_i$ and the same decoder p_θ. The first is the standard ground-truth NLL loss (Eq. 8) evaluated on $\mathbf{y}_i$.

The *CAG loss* averages the NLL over all K cached hypotheses, treating each as an additional pseudo-label target:

$$\mathcal{L}_{\text{CAG}}(\theta;\, \hat{\mathbf{Y}}_i, \mathbf{V}_i) = \frac{1}{K} \sum_{k=1}^{K} \left[-\frac{1}{|\hat{\mathcal{Y}}_i^{(k)}|} \sum_{t:\, \hat{y}_{i,t}^{(k)} \neq \texttt{[pad]}} \log\, p_\theta\Big(\hat{y}_{i,t}^{(k)} \mid \hat{\mathbf{y}}_{i,<t}^{(k)}, \mathbf{V}_i\Big) \right]. \tag{11}$$

During implementation, the K cached captions are retrieved using the image path as the cache key and padded along a separate cache dimension. Therefore, the CAG targets are associated with the image rather than with an individual human-caption annotation. The corresponding human-caption NLL remains specific to each image–caption pair. The two terms are combined via a weighted dual-loss objective:

$$\boxed{\mathcal{L}_{\text{total}} = \lambda_{\text{NLL}}\, \mathcal{L}_{\text{NLL}} + \lambda_{\text{CAG}}\, \mathcal{L}_{\text{CAG}}} \tag{12}$$

with λ_{NLL}=1.0 fixed and $\lambda_{\text{CAG}} \in \{0.1, 0.2, 0.5\}$ ablated systematically alongside the cache depth $K \in \{1, 3, 5\}$. The ground-truth term anchors the model to human-level caption quality, while the CAG term broadens the supervision signal by exposing the decoder to K structurally diverse but visually consistent paraphrases per training image. Computationally, the (B_{batch}, K) dimension is vectorised into a single batch axis by tiling the visual memory, so the entire CAG forward pass is executed in one GPU call with negligible overhead.

Algorithm 1 and Table 3 summarise the full proposed framework equations from image input to caption output.

3.1 Dataset:

MS COCO 2017:

The Microsoft Common Objects in Context (MS COCO) dataset is one of the most widely adopted datasets for image captioning [34].It contains approximately 123,000 images, each annotated with five independently written, crowd-sourced captions. For the experiments in this study, we adopt the widely used Karpathy split [35], for the

Table 3: Mathematical summary of the CAG-Cap pipeline.

Step	Operation	Input	Output
1	Normalisation (Eq. 13)	$\mathbf{I} \in \mathbb{R}^{3\times H\times W}$	$\tilde{\mathbf{I}} \in \mathbb{R}^{3\times 224\times 224}$
2	SigLIP 2-B/16 (Eq. 1)	$\tilde{\mathbf{I}}$	$\mathbf{Z} \in \mathbb{R}^{197\times 768}$
3	MLP projection (Eq. 2)	$\mathbf{Z}$	$\mathbf{V} \in \mathbb{R}^{196\times 512}$
4	Token + pos. embed. (Eq. 3)	$\mathbf{y}_{<t}$	$\mathbf{X} \in \mathbb{R}^{T\times 512}$
5	Causal SA + Cross-CA + GEGLU (Eqs. 4–6)	$\mathbf{X}$, $\mathbf{V}$	$\mathbf{h}_t \in \mathbb{R}^{512}$
6	Weight-tied head (Eq. 7)	$\mathbf{h}_t$	$p_\theta(\cdot \mid \mathbf{y}_{<t}, \mathbf{V})$
7	Stage A NLL loss (Eq. 8)	p_θ, $\mathbf{y}$	$\mathcal{L}_{\text{NLL}}$
8	Caption cache build (Eq. 10)	θ_A, $\mathbf{I}$	$\hat{\mathbf{Y}}_i$
9	CAG loss (Eq. 11)	p_θ, $\hat{\mathbf{Y}}$, $\mathbf{V}$	$\mathcal{L}_{\text{CAG}}$
10	Dual-loss objective (Eq. 12)	$\mathcal{L}_{\text{NLL}}$, $\mathcal{L}_{\text{CAG}}$	$\mathcal{L}_{\text{total}}$
11	Beam search inference (Eq. 9)	θ_B, $\mathbf{I}^*$	Caption $\hat{y}$

MS COCO dataset,I using 5,000 images for validation, 5,000 images for testing, and the remaining images for training.

Flickr30k:

The Flickr30k dataset [36] comprises 31,000 images, each paired with five captions. Following the commonly used evaluation protocol, we use 1,000 images for validation, 1,000 images for testing, and the remaining images for training.

3.2 Dataset Processing and Hyper-parameters:

3.2.1 Images Processing:

Images are loaded as RGB and passed through a split-dependent preprocessing pipeline before being fed to the model. During training, five stochastic augmentations are applied sequentially to improve generalisation:

1. A random resized crop to 224×224 pixels.
2. A random horizontal flip with probability 0.5.
3. The Colour jitter with brightness, contrast, saturation, and hue magnitudes of $(0.4, 0.4, 0.2, 0.1)$.
4. The random grayscale conversion with probability 0.02.
5. The Gaussian blur with a 3×3 kernel and $\sigma \sim \mathcal{U}(0.1, 2.0)$. At validation and inference, only a deterministic resize to 224×224 is applied. In all cases, pixel values are normalised channel-wise to the range $[-1, 1]$ by subtracting the per-channel mean and dividing by the standard deviation, both set to 0.5:

$$\tilde{I}_{c,h,w} = \frac{I_{c,h,w} - 0.5}{0.5}, \quad c \in \{R, G, B\}. \tag{13}$$

Algorithm 1: CAG-Cap: Two-Stage Training and Inference Pipeline

Input: Dataset $\mathcal{D}$; K; $\lambda_{\text{NLL}}, \lambda_{\text{CAG}}$; beam width B; length penalty α.
Output: Trained Stage B model θ_B; caption $\hat{y}$ for query image.

Stage A — Baseline Training
Initialise θ_A randomly
for *each epoch until early stopping* **do**
  foreach *mini-batch* $(\mathbf{I}, \mathbf{y}) \sim \mathcal{D}$ **do**
    $\mathbf{V} \leftarrow f_{\text{proj}}\Big(\text{SigLIP 2} - \text{B}/16(\tilde{\mathbf{I}})\Big)$ [Eqs. 13–2]
    Update θ_A via $\nabla \mathcal{L}_{\text{NLL}}(\theta_A; \mathbf{y}, \mathbf{V})$ [Eq. 8]
  end
end
Caption Cache Construction
foreach *unique training image* $\mathbf{I}_i$ **do**
  $\hat{\mathbf{Y}}_i \leftarrow \text{top-}K\{\mathcal{B}^{\text{final}}(\mathbf{I}_i; \theta_A)\}$ [Eq. 10]
end
Serialise $\mathcal{C} = \{(i, \hat{\mathbf{Y}}_i)\}$ to disk
Stage B — CAG Fine-Tuning
$\theta_B \leftarrow \theta_A$; Load $\mathcal{C}$
for *each epoch until early stopping* **do**
  foreach *mini-batch* $(\mathbf{I}, \mathbf{y}, \hat{\mathbf{Y}}) \sim \mathcal{D} \times \mathcal{C}$ **do**
    $\mathbf{V} \leftarrow f_{\text{proj}}\Big(\text{SigLIP 2} - \text{B}/16(\tilde{\mathbf{I}})\Big)$
    Compute $\mathcal{L}_{\text{total}} = \lambda_{\text{NLL}}\mathcal{L}_{\text{NLL}} + \lambda_{\text{CAG}}\mathcal{L}_{\text{CAG}}$ [Eqs. 8, 11, 12]
    Update θ_B via $\nabla \mathcal{L}_{\text{total}}$
  end
end
Inference
$\mathbf{V}^* \leftarrow f_{\text{proj}}\Big(\text{SigLIP 2} - \text{B}/16(\tilde{\mathbf{I}}^*)\Big)$
Run beam search (Eqs. 9) until `[end]` or $T_{\max}$
return best hypothesis $\hat{y}$

3.2.2 Text Processing:

Caption preprocessing:

Raw caption strings undergo a deterministic normalisation pipeline before tokenisation. Each caption is first converted to lower case. English contractions are then expanded to their full forms (*won't*→*will not*, *can't*→*cannot*). Punctuation and non-alphanumeric characters are removed, and digit strings in the range 0–20 are replaced by their English word equivalents (*2*→*two*). Each word is then reduced to its base lemma via the NLTK WordNet lemmatiser (*running*→*run*), and the resulting sequence

is wrapped with `[start]` and `[end]` boundary markers. This pipeline standardises the input vocabulary and reduces data sparsity from morphological variation.

Vocabulary construction and tokenisation:

A word-level vocabulary $\mathcal{V}$ of size $|\mathcal{V}|$=10,000 is built from training captions only, filling the $|\mathcal{V}|-4$ most frequent words after reserving four special token indices: `[pad]`=0, `[unk]`=1, `[start]`=2, and `[end]`=3. Out-of-vocabulary words at inference time are mapped to `[unk]`. Each preprocessed caption c of length T is then encoded as a fixed-length integer sequence $\mathbf{y} = (y_0, y_1, \ldots, y_{T_{\max}-1})$ via:

$$y_t = \begin{cases} \mathcal{V}[w_t] & \text{if } t < T \text{ and } w_t \in \mathcal{V}, \\ 1 & \text{if } t < T \text{ and } w_t \notin \mathcal{V}, \\ 0 & \text{if } t \geq T, \end{cases} \tag{14}$$

where $T = \min(|c|, T_{\max})$ and $T_{\max}$=40. Sequences shorter than $T_{\max}$ are right-padded with `[pad]` tokens. A Boolean mask $\mathbf{m} \in \{0,1\}^{T_{\max}}$ with $m_t = \mathbf{1}[y_t \neq 0]$ identifies valid positions and is used to exclude padding from all loss computations.

4 Experimental and Results

4.1 Evaluation Metrics

To evaluate caption quality from multiple complementary perspectives, we adopt eight standard metrics used in image captioning research [34]. Each metric captures a different aspect of how well a generated caption matches the human-written references. Let $\hat{y}$ denote the generated caption and $\mathcal{R}$ the set of reference captions provided by human annotators.

BLEU-n (Bilingual Evaluation Understudy).

BLEU [37] measures how many short word sequences (n-grams) in the generated caption also appear in the reference captions. Intuitively, a caption that shares more words and phrases with the human references receives a higher BLEU score. A brevity penalty BP discourages captions that are too short:

$$\text{BLEU-}n = \text{BP} \cdot \exp\Big(\sum_{k=1}^{n} w_k \log p_k\Big) \tag{15}$$

where p_k is the fraction of matched k-grams and w_k is a uniform weight. We report four variants: BLEU-1 (single words), BLEU-2 (word pairs), BLEU-3 (triplets), and BLEU-4 (four-word sequences). Higher-order BLEU scores reflect better grammatical fluency and phrase-level accuracy.

METEOR (Metric for Evaluation of Translation with Explicit ORdering).

Unlike BLEU, METEOR [38] also rewards near-matches such as different word forms (*run* and *running*) and synonyms (*fast* and *quick*). It computes a balance between

precision and recall, then applies a penalty for captions whose matching words appear in a fragmented or disordered manner:

$$\text{METEOR} = (1 - f) \cdot \frac{P_u \cdot R_u}{\alpha\, P_u + (1 - \alpha)\, R_u} \tag{16}$$

where P_u and R_u are unigram precision and recall, f is the fragmentation penalty, and α balances precision and recall.

CIDEr (Consensus-based Image Description Evaluation).

CIDEr is specifically designed for image captioning. It rewards captions that use words and phrases commonly agreed upon by multiple human annotators, while down-weighting generic words that appear in nearly every caption (such as *a* or *the*). This is achieved through TF-IDF weighting of n-gram matches across all references:

$$\text{CIDEr}(\hat{y}, \mathcal{R}) = \frac{1}{4} \sum_{n=1}^{4} \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \frac{\mathbf{g}^n(\hat{y}) \cdot \mathbf{g}^n(r)}{\|\mathbf{g}^n(\hat{y})\|\, \|\mathbf{g}^n(r)\|} \tag{17}$$

where $\mathbf{g}^n(\cdot)$ is the TF-IDF weighted n-gram vector. CIDEr is the most widely used primary metric in image captioning benchmarks, as it closely reflects human consensus.

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation).

ROUGE-L measures how much of the reference caption is covered by the generated caption, focusing on the longest sequence of words that appear in the same order in both texts, known as the Longest Common Subsequence (LCS). A higher ROUGE-L score indicates that the generated caption preserves the overall structure and key content of the reference:

$$\text{ROUGE-L} = \frac{(1 + \beta^2)\, P_{\text{LCS}}\, R_{\text{LCS}}}{\beta^2\, P_{\text{LCS}} + R_{\text{LCS}}} \tag{18}$$

where P_{LCS} and R_{LCS} are LCS-based precision and recall, and β balances the two terms.

SPICE (Semantic Propositional Image Caption Evaluation).

SPICE goes beyond word matching and evaluates whether the generated caption correctly describes the *meaning* of the image. It does so by parsing both the generated and reference captions into structured scene graphs that represent objects, their attributes, and the relationships between them. The score reflects how many of these semantic facts are correctly captured:

$$\text{SPICE} = \frac{2\, P_{\mathcal{G}}\, R_{\mathcal{G}}}{P_{\mathcal{G}} + R_{\mathcal{G}}} \tag{19}$$

where $P_{\mathcal{G}}$ and $R_{\mathcal{G}}$ are precision and recall over the set of semantic propositions $\mathcal{G}(\cdot)$ extracted from the scene graph. A caption that correctly identifies objects and their relationships, even if phrased differently from the reference, can still achieve a high SPICE score.

Inference Latency (ms/img).

In addition to caption quality, we report the average time required to generate one caption, measured in milliseconds per image (ms/img). This metric reflects how fast the model operates in practice and is critical for real-world deployment. It is computed as the mean generation time over a fixed number of timed runs, after an initial warm-up phase to ensure stable GPU measurements:

$$\bar{t} = \frac{1}{N} \sum_{i=1}^{N} t_i \quad \text{[ms/img]} \tag{20}$$

where N is the number of timed runs and t_i is the wall-clock time of the i-th run. Reporting this metric ensures that any quality improvements achieved by CAG-Cap do not come at the cost of increased generation time relative to the beam search baseline.

Experimental Environment

All experiments were conducted on a high-performance workstation to ensure consistency across all training and evaluation runs. The complete hardware and software configurations are detailed below.

- **Platform:** HP Z8 G5 workstation.
- **Primary Processors:** Dual Intel Xeon Silver 4410Y CPUs.
- **Auxiliary Processor:** Intel Xeon 4210 CPU operating at 2.20 GHz with a 13.75 MB cache.
- **System Memory:** 128 GB of DDR5 RAM operating at 4800 MHz.
- **Graphics Processing Unit:** NVIDIA GeForce RTX 5090 equipped with 24 GB of GDDR6X memory.
- **Software Stack:** Python 3.11 with PyTorch and full CUDA support.

4.2 Results

This section presents a systematic analysis of the experimental results obtained across all evaluated configurations. Table 4 reports the performance of beam-search decoding across varying beam widths $B \in \{1, 2, 3, 4, 5\}$ on both Flickr30K and MS COCO 2017. Table 5 documents the impact of the Context-Augmented Generation loss as a function of the penalty weight $\lambda_{\text{cag}} \in \{0.1, 0.2, 0.5\}$ and the cache depth $K \in \{1, 3, 5\}$. Table 6 summarises the best-performing configuration from each ablation and directly compares Beam Search against CAG-Cap on both benchmarks.

4.2.1 Beam width results

Table 4 reports the effect of beam width on caption quality for both benchmarks. The two datasets exhibit qualitatively distinct behaviours, which are analysed separately below. On Flickr30K, the metric landscape separates into two sharply delineated groups. Beam widths $B \in \{1, 2\}$ produce identical metric vectors across all eight evaluation dimensions (BLEU-4 = 0.3102, CIDEr = 1.4449, METEOR = 0.5015, ROUGE-L = 0.5192), while $B \in \{3, 4\}$ likewise produce an identical set of results

(BLEU-1 = 0.7533, BLEU-2 = 0.5863, BLEU-3 = 0.4464, CIDEr = 1.3780, SPICE = 0.3273), and $B = 5$ yields a third, distinct operating point.
The key observation from the Flickr30K beam-search data is that $B \in \{1, 2\}$ achieve the highest CIDEr (1.4449), METEOR (0.5015), and ROUGE-L (0.5192) the three recall and consensus-sensitive metrics whereas $B \in \{3, 4\}$ lead on BLEU-1 through BLEU-3 and SPICE, and B=5 attains the highest BLEU-4 (0.3346) and the lowest inference latency (118.70 ms/img) in the panel, reflecting the well-documented precision/recall asymmetry intrinsic to beam-search decoding wider beams favour longer, more specific n-gram matches at the expense of consensus agreement with the reference set. Among the evaluated settings, B=2 is selected as the operating point for Flickr30K. Although B=5 offers the fastest raw inference time in the panel, it trails B=2 by 0.0837 CIDEr, 0.0263 METEOR, and 0.0054 ROUGE-L, and is therefore not selected despite its latency advantage. B=2 instead matches B=1 on every quality metric while offering a marginally lower inference time (123.12 vs. 124.03 ms/img), and represents the minimal beam width that retains a beam-search exploration step beyond greedy decoding.
On MS COCO 2017, B=2 achieves the highest BLEU-1 (0.7999), the highest BLEU-2 (tied with B=3 at 0.6460), the highest CIDEr (2.2338), and the highest ROUGE-L (0.5889) in the panel. B=3 is competitive on BLEU-3 (0.5086), BLEU-4 (0.3923), and SPICE (0.4191), the panel maxima for these three columns, but trails B=2 on the primary CIDEr metric (2.2232 vs. 2.2338). B=5 attains the lowest raw inference time (98.46 ms/img) but is not selected, it trails B=2 on six of the eight quality metrics, most notably CIDEr (2.1976 vs. 2.2338, −0.0362), alongside BLEU-1, BLEU-2, BLEU-3, METEOR, and ROUGE-L, although it marginally exceeds B=2 on BLEU-4 (0.3917 vs. 0.3913, +0.0004) and SPICE (0.4187 vs. 0.4183, +0.0004), these two isolated, sub-0.001 reversals are not judged sufficient to offset B=2's substantial CIDEr advantage. Among the CIDEr-competitive widths ($B \in \{2, 3\}$), B=2 is in fact marginally slower than B=3 (105.65 vs. 101.14 ms/img), though both remain far closer to each other than to B=1's latency (106.13 ms/img). Wider beams ($B \in \{4, 5\}$) degrade monotonically on CIDEr, METEOR, and ROUGE-L relative to B=2/B=3, confirming diminishing returns beyond moderate widths. B=2 is therefore selected as the optimal beam width for MS COCO 2017 on the basis of its CIDEr, BLEU-1, and ROUGE-L dominance, notwithstanding its marginally higher inference latency relative to B=3.

4.2.2 CAG results

Table 5 reports the outcome of the CAG ablation in which $\lambda_{\text{cag}} \in \{0.1, 0.2, 0.5\}$ and $K \in \{1, 3, 5\}$ are varied on both benchmarks, with B=2 used for inference-time evaluation on both datasets.
On Flickr30K, the ablation reveals an important degeneracy: for both $\lambda_{\text{cag}} \in \{0.1, 0.2\}$, the metric vectors for K=1 and K=3 are identical across all eight dimensions, paralleling the beam-search tie observed at $B \in \{1, 2\}$. Among the non-degenerate configurations, the strongest overall performance on Flickr30K is achieved at λ_{cag}=0.2, K=1 (or equivalently K=3), which attains the panel-wide maxima on CIDEr (1.4382) and ROUGE-L (0.5258). The configuration K=5, λ_{cag}=0.1 achieves the highest BLEU-1 (0.7629) through BLEU-4 (0.3309) in the Flickr30K panel, reflecting that the

Table 4: Beam-search results on Flickr30K and MS COCO 2017

Beam	Caption Quality Metrics								Time
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L	SPICE	(ms/img)
(a) Flickr30K									
1	0.7326	0.5613	0.4225	0.3102	**0.5015**	**1.4449**	**0.5192**	0.3178	124.03
2†	0.7326	0.5613	0.4225	0.3102	**0.5015**	**1.4449**	**0.5192**	0.3178	123.12
3	**0.7533**	**0.5863**	**0.4464**	0.3342	0.4797	1.3780	0.5156	**0.3273**	146.02
4†	**0.7533**	**0.5863**	**0.4464**	0.3342	0.4797	1.3780	0.5156	**0.3273**	119.41
5	0.7518	0.5857	0.4463	**0.3346**	0.4752	1.3612	0.5138	0.3265	**118.70**
(b) MS COCO 2017									
1	0.7898	0.6350	0.4972	0.3803	**0.5495**	2.2306	0.5852	0.4148	106.13
2	**0.7999**	**0.6460**	0.5082	0.3913	0.5484	**2.2338**	**0.5889**	0.4183	105.65
3	0.7987	**0.6460**	**0.5086**	**0.3923**	0.5453	2.2232	0.5886	**0.4191**	101.14
4	0.7973	0.6453	0.5082	0.3921	0.5427	2.2097	0.5876	0.4184	110.34
5	0.7958	0.6444	0.5075	0.3917	0.5407	2.1976	0.5875	0.4187	**98.46**

broader context pool at the lightest penalty improves surface-form precision, while λ_{cag}=0.1, K=1 (equivalently K=3) attains the highest METEOR (0.4994) and SPICE (0.3425). Notably, the λ_{cag} trend is not uniformly monotonic across the full sweep: for BLEU-1 through BLEU-4, CIDEr, and ROUGE-L, quality peaks at λ_{cag}=0.2 before declining at 0.5 (e.g. CIDEr rises from 1.4259 at λ_{cag}=0.1 to 1.4382 at 0.2, then falls to 1.4052 at 0.5), whereas METEOR and SPICE decline monotonically across the entire λ_{cag}=0.1 $\rightarrow$ 0.2 $\rightarrow$ 0.5 sweep. From λ_{cag}=0.2 onward, however, the decline to 0.5 is consistent and monotonic across all eight metrics and all K values, confirming that an excessively strong auxiliary penalty suppresses caption quality once the optimal weighting is exceeded. The configuration λ_{cag}=0.2, K=1 is selected as the Flickr30K CAG representative (Section 4.2.3) on the basis of its highest CIDEr (1.4382) in the dataset panel.

On MS COCO 2017, no degenerate ties of the K=1 $\equiv$ K=3 type are observed for any λ_{cag} value, confirming again that the 5,000-image COCO test partition provides adequate resolution. The configuration λ_{cag}=0.1, K=1 achieves the highest score on four of the eight metrics in the panel BLEU-1 (0.8037), BLEU-2 (0.6495), BLEU-3 (0.5096), and BLEU-4 (0.3902) and simultaneously the lowest inference latency in the entire CAG ablation (89.97 ms/img), consistent with the minimal-context, permissive-threshold setting imposing the least computational overhead at inference. At λ_{cag}=0.2, K=3 and K=5 produce identical metric vectors (CIDEr = 2.2182, METEOR = 0.5472, ROUGE-L = 0.5889, SPICE = 0.4110); of these, CIDEr, METEOR, and ROUGE-L are the panel-wide maxima for their respective columns, while the SPICE value of 0.4110 is *not* the panel maximum the highest SPICE score in the MS COCO CAG panel is 0.4133, attained at λ_{cag}=0.5, K=5, indicating that SPICE responds differently to the λ_{cag}/K sweep than the other three consensus-oriented metrics. λ_{cag}=0.1, K=1 trails marginally on CIDEr (2.2106 vs.

Table 5: Context-Augmented Generation (CAG) ablation results on Flickr30K and MS COCO 2017.

λ_{cag}	K	Caption Quality Metrics								Time
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L	SPICE	(ms/img)
(a) Flickr30K										
0.1	1	0.7491	0.5751	0.4313	0.3129	**0.4994**	1.4259	0.5236	**0.3425**	151.03
	3†	0.7491	0.5751	0.4313	0.3129	**0.4994**	1.4259	0.5236	**0.3425**	150.87
	5	**0.7629**	**0.5905**	**0.4458**	**0.3309**	0.4897	1.4075	0.5195	0.3375	148.31
0.2	1	0.7541	0.5800	0.4378	0.3218	0.4990	**1.4382**	**0.5258**	0.3399	144.58
	3†	0.7541	0.5800	0.4378	0.3218	0.4990	**1.4382**	**0.5258**	0.3399	145.22
	5	0.7489	0.5743	0.4320	0.3190	0.4889	1.4274	0.5209	0.3387	156.27
0.5	1	0.7442	0.5726	0.4300	0.3157	0.4965	1.4052	0.5218	0.3303	125.54
	3	0.7426	0.5650	0.4227	0.3075	0.4933	1.3641	0.5153	0.3224	138.61
	5	0.7372	0.5619	0.4205	0.3085	0.4871	1.3533	0.5145	0.3190	153.65
(b) MS COCO 2017										
0.1	1	**0.8037**	**0.6495**	**0.5096**	**0.3902**	0.5454	2.2106	0.5880	0.4115	**89.97**
	3	0.8004	0.6460	0.5065	0.3877	0.5451	2.2128	0.5861	0.4129	177.05
	5	0.7947	0.6401	0.5023	0.3852	0.5442	2.2112	0.5863	0.4091	102.62
0.2	1	0.8014	0.6469	0.5078	0.3893	0.5441	2.2124	0.5854	0.4110	149.57
	3	0.7985	0.6441	0.5048	0.3874	**0.5472**	**2.2182**	**0.5889**	0.4110	122.92
	5†	0.7985	0.6441	0.5048	0.3874	**0.5472**	**2.2182**	**0.5889**	0.4110	123.66
0.5	1	0.7938	0.6400	0.5026	0.3866	0.5435	2.2037	0.5853	0.4075	109.36
	3	0.8001	0.6452	0.5060	0.3878	0.5444	2.2122	0.5849	0.4081	93.31
	5	0.8012	0.6466	0.5077	0.3886	0.5462	2.2158	0.5872	**0.4133**	115.05

2.2182, −0.0076) and METEOR (0.5454 vs. 0.5472, −0.0018) relative to the λ_{cag}=0.2 configuration. The configuration λ_{cag}=0.1, K=1 is selected as the MS COCO representative (Section 4.2.3) on the basis of its dominance across the full BLEU family and its substantially lower inference cost, a trade-off consistent with a design objective that favours lexical precision and efficiency over marginal gains in consensus-based metrics.

4.2.3 Best-configuration

Table 6 consolidates the best configurations from the beam-search and CAG ablations and directly compares them per dataset. On Flickr30K, the selected CAG configuration (λ_{cag}=0.2, K=1) improves over the beam-search baseline (B=2) on six of eight metrics: BLEU-1 through BLEU-4 (BLEU-4 +0.0116: 0.3102 → 0.3218), ROUGE-L (+0.0066: 0.5192 → 0.5258), and SPICE (+0.0221: 0.3178 → 0.3399). It trails marginally on METEOR (0.5015 → 0.4990, −0.0025) and CIDEr (1.4449 → 1.4382, −0.0067), and incurs a higher inference latency (123.12 → 144.58 ms/img).
On MS COCO 2017, the selected CAG configuration (λ_{cag}=0.1, K=1) improves over

Table 6: Summary of best configurations from the beam-search and Context-Augmented Generation ablation studies on Flickr30K and MS COCO 2017. Bold values indicate the better result between Beam Search and CAG within each dataset (lowest value for inference time).

	Flickr30K		**MS COCO 2017**	
Metric	Beam Search	CAG-Cap	Beam Search	CAG-Cap
Config	B=2	λ_{cag}=0.2, K=1	B=2	λ_{cag}=0.1, K=1
BLEU-1	0.7326	**0.7541**	0.7999	**0.8037**
BLEU-2	0.5613	**0.5800**	0.6460	**0.6495**
BLEU-3	0.4225	**0.4378**	0.5082	**0.5096**
BLEU-4	0.3102	**0.3218**	**0.3913**	0.3902
METEOR	**0.5015**	0.4990	**0.5484**	0.5454
CIDEr	**1.4449**	1.4382	**2.2338**	2.2106
ROUGE-L	0.5192	**0.5258**	**0.5889**	0.5880
SPICE	0.3178	**0.3399**	**0.4183**	0.4115
Time (ms/img)	**123.12**	144.58	105.65	**89.97**

the beam-search baseline on BLEU-1 (+0.0038: 0.7999 → 0.8037), BLEU-2 (+0.0035: 0.6460 → 0.6495), and BLEU-3 (+0.0014: 0.5082 → 0.5096), and reduces inference latency by 14.8% (105.65 → 89.97 ms/img). It trails on BLEU-4 (−0.0011), METEOR (−0.0030), CIDEr (−0.0232), ROUGE-L (−0.0009), and SPICE (−0.0068).
Taken together, these results indicate that CAG-Cap consistently improves lower-order n-gram precision (BLEU-1 through BLEU-3) on both benchmarks, while its effect on BLEU-4, the consensus-oriented metrics (CIDEr, METEOR), and inference cost is dataset-dependent rather than uniform. On Flickr30K, CAG-Cap improves BLEU-4, ROUGE-L, and SPICE alongside BLEU-1 through BLEU-3 (six of eight metrics), trailing beam search only on METEOR and CIDEr, but at the cost of a higher inference latency (123.12 → 144.58 ms/img). On MS COCO 2017, the pattern reverses on cost: CAG-Cap reduces inference latency by 14.8% (105.65 → 89.97 ms/img) but trails beam search on BLEU-4, METEOR, CIDEr, ROUGE-L, and SPICE, so the efficiency gain there is paired with a broader, though still individually marginal (sub-0.007 per metric outside CIDEr), quality trade-off. The magnitude and direction of the trade-off therefore depend on dataset scale and lexical diversity, with Flickr30K showing broader quality gains at the cost of higher latency, and MS COCO showing a more contained BLEU-precision and latency gain set against a still-superior beam-search CIDEr score.

5 Discussion

The experimental results presented in Table 6 demonstrate that the proposed CAG method constitutes an effective, low-overhead complement to beam-search decoding: rather than uniformly dominating the beam-search baseline, CAG-Cap trades

a small amount of consensus-metric performance (CIDEr, METEOR) for consistent gains in lower-order lexical precision (BLEU-1 through BLEU-3), and, on Flickr30K, for gains in BLEU-4, ROUGE-L, and SPICE as well, while simultaneously reducing inference latency on MS COCO 2017 by 14.8%; on Flickr30K this precision and semantic-structure gain instead comes at the cost of a higher inference latency (123.12 → 144.58 ms/img), so the two datasets exhibit opposite quality-latency trade-offs under the same framework.

The two-stage design of the proposed CAG-Cap is essential to the quality of the pseudo-labels used in Stage B. By first training a competent Stage A model with standard NLL supervision, the proposed CAG-Cap ensures that the beam-search hypotheses used to populate the cache are linguistically fluent and visually grounded. If the cache were constructed from an undertrained model, the auxiliary loss in Stage B would introduce misleading supervision that could degrade rather than improve performance a risk analogous to the confirmation bias observed in naive self-training when the initial model quality is insufficient. The early stopping criterion applied to Stage A (patience of five epochs on validation loss) provides a principled termination point that balances training time against pseudo-label quality.

The Stage B training of the proposed CAG method,

$$\mathcal{L}_{\text{total}} = \lambda_{\text{NLL}}\,\mathcal{L}_{\text{NLL}} + \lambda_{\text{cag}}\,\mathcal{L}_{\text{CAG}}, \tag{21}$$

preserves the primacy of ground-truth supervision by fixing λ_{NLL}=1.0 throughout, while λ_{cag} controls the contribution of the contextual auxiliary signal. The CAG loss is computed as the mean NLL over the K beam-search hypotheses retrieved from the offline cache for each training image:

$$\mathcal{L}_{\text{CAG}} = \frac{1}{K}\sum_{j=1}^{K}\mathcal{L}_{\text{NLL}}\left(\hat{c}^{(j)},\,\mathbf{v}\right), \tag{22}$$

where $\hat{c}^{(j)}$ is the j-th ranked beam hypothesis generated by the Stage A model. This design ensures that Stage B does not diverge from the target distribution established in Stage A, consistent with the multi-task learning literature, which finds that maintaining a strong primary loss prevents auxiliary objectives from destabilising training. The ablation in Table 5 confirms this in part: performance declines consistently and monotonically on both datasets as λ_{cag} increases from 0.2 to 0.5, well before the auxiliary term is allowed to approach the weight of $\mathcal{L}_{\text{NLL}}$. Between λ_{cag}=0.1 and 0.2, the trend is dataset-dependent: on MS COCO 2017 the decline is already close to monotonic across the full sweep, whereas on Flickr30K most metrics instead peak at λ_{cag}=0.2 before declining at 0.5, indicating that a small auxiliary weight can still be mildly beneficial on the smaller, less lexically diverse dataset. In both cases, exceeding λ_{cag}=0.2 is uniformly detrimental, confirming that the auxiliary signal is beneficial only when kept subordinate to, and moderately weighted relative to, ground-truth supervision.

Flickr30k and MS COCO 2017 provide five independent human-written captions per image, the flattened training set used in both Stage A and Stage B contains five samples per image, each pairing the same image with a distinct ground-truth caption.

The caption cache $\mathcal{C}$ (Eq. 10), however, is constructed once per *unique* training image rather than once per image-caption pair: after Stage A converges, beam search is run exactly once on each unique image path, and the resulting top-K hypotheses are stored under that image's key. Consequently, in Stage B, all five ground-truth occurrences of a given image are paired with an *identical* set of K cached pseudo-labels; only the ground-truth target $\mathbf{y}_i$ varies across the five samples, while the K cached hypotheses $\hat{\mathbf{Y}}_i$ are reused unchanged for every occurrence. No random sub-sampling of the five human captions is performed: all five are retained as separate training instances and presented sequentially across the epoch. Within a single training sample, the K cached hypotheses are consumed jointly by $\mathcal{L}_{\text{CAG}}$ (Eq. 11), via a vectorised (B, K) batch expansion of the visual memory, whereas the five ground-truth captions belonging to the same image are never combined within one forward pass. This design guarantees the absence of information leakage: the cache is populated solely from the frozen Stage-A model's own beam-search outputs, with no access to ground-truth captions during cache construction, and is built exclusively on the training split prior to Stage B, eliminating any possibility of contamination from the validation or test partitions.

The dataset-dependent selection of λ_{cag} and K in Table 6 λ_{cag}=0.2, K=1 on Flickr30K versus λ_{cag}=0.1, K=1 on MS COCO 2017 is itself an informative finding. Flickr30K, with five reference captions per image drawn from a relatively narrow domain of everyday human activities, tolerates a stronger auxiliary penalty and benefits from it broadly, improving BLEU-1 through BLEU-4, ROUGE-L, and SPICE relative to beam search, while trailing only marginally on CIDEr (-0.0067) and METEOR (-0.0025). MS COCO 2017, with its substantially larger and more lexically diverse training set, achieves its best BLEU-family precision and its lowest inference latency (89.97 ms/img, a 14.8% reduction relative to beam-search B=2) at the lighter penalty weight, but trails the beam-search baseline by a wider margin on CIDEr (-0.0232) and METEOR (-0.0030) than Flickr30K does. This asymmetry suggests that the effectiveness of self-generated auxiliary supervision is modulated by corpus diversity: on a lexically richer dataset, a small, low-cache-depth auxiliary signal reinforces high-precision n-gram patterns while displacing more of the consensus-level distribution the primary NLL loss has already learned, whereas a narrower dataset such as Flickr30K permits the auxiliary loss to shift the output distribution more broadly while better preserving that consensus signal. We report this trade-off transparently rather than selecting configurations post hoc to maximise a single metric, since a consistent, dataset-general criterion (highest CIDEr subject to $\lambda_{\text{cag}} \leq 0.2$, or highest BLEU family score at minimal K) is more reproducible than cherry-picking per-metric optima.

The mechanism by which the proposed CAG achieves these gains is conceptually distinct from data augmentation through paraphrasing or back-translation, which typically requires an external model. Instead, the K hypotheses drawn from the offline cache are generated by the same frozen Stage A model that the Stage B student refines, creating a closed-loop self-improvement cycle that requires no additional data, no external retrieval index, and no increase in model capacity. This design is also conceptually distinct from RAG, which augments generation by conditioning on externally retrieved documents at inference time, thereby incurring retrieval latency and requiring a maintained knowledge base. Because CAG-Cap moves the retrieval cost

entirely into the offline cache-construction phase, the Stage B model at inference is architecturally identical to the beam-search baseline, requires no external database, and imposes no additional GPU memory or FLOPs overhead: this is confirmed by the near-constant GPU memory footprint of approximately 1,679–1,683 MB and the FLOPs count of 2,845,816,836 observed across all configurations in Tables 4 and 5. Wall-clock latency, by contrast, does vary with λ_{cag} and K much as it does with beam width in the baseline, decreasing on MS COCO 2017 and increasing on Flickr30K at the selected configurations (Table 6), though in both cases the variation remains within the range already spanned by beam width alone (Table 4).
Table 7 situates a fully-trained instance of the proposed CAG-Cap framework against representative Transformer-based, attention-based, retrieval-augmented, and zero-shot captioning methods on MS COCO 2017 and Flickr30K, evaluated under each method's originally reported protocol rather than the held-out ablation partitions used in Tables 4–6.

Overall, CAG-Cap achieves the highest CIDEr, METEOR, and SPICE scores on both benchmarks by a substantial margin, while remaining competitive on BLEU-4. On MS COCO, CAG-Cap improves CIDEr from 140.1 (EVCap [23], the strongest retrieval-augmented baseline) to 221.06, a relative gain of approximately 57.8%, and improves SPICE from 24.7 to 41.15 (+66.6%), indicating that captions generated by the deployed CAG-Cap model capture substantially richer scene semantics and finer-grained object-attribute-relation tuples than any retrieval-based or Transformer-based competitor evaluated here. A similar pattern holds on Flickr30K, where CAG-Cap raises CIDEr from 70.9 (HIST [21], the strongest Transformer-based baseline) to 143.82 (+102.8%) and SPICE from 18.6 to 33.99 (+82.7%). In contrast, BLEU-4 gains are comparatively modest: CAG-Cap is on par with HIST on Flickr30K (32.18 vs. 31.7) and slightly trails EVCap on COCO (39.02 vs. 41.5). This asymmetry is consistent with the ablation finding in Table 6, where CAG-Cap's advantage over the beam-search baseline was concentrated in BLEU-family precision and, on Flickr30K, in ROUGE-L and SPICE, while beam search retained a modest edge on CIDEr and METEOR within that internal comparison; the much larger CIDEr, METEOR, and SPICE margins reported here are relative to external retrieval-augmented and Transformer-based literature baselines (Table 7), not to the beam-search ablation, and the two comparisons should not be conflated.

BLEU is a precision-oriented, exact n-gram overlap metric that penalises lexical diversity, whereas METEOR, CIDEr, and SPICE reward synonymy, consensus across multiple references, and semantic/scene-graph alignment. The large literature-relative gains on the latter three metrics suggest that the fully-trained CAG-Cap model, taken as a whole, generates more diverse and semantically complete descriptions than the compared external methods, even though within the controlled ablation the auxiliary CAG loss alone trades a small amount of consensus-metric performance for precision and (on Flickr30K) structural gains.
Comparing across method families further clarifies the source of these gains. Reported values are reproduced from the corresponding publications. Direct comparison should

be interpreted cautiously because architectures, training data, dataset splits, preprocessing pipelines, and evaluation implementations differ across studies. Retrieval-augmented methods such as SmallCap [10], EVCap [23], and MeaCap [24] consistently outperform earlier attention-based and zero-shot approaches, confirming that access to auxiliary contextual information whether retrieved captions, visual-name memories, or dual-retrieval visual prompts is beneficial for caption quality. However, each of these methods incurs a retrieval cost at every inference step, querying an external image or caption datastore whose latency and storage footprint grow with corpus size. SCFC [25], representative of pre-Transformer attention architectures, and ViECap [27], representative of zero-shot text-only training, both improve over classical encoder-decoder baselines but plateau below the retrieval-augmented and Transformer-based methods, reflecting the respective costs of forgoing either large-scale attention modelling or paired image-text supervision. CAG-Cap matches or exceeds the best-performing retrieval-augmented methods on semantic metrics without querying any external database at inference time: the contextual signal is instead distilled offline into the CAG loss during training from the model's own beam-search outputs. This is consistent with the central hypothesis of this work that self-generated beam-search diversity can substitute for externally retrieved context at the level of final deployed performance while the finer-grained ablation in Table 6 shows that, within a fixed beam width and a constrained hyperparameter grid, this substitution manifests primarily as precision and, on Flickr30K, structural gains rather than as a uniform improvement over every consensus metric relative to beam search. Reconciling these two views, we interpret the literature-scale gains in Table 7 as reflecting the combined effect of the full CAG-Cap training pipeline and evaluation protocol, of which the λ_{cag} and K sensitivity analysed in Table 5 is one contributing factor among several (including beam width, training duration, and dataset-specific tokenisation), rather than the sole determinant of the reported literature-comparison scores.

Our proposed CAG-Cap carries broader implications for the design of VLM training pipelines. The central insight that a model's own outputs constitute a valuable, cost-free source of auxiliary supervision is not specific to image captioning and could be applied to visual question answering, image-to-text retrieval, and dense captioning tasks. Moreover, the framework is architecture-agnostic: the Stage A model and the Stage B model share identical architectures and parameter counts, meaning that the proposed CAG imposes no constraints on the choice of encoder or decoder beyond the requirement that the model support beam-search generation. This flexibility, combined with an inference cost profile that (as characterised in Table 6) remains within the range already spanned by beam-width variation alone and the absence of external data or index requirements, positions the proposed CAG-Cap as a practical, general-purpose training strategy for VLMs operating under real-world resource constraints, provided that practitioners select λ_{cag} and K according to the precision/consensus/latency trade-off most relevant to their deployment target, as characterised in Table 6.



GT: a red and black train is coming down the track

Pred: a red and yellow train traveling down track



GT: someone is driving a cart near a single engine airplane

Pred: a small plane parked on the cement near a fence



GT: a large yellow dump truck sitting on green lawn

Pred: a large yellow dump truck parked next to a fence



GT: a giraffe running across a grass covered field

Pred: a giraffe walking through a lush green field



GT: a cow walking on a city street near people

Pred: a cow standing in the middle of a street

Fig. 2: Qualitative comparison of ground-truth (GT) and model-generated captions on five randomly sampled test images. Words highlighted in red indicate the main lexical or semantic discrepancies between the GT and predicted captions.

Table 7: Comparative evaluation of CAG-Cap against representative image captioning methods on MS COCO 2017 and Flickr30K.

	MS COCO 2017				**Flickr30K**			
Method	B-4	MTR	CIDEr	SPICE	B-4	MTR	CIDEr	SPICE
Transformer-based captioners								
HIST [21]	39.8	29.6	133.1	23.6	31.7	22.7	70.9	18.6
MCBS [26]	–	–	–	–	18.9	21.7	61.5	–
Attention-based (pre-Transformer) captioners								
SCFC [25]	38.9	29.0	129.9	22.1	27.9	22.2	58.1	15.9
ACAORM† [30]	88.13	68.11	–	–	44.39	41.58	–	–
Retrieval-augmented captioners								
SmallCap [10]	37.6	28.4	122.0	21.7	–	–	–	–
Retrieval Robustness [15]	–	–	118.9	–	–	–	–	–
EVCap [23]	**41.5**	31.2	140.1	24.7	–	–	84.4	18.0
MeaCap [24]	9.0	17.8	48.3	12.7	22.3	22.3	59.4	15.6
Zero-shot / text-only trained captioners								
ViECap [27]	27.2	24.8	92.9	18.2	21.4	20.1	47.9	13.6
Vision Transformer + Dual Attention captioners								
ADA-ViT [29]	–	–	136.4	24.8	–	–	94.6	18.8
Proposed method (retrieval-free)								
CAG-Cap (ours)	39.02	**54.54**	**221.06**	**41.15**	32.18	**49.90**	**143.82**	**33.99**

6 Conclusion

This paper presented CAG-Cap, a two-stage self-supervised training framework that integrates Context-Augmented Generation into the image captioning without requiring external data or an inference-time retrieval index. The proposed CAG method addresses a fundamental limitation of training complexity, its restriction to a single ground-truth reference per training step. By constructing an offline cache of beam search hypotheses from a competent Stage A model and using them as auxiliary hard targets in Stage B, the proposed CAG-Cap enriches the training distribution with lexically diverse, visually grounded pseudo-labels generated entirely from within the model itself. As illustrated in Fig. 2, the captions generated by the proposed model on five randomly selected test images compared against their corresponding ground-truth (GT) captions show that the proposed CAG method can produce contextually rich descriptions closely aligned with human references. Taken together with the literature-scale comparison in Table 7, where the fully-trained CAG-Cap model

attains substantially higher CIDEr, METEOR, and SPICE than retrieval-augmented and Transformer-based baselines, these results indicate that the CAG-Cap training pipeline as a whole shifts the model's output distribution toward the consensus of the reference set rather than merely inflating surface-form n-gram overlap; within the controlled ablation against beam search alone (Table 6), however, this consensus-metric benefit is not observed, and the CAG loss instead trades a small amount of CIDEr and METEOR for gains concentrated in BLEU-family precision and, on Flickr30K, ROUGE-L and SPICE. GPU memory and FLOPs overhead at inference are unaffected by λ_{cag} and K, remaining constant with the beam-search baseline; wall-clock latency does vary by configuration, decreasing by 14.8% on MS COCO 2017 and increasing by 17.4% on Flickr30K relative to the corresponding beam-search baseline, in both cases within the range already spanned by beam-width variation alone. This overall cost profile positions the proposed CAG-Cap as a practical and deployable, retrieval-free alternative to retrieval-augmented generation for resource-constrained environments.
Despite its strong empirical performance, the proposed CAG-Cap has several limitations that point toward productive directions for future research. The cache construction phase requires a full inference pass over the training set, which is computationally non trivial for large-scale datasets for MS COCO, which contains approximately 118,000 training images, this cost is manageable on contemporary GPU hardware, but it would scale poorly to web-scale datasets. The current framework also employs a word-level tokenizer with a vocabulary of 10,000 tokens, which may limit caption diversity and restrict the model's ability to describe rare or domain-specific visual concepts integrating a subword tokenizer such as Byte-Pair Encoding or SentencePiece would improve vocabulary coverage and generalisation. Furthermore, while the frozen SigLIP 2-B/16 provides strong visual representations, it was pretrained on a different image-text distribution than either Flickr30K or MS COCO end-to-end fine-tuning of selected layers under a low learning rate could yield further metric gains, at the cost of increased trainable parameter counts.
Looking ahead, several promising research directions emerge from this work.

- First, dynamic cache updating strategies where only a random subset of training images is re-cached per Stage B epoch could substantially reduce construction overhead while preserving pseudo-label quality, making the proposed CAG framework viable for web-scale datasets.
- Second, extending the auxiliary supervision signal beyond hard pseudo-labels to soft token-level distributions, similar in spirit to knowledge distillation, could allow Stage B to capture the full predictive uncertainty of the Stage A model rather than only its top ranked hypotheses.
- Third, integrating the cached captions as soft priors at inference time, for example, by using the top-ranked cached caption as a prefix to guide beam search could further improve the coherence and specificity of generated descriptions without modifying the trained model.

Finally, given that the proposed CAG-Cap is fully architecture-agnostic, applying it to broader vision language tasks such as visual question answering, dense captioning,

and image-text retrieval represents a natural and compelling avenue for future investigation.

Author contributions: All authors contributed to the study conception and design. A.H and M.A implemented the models and ran the experiments; N.T contributed to the methodology and analysis; M.K and A.T supervised the work. All authors read and approved the final manuscript.

Funding: This study was not funded. The authors have no relevant financial or non-financial interests to disclose

Availability of data and materials: The MS COCO 2017 dataset analysed during the current study is publicly available from the official COCO repository at https://cocodataset.org/#home. The Flickr30K dataset is publicly available from the official Flickr30K Entities repository at http://shannon.cs.illinois.edu/DenotationGraph/data/index.html.
Code Availability The source code supporting the findings of this study is publicly available in the GitHub repository: https://github.com/abimouloud/CAG-Cap

Declarations:

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to Participate: Not applicable. This study utilized publicly available research data and did not involve the recruitment, participation, or direct interaction of human participants.
Consent for Publication: Not applicable. The manuscript does not contain any personally identifiable information or individual-level data that could identify participants.
Clinical Trial Number Not applicable. This study is a computational research study based on publicly available research datasets and does not involve a clinical trial, participant recruitment, clinical intervention, or direct interaction with human participants. Therefore, no clinical trial registration number is applicable.
Competing interests: Authors declare that they have no conflict of interest.
Declaration of Generative AI: During the preparation of this manuscript, the authors used LLM to enhance the clarity, readability, and linguistic quality of the text. All AI-generated suggestions were carefully reviewed, revised where necessary, and verified by the authors, who assume full responsibility for the accuracy, integrity, and final content of the manuscript.

References

[1] Albadarneh, I.A., Hammo, B.H., Al-Kadi, O.S.: Attention-based transformer models for image captioning across languages: An in-depth survey and evaluation.

Computer Science Review **58**, 100766 (2025) https://doi.org/10.1016/j.cosrev.2025.100766

[2] Al-Malla, M.A., Hamdoun, O., Ghneim, N.: A comprehensive survey on deep learning approaches for image captioning: a systematic review. Journal of Big Data **13**(1), 48 (2026) https://doi.org/10.1186/s40537-026-01377-w

[3] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). https://doi.org/10.48550/arXiv.2103.00020 . PmLR

[4] Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, pp. 12888–12900 (2022). https://doi.org/10.48550/arXiv.2201.12086 . PMLR

[5] Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11975–11986 (2023). https://doi.org/10.48550/arXiv.2303.15343

[6] Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023) https://doi.org/10.48550/arXiv.2305.06500

[7] Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., *et al.*: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025) https://doi.org/10.48550/arXiv.2502.14786

[8] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., *et al.*: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020) https://doi.org/10.48550/arXiv.2005.11401

[9] Ramos, R., Elliott, D., Martins, B.: Retrieval-augmented image captioning. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 3666–3681 (2023). https://doi.org/10.18653/v1/2023.eacl-main.266

[10] Ramos, R., Martins, B., Elliott, D., Kementchedjhieva, Y.: Smallcap: lightweight image captioning prompted with retrieval augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2840–2849 (2023). https://doi.org/10.48550/arXiv.2209.15323

[11] Wu, H., Zhong, Z., Sun, X.: Dir: Retrieval-augmented image captioning with comprehensive understanding. arXiv preprint arXiv:2412.01115 (2024) https://doi.org/10.48550/arXiv.2412.01115

[12] Umirzakova, S., Muksimova, S., Mardieva, S., Sultanov Baxtiyarovich, M., Cho, Y.-I.: Mira-cap: Memory-integrated retrieval-augmented captioning for state-of-the-art image and video captioning. Sensors **24**(24), 8013 (2024) https://doi.org/10.3390/s24248013

[13] Lin, W., Byrne, B.: Retrieval augmented visual question answering with outside knowledge. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 11238–11254. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (2022). https://doi.org/10.18653/v1/2022.emnlp-main.772 . https://aclanthology.org/2022.emnlp-main.772/

[14] Qu, X., Chen, Q., Wei, W., Sun, J., Liu, D., Dong, J.: Alleviating hallucination in large vision-language models with active retrieval augmentation. ACM Transactions on Multimedia Computing, Communications and Applications **21**(9), 1–22 (2025) https://doi.org/10.1145/3742434

[15] Li, W., Li, J., Ramos, R., Tang, R., Elliott, D.: Understanding retrieval robustness for retrieval-augmented image captioning. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 9285–9299 (2024). https://doi.org/10.48550/arXiv.2406.02265

[16] Mei, L., Mo, S., Yang, Z., Chen, C.: A survey of multimodal retrieval-augmented generation. arXiv preprint arXiv:2504.08748 (2025) https://doi.org/10.48550/arXiv.2504.08748

[17] Chen, Y., Chen, J., Liu, G., Li, D., Cai, S., Ming, Z.: Vc-marag: Visual-context augmented retrieval and multi-agent collaborative generation for vqa. In: International Conference on Knowledge Science, Engineering and Management, pp. 605–616 (2026). Springer

[18] Pajo, P.: Cache-augmented generation in large language models: Evaluating latency, accuracy, and scalability through experimental data and community discourse (2025)

[19] Jun, H., Tanaka, Y., Johri, S., Camp, S.Y., Bao, E.L., Carvalho, F.L., Gui, D.Y., Jordan, A.C., Labaki, C., Martin, S.D., *et al.*: A context-augmented large language model for accurate precision oncology medicine recommendations. Cancer cell **44**(3), 676–685 (2026)

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) https://doi.org/10.48550/arXiv.1706.03762

[21] Lv, F., Wang, R., Jing, L., Dai, P.: Hist: Hierarchical and sequential transformer for image captioning. IET Computer Vision **18**(7), 1043–1056 (2024) https://doi.org/10.1049/cvi2.12305

[22] Abdulgalil, H.D., Basir, O.A.: Next-generation image captioning: A survey of methodologies and emerging challenges from transformers to multimodal large language models. Natural Language Processing Journal **12**, 100159 (2025) https://doi.org/10.1016/j.nlp.2025.100159

[23] Li, J., Vo, D.M., Sugimoto, A., Nakayama, H.: Evcap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13733–13742 (2024). https://doi.org/10.1109/CVPR52733.2024.01303

[24] Zeng, Z., Xie, Y., Zhang, H., Chen, C., Chen, B., Wang, Z.: Meacap: Memory-augmented zero-shot image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14100–14110 (2024). https://doi.org/10.1109/CVPR52733.2024.01337

[25] Pourkeshavarz, M., Nabavi, S., Moghaddam, M.E., Shamsfard, M.: Stacked cross-modal feature consolidation attention networks for image captioning. Multimedia Tools and Applications **83**(4), 12209–12233 (2024) https://doi.org/10.1007/s11042-023-15869-x

[26] Zhu, D., Shen, X., Li, X., Elhoseiny, M., *et al.*: Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: International Conference on Learning Representations, vol. 2024, pp. 18378–18394 (2024)

[27] Fei, J., Wang, T., Zhang, J., He, Z., Wang, C., Zheng, F.: Transferable decoding with visual entities for zero-shot image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3136–3146 (2023). https://doi.org/10.48550/arXiv.2307.16525

[28] Zhao, F., Yu, Z., Wang, T., Zhao, H.: Meshed context-aware beam search for image captioning. Entropy **26**(10), 866 (2024) https://doi.org/10.3390/e26100866

[29] Patra, B., Kisku, D.R.: Enhancing image captioning with asynchronous dual attention vision transformer. Intelligent Data Analysis **30**(2), 388–414 (2026) https://doi.org/10.1177/1088467X251351627

[30] Patel, M., Vaka, P.D.R., Singh, D.P., Choudhary, J., Solanki, S.: Enhanced image captioning with advanced context-aware object relational model. Discover Computing **28**(1), 337 (2025) https://doi.org/10.1007/s10791-025-09839-0

[31] Chan, B.J., Chen, C.-T., Cheng, J.-H., Huang, H.-H.: Don't do rag: When

cache-augmented generation is all you need for knowledge tasks. In: Companion Proceedings of the ACM on Web Conference 2025, pp. 893–897 (2025). https://doi.org/10.1145/3701716.3715490

[32] Shazeer, N.: Glu variants improve transformer. arXiv preprint arXiv:2002.05202 (2020) https://doi.org/10.48550/arXiv.2002.05202

[33] Inan, H., Khosravi, K., Socher, R.: Tying word vectors and word classifiers: A loss framework for language modeling. (2016). https://doi.org/10.48550/arXiv.1611.01462

[34] Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015) https://doi.org/10.48550/arXiv.1504.00325

[35] Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)

[36] Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the association for computational linguistics **2**, 67–78 (2014) https://doi.org/10.1162/tacl_a_00166

[37] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002). https://doi.org/10.3115/1073083.1073135

[38] Banerjee, S., Lavie, A.: 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, pp. 65–72 (2005). https://doi.org/10.3115/1626355.1626389