

Article

Global-to-Local Visual Conditioning for Image Captioning with a Frozen Vision–Language Model

Sohyun Lee  and Daeyoung Heo *

Department of Computer Science, Kookmin University, Seoul 08826, Republic of Korea;
sohyunlee@kookmin.ac.kr

* Correspondence: dyheo@kookmin.ac.kr

Featured Application

The proposed framework provides a lightweight image captioning solution for vision–language systems that employ frozen pretrained vision and language models. It is applicable to parameter-efficient adaptation scenarios where only a small number of trainable parameters can be updated while preserving the pretrained backbone architecture.

Abstract

Image captioning with pretrained vision and language models often requires substantial model adaptation, while lightweight settings restrict the number of components that can be updated. This study investigates an input-level visual conditioning approach for image captioning using a frozen ResNet-50 image encoder and a frozen Phi-3 Mini language model. A global image representation is first projected into ten continuous visual tokens. A gated local residual module is then introduced to incorporate spatial features from the final ResNet feature map without modifying either pretrained backbone. The framework was evaluated on the MS COCO Karpathy split under a fixed training and decoding protocol. The global visual-conditioning baseline achieved a CIDEr score of 0.8382 on the test split. Adding the gated local residual module increased CIDEr to 0.8693, while adding only 3.95 million stage-specific trainable parameters. Under a near-matched trainable-parameter budget, a rank-10 QKV-only LoRA control achieved a CIDEr score of 0.8439, whereas direct and gated local residual conditioning achieved 0.8611 and 0.8693, respectively. These results are limited to the evaluated frozen ResNet-50–Phi-3 configuration and do not imply a general advantage over LoRA variants. As an additional sequence-level optimization experiment, self-critical sequence training increased CIDEr marginally from 0.8693 to 0.8704 without changing the inference architecture. Caption quality was further examined using a structured LLM-based evaluation over all 5000 images in the Karpathy test split, assessing semantic accuracy, information coverage, fluency, and reference-based unsupported content. A zero-shot evaluation on Flickr30k showed a marked performance decrease, indicating limited cross-dataset generalization. Overall, the experiments provide an empirical analysis of global and local visual conditioning for parameter-efficient caption generation with frozen pretrained components.



Academic Editor: João M. F. Rodrigues

Received: 28 May 2026

Revised: 22 July 2026

Accepted: 24 July 2026

Published: 28 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: image captioning; visual conditioning; small language model; parameter-efficient adaptation; gated residual fusion; frozen language model; qualitative evaluation

1. Introduction

Image captioning aims to generate natural language descriptions that summarize the semantic content of an image by jointly modeling visual and linguistic information. Recent advances in pretrained vision–language models have substantially improved caption generation by combining powerful visual encoders with large language models. Rather than training these models from scratch, many recent studies employ pretrained vision and language components and adapt them to downstream tasks through lightweight trainable modules.

As pretrained vision–language models become increasingly larger, parameter-efficient adaptation has emerged as a practical approach for reducing the number of trainable parameters while preserving pretrained knowledge. Under such settings, the effectiveness of image captioning depends not only on the pretrained models themselves but also on how visual information is transferred from the vision encoder to the language model. Designing an efficient visual conditioning mechanism therefore becomes an important problem when both pretrained backbones remain frozen.

While compact global image representations provide high-level semantic information, they may not always sufficiently capture fine-grained spatial evidence that contributes to object attributes, local interactions, and complex scene composition. This motivates the following research question:

“Can complementary local visual information improve caption generation while maintaining a frozen vision–language architecture and a parameter-efficient adaptation setting?”

To investigate this question, this work studies a lightweight image captioning framework based on a frozen ResNet-50 [1] image encoder and a frozen Phi-3 Mini [2] language model. The study examines how different visual conditioning strategies influence caption generation under a unified experimental setting while keeping both pretrained backbones unchanged. In addition to global and local visual conditioning, decoder adaptation using Low-Rank Adaptation (LoRA) [3] and sequence-level optimization through Self-Critical Sequence Training (SCST) [4] are evaluated to better understand their respective contributions under the same frozen architecture.

The proposed framework is evaluated on the MS COCO Karpathy split [5,6] using standard automatic captioning metrics. To complement quantitative evaluation, we further conduct a structured qualitative analysis over all 5000 images in the Karpathy test split using a fixed large language model (LLM)-based evaluation protocol with predefined assessment criteria. Cross-domain generalization is additionally examined through zero-shot evaluation on Flickr30k [7].

The main contributions of this work are summarized as follows:

- We investigate a lightweight image captioning framework that combines a frozen ResNet-50 image encoder and a frozen Phi-3 Mini language model through parameter-efficient visual conditioning.
- We experimentally compare global and local visual conditioning with a parameter-matched LoRA control and evaluate the effect of sequence-level optimization under a unified experimental setting.
- We provide a comprehensive evaluation using both standard automatic metrics and a structured qualitative analysis over the complete MS COCO Karpathy test split, including representative examples and systematic observations of generated caption quality.

2. Related Work

Image captioning research has progressed from task-specific encoder–decoder architectures to large-scale vision–language pretraining and parameter-efficient adaptation. Within this broader landscape, recent approaches can be distinguished by whether they adapt the

language model or improve the visual-conditioning interface. The following subsections review these developments and clarify the position of the proposed framework.

2.1. Evolution of Image Captioning

Early neural image-captioning systems were commonly formulated as encoder–decoder models. Show and Tell combined a convolutional image encoder with a recurrent language decoder, establishing an end-to-end framework for mapping images to natural-language descriptions [8]. Show, Attend and Tell subsequently introduced visual attention, allowing the language decoder to emphasize different spatial regions of the image at individual generation steps rather than relying exclusively on a single global representation [9].

The introduction of the Transformer [10] shifted sequence generation away from recurrent processing toward attention-based modeling. Transformer-based captioning architectures, such as the Meshed-Memory Transformer, further improved the modeling of relationships among image regions and enabled the decoder to use representations from multiple visual-encoding levels [11]. These developments established two principles that remain relevant to recent captioning systems: effective caption generation depends on both strong linguistic modeling and the preservation of sufficiently detailed visual information.

More recently, the field has moved from training captioning architectures primarily on task-specific datasets toward adapting large pretrained vision and language components. This shift has expanded image captioning from a narrowly defined generation task into part of a broader vision–language modeling paradigm.

2.2. Vision–Language Pretraining and Frozen Backbone Integration

Large-scale vision–language pretraining has substantially influenced recent caption-generation systems. BLIP [12] introduced a unified framework for vision–language understanding and generation and used bootstrapped caption generation and filtering to improve learning from noisy image–text data. BLIP-2 reduced the cost of multimodal pretraining by connecting a frozen image encoder and a frozen large language model through a lightweight Querying Transformer, or Q-Former [13]. Instruct BLIP [14] extended this design through vision–language instruction tuning and an instruction-aware Q-Former.

Other multimodal systems have pursued broader task generality. Flamingo connects pretrained visual and language components through cross-attention layers and supports few-shot inference over interleaved visual and textual inputs [15]. LLaVA uses a lightweight visual-to-language projection followed by visual instruction tuning to connect visual features with a pretrained language model [16]. Kosmos-2 extends multimodal generation toward explicit phrase and region grounding [17]. Although these systems differ in training scale, architecture, and task scope, they demonstrate the importance of the interface that maps visual representations into the language-model space.

A related line of research studies multimodal learning while preserving pretrained language-model parameters. Frozen transforms visual inputs into continuous representations that can condition a frozen autoregressive language model [18]. Such approaches suggest that useful multimodal behavior can be obtained without updating the complete language model, provided that the visual-to-language interface is sufficiently expressive.

Compact language models also make frozen-backbone integration increasingly practical. Phi-3 Mini is a 3.8-billion-parameter language model designed to provide strong language capability at a substantially smaller scale than many large language models [2]. However, Phi-3 Mini is a text language model rather than a pretrained vision–language model. The present study therefore combines it with an independently pretrained visual encoder and explicitly learns the visual-conditioning interface between the two frozen components.

Figure 1 summarizes the reviewed research landscape and positions the proposed framework between language-side adaptation and visual conditioning under frozen pre-trained backbones. The proposed framework keeps both backbones frozen and evaluates global-only conditioning, direct and gated local residual conditioning, and a QKV-LoRA control under the same MS COCO Karpathy protocol.

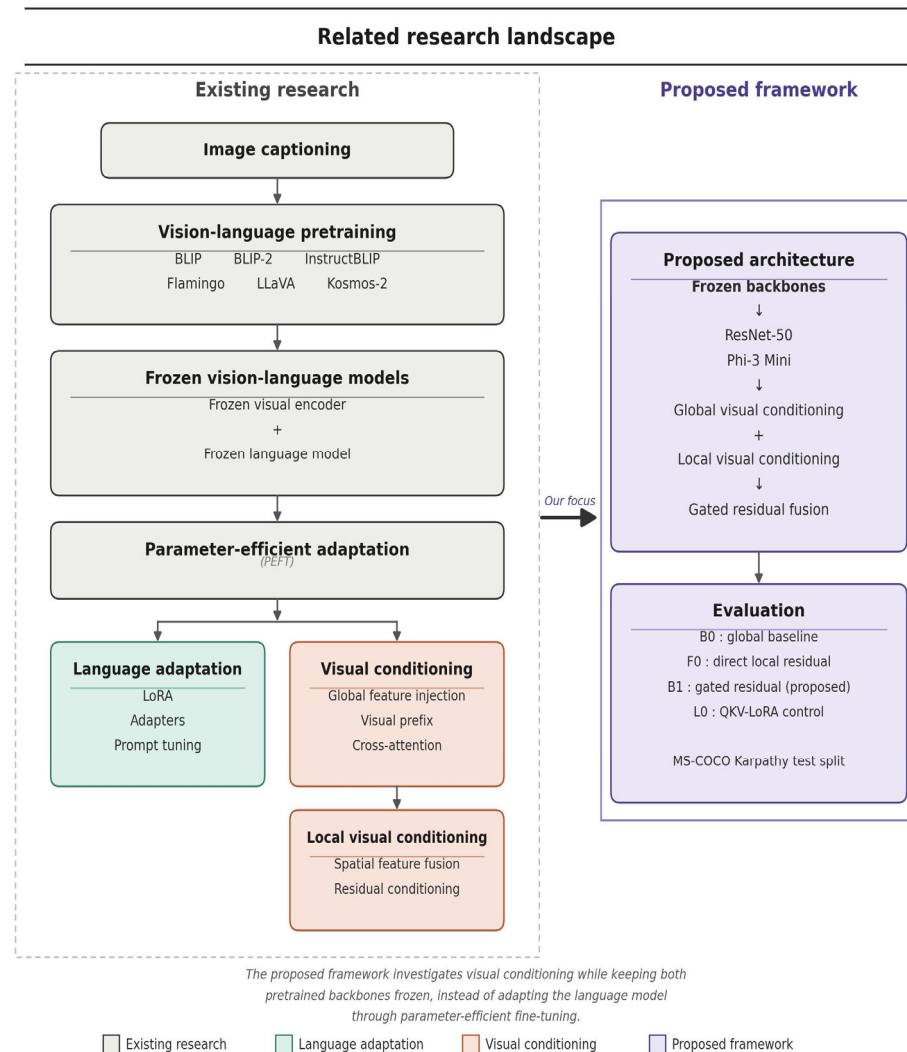


Figure 1. Research landscape of image captioning and parameter-efficient vision–language adaptation, with the position of the proposed framework.

2.3. Parameter-Efficient Language-Side Adaptation

Parameter-efficient fine-tuning reduces adaptation cost by updating only a small number of task-specific parameters. Adapter-based methods insert lightweight bottleneck modules into pretrained networks while leaving the original model parameters unchanged [19]. Prefix Tuning optimizes continuous task-specific representations that are injected into Transformer layers while the underlying language model remains frozen [20]. Prompt Tuning instead learns continuous input prompts that condition a frozen model on a downstream task [21].

LoRA represents weight updates through trainable low-rank matrices added to selected pretrained layers [3]. Unlike soft-prompt approaches, LoRA directly adapts designated projection matrices while retaining the original pretrained weights. It has consequently become a representative control for evaluating whether performance improvements arise from language-model adaptation.

The objective of the present study is not to compare the full range of PEFT algorithms. Rather, it is to isolate the contribution of visual conditioning under a fixed frozen-backbone configuration. A QKV-only LoRA model is therefore used as a representative language-side control. This comparison asks whether a similar stage-specific trainable-parameter budget is more beneficial when assigned to decoder adaptation or to the proposed local visual-conditioning module. The conclusions are restricted to the evaluated LoRA configuration and are not generalized to other ranks, target modules, or PEFT methods.

2.4. Visual Conditioning for Frozen Language Models

Visual conditioning provides an alternative to adapting the internal parameters of a language model. Instead of changing the language model itself, visual representations are transformed into continuous embeddings or tokens that can be processed through its existing input or attention interfaces. Frozen [18], BLIP-2 [13], and related architectures demonstrate different forms of visual feature projection, query-based extraction, and cross-attention-based integration.

For image captioning specifically, ClipCap maps a pretrained visual representation to a sequence of continuous embeddings used to condition a pretrained language model [22]. This approach showed that caption generation can remain viable even when only the visual mapping network is trained and both the image encoder and language model are frozen. However, methods based predominantly on pooled global representations may compress or discard spatial information needed to distinguish small objects, attributes, and local relationships.

Visual conditioning mechanisms can therefore be considered at two complementary levels. Global conditioning provides an overall semantic representation of the image, whereas local conditioning retains spatially distributed evidence from the visual feature map. Directly adding local information may improve detail coverage, but an unconstrained local residual may also disrupt the stable global representation. This motivates mechanisms that regulate how strongly local evidence modifies global conditioning.

The proposed framework addresses this problem using a frozen ResNet-50 visual encoder and a frozen Phi-3 Mini language model [1,2]. A global image representation is first mapped to continuous visual-conditioning tokens. Spatial ResNet features are then processed through learnable queries and multi-head cross-attention, after which the resulting local representation is incorporated as a gated residual update. The global-only baseline (B0), direct local residual configuration (F0), and gated local residual configuration (B1) separate the contribution of local evidence from that of the gating mechanism. The QKV-LoRA control (L0) further distinguishes visual-side conditioning from language-side adaptation under the same experimental framework.

Accordingly, this study does not aim to establish a general ranking among contemporary vision–language models or PEFT methods. Its narrower objective is to determine whether lightweight global-to-local visual conditioning improves caption generation when the visual encoder and language model remain completely frozen.

3. Proposed Method

3.1. Overall Architecture

Figure 2 illustrates the overall architecture of the proposed image captioning framework. The framework consists of three principal components: (1) a frozen ResNet-50 [1] image encoder for visual feature extraction, (2) lightweight trainable visual conditioning modules, and (3) a frozen Phi-3 Mini [2] language model for caption generation. During training, only the visual conditioning modules are optimized, while all parameters of the pretrained vision encoder and language model remain fixed.

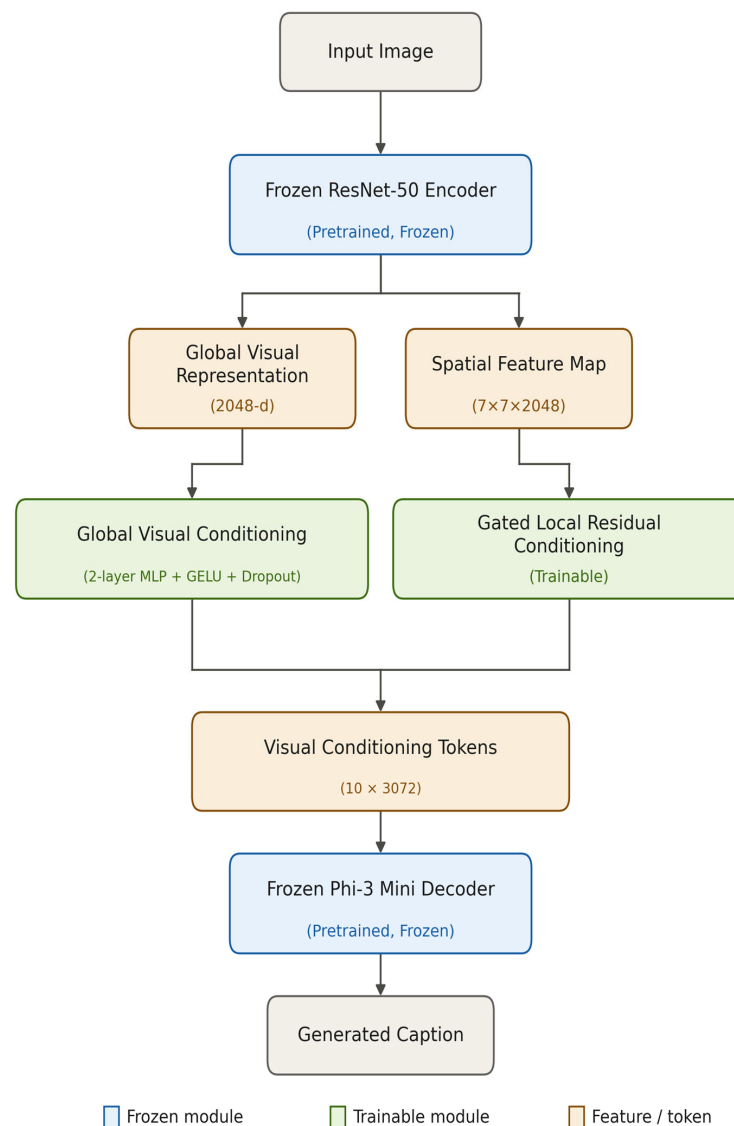


Figure 2. Overall architecture of the proposed framework with frozen ResNet-50 and Phi-3 Mini backbones.

Given an input image, the frozen ResNet-50 encoder first extracts both a global image representation and a spatial feature map from its final convolutional stage. The global representation provides compact semantic information describing the overall image content, whereas the spatial feature map preserves local visual evidence corresponding to object regions and scene structure. These complementary visual representations are processed by separate trainable conditioning modules before being provided to the language model.

The proposed framework adopts a two-stage visual conditioning strategy. First, the global image representation is transformed into a sequence of continuous visual tokens through a lightweight projection network. These visual tokens are prepended to the textual input and serve as the initial visual context for caption generation. Subsequently, local spatial features extracted from the final convolutional feature map are incorporated through a gated residual conditioning module. This module introduces complementary local visual information while preserving the semantic context established by the global visual tokens.

Unlike conventional Prefix Tuning [20], which injects trainable prefix representations into multiple Transformer layers, the proposed framework performs visual conditioning only at the input interface between the image encoder and the language model. Consequently, visual information is provided before language decoding begins, while the internal

architecture and parameters of both pretrained backbones remain unchanged. This design allows the framework to investigate the effectiveness of lightweight visual conditioning independently of modifications to the pretrained vision or language models.

The proposed architecture was designed to systematically evaluate different visual conditioning strategies under a unified experimental setting. The global visual conditioning model serves as the baseline configuration, while the gated local residual module extends the baseline by incorporating additional spatial information. Decoder adaptation using LoRA [3] and sequence-level optimization through Self-Critical Sequence Training (SCST) [4] are evaluated separately to analyze their respective contributions without altering the overall architecture.

3.2. Global Visual Conditioning

The global visual conditioning module converts an image-level representation extracted by the frozen ResNet-50 encoder into a sequence of continuous visual tokens that provide the initial image context to the frozen language model. This module operates only at the input interface of Phi-3 Mini and does not modify the internal layers of either pretrained backbone.

$$\mathbf{v}_g \in \mathbb{R}^{2048}$$

A trainable projection network maps $\mathbf{v}_g$ to the embedding dimension of Phi-3 Mini. The projector consists of a fully connected layer preserving the 2048-dimensional feature size, a GELU activation, dropout, and a second fully connected layer producing $L_v d$ output values:

$$\mathbf{h}_g = \text{Dropout}(\text{GELU}(\mathbf{W}_1 \mathbf{v}_g + \mathbf{b}_1)),$$

$$\mathbf{z}_g = \mathbf{W}_2 \mathbf{h}_g + \mathbf{b}_2,$$

where

$$\mathbf{W}_1 \in \mathbb{R}^{2048 \times 2048}, \quad \mathbf{W}_2 \in \mathbb{R}^{(L_v d) \times 2048}.$$

The projected vector is reshaped into a sequence of global visual conditioning tokens:

$$\mathbf{T}_g = \text{reshape}(\mathbf{z}_g) \in \mathbb{R}^{L_v \times d},$$

where $L_v = 10$ is the number of visual conditioning tokens and $d = 3072$ is the hidden dimension of Phi-3 Mini. Dropout is set to 0.1 during training.

The resulting visual tokens are prepended to the embedding sequence of the fixed textual prompt, “Describe the image in one concise sentence.” During cross-entropy training, the target-caption embeddings follow the visual tokens and prompt embeddings:

$$\mathbf{X}_{train} = [\mathbf{T}_g; \mathbf{E}_p; \mathbf{E}_y],$$

where $\mathbf{E}_p$ denotes the prompt embeddings and $\mathbf{E}_y$ denotes the target-caption embeddings. The loss is computed only over the target-caption positions; the visual-token and prompt positions are masked from the language-modeling objective. During inference, only the visual tokens and prompt embeddings are supplied:

$$\mathbf{X}_{infer} = [\mathbf{T}_g; \mathbf{E}_p].$$

The complete training procedure for the global visual-prefix anchor is provided in Algorithm A1 in Appendix B.

3.3. Gated Local Residual Conditioning

While the global visual conditioning module provides an image-level semantic representation, it does not explicitly preserve spatial information contained in the intermediate feature map of the visual encoder. To complement the global representation without modifying either pretrained backbone, we introduce a lightweight gated local residual conditioning module. As illustrated in Figure 3, spatial features are projected into local embeddings and summarized by learnable queries through multi-head cross-attention. The resulting representation is mapped to the Phi-3 embedding space through a zero-initialized residual projection and added to the global visual tokens after modulation by a learnable scalar gate.

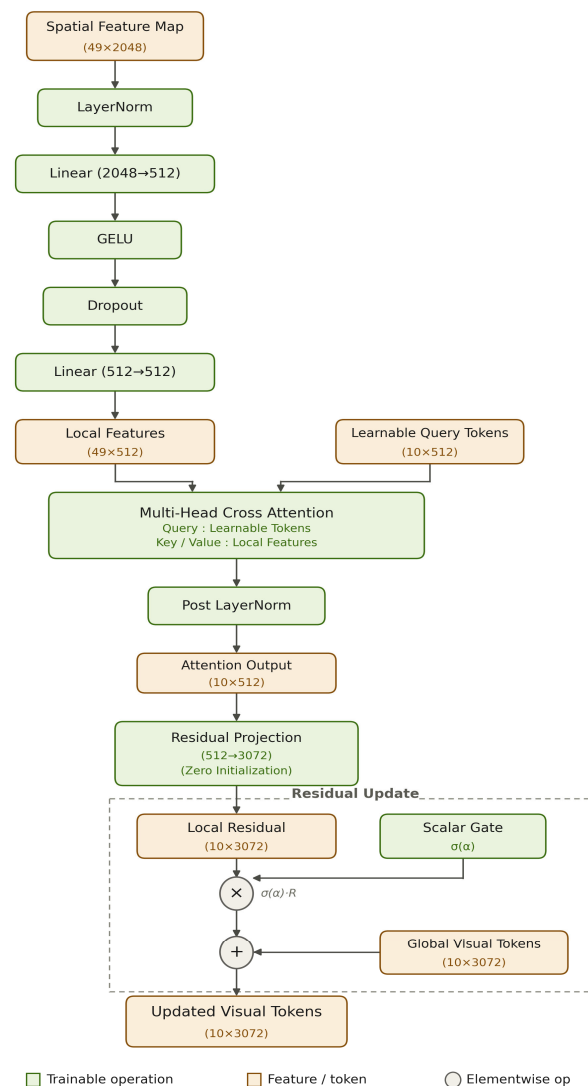


Figure 3. Architecture of the proposed gated local residual conditioning module with cross-attention and gated residual fusion.

Let $F \in \mathbb{R}^{49 \times 2048}$ denote the spatial feature map obtained from the final convolutional stage of the frozen ResNet-50 encoder after flattening the 7×7 spatial locations. The spatial features are first normalized and projected into a lower-dimensional representation,

$$F_l = f_{local}(F) \in \mathbb{R}^{49 \times 512},$$

where f_{local} consists of Layer Normalization followed by two linear layers with GELU activation and dropout.

A set of learnable query tokens,

$$\mathbf{Q} \in \mathbb{R}^{L_v \times 512},$$

is introduced to summarize the spatial features through multi-head cross-attention,

$$\mathbf{A} = \text{CrossAttn}(\mathbf{Q}, \mathbf{F}_l, \mathbf{F}_l),$$

where the learnable query tokens serve as the query, while the projected local features are used as both the keys and values. The attention output is subsequently normalized and projected into the embedding dimension of the frozen Phi-3 Mini decoder,

$$\mathbf{R} = f_{\text{proj}}(\mathbf{A}) \in \mathbb{R}^{L_v \times 3072},$$

forming a local residual representation.

Unlike directly replacing the global conditioning tokens, the proposed method updates them through a gated residual formulation,

$$\mathbf{T}_{\text{updated}} = \mathbf{T}_g + \sigma(\alpha)\mathbf{R},$$

where $\mathbf{T}_g$ denotes the global visual conditioning tokens, $\mathbf{R}$ is the local residual representation, $\sigma(\cdot)$ is the sigmoid function, and α is a learnable scalar parameter controlling the contribution of the local branch.

To preserve the behavior of the global baseline at the beginning of optimization, the weights and biases of the residual projection layer are initialized to zero. Consequently, the initial residual output becomes zero, and the updated visual tokens are initially identical to the global visual conditioning tokens. During training, the local branch gradually learns residual corrections without modifying either pretrained backbone.

The resulting updated visual conditioning tokens are subsequently concatenated with the fixed textual prompt and supplied to the frozen Phi-3 Mini decoder for caption generation. The training procedures for the direct and gated local residual conditioning variants are detailed in Algorithm A2 and Algorithm A3, respectively, in Appendix B.

3.4. Training Strategy

The proposed framework is trained by optimizing only the lightweight visual conditioning modules while keeping both the pretrained ResNet-50 image encoder and the Phi-3 Mini language model frozen throughout training. Specifically, the trainable parameters consist of the global visual conditioning projector described in Section 3.2 and the gated local residual conditioning module introduced in Section 3.3. This optimization strategy isolates the learning process to the visual conditioning interface while preserving the pretrained visual and linguistic representations of both backbones.

During supervised training, the updated visual conditioning tokens are concatenated with a fixed textual prompt and the ground-truth caption embeddings before being fed into the frozen language model. The textual prompt is fixed as: “Describe the image in one concise sentence.” for all training and evaluation experiments. The complete input embedding sequence is therefore constructed as

$$\mathbf{X}_{\text{train}} = [\mathbf{T}_{\text{updated}}; \mathbf{E}_{\text{prompt}}; \mathbf{E}_{\text{caption}}],$$

where $\mathbf{T}_{\text{updated}}$ denotes the updated visual conditioning tokens, $\mathbf{E}_{\text{prompt}}$ denotes the embeddings of the fixed textual prompt, and $\mathbf{E}_{\text{caption}}$ denotes the embeddings of the ground-truth caption.

Training follows the standard autoregressive teacher-forcing paradigm, in which each target token is predicted conditioned on the updated visual conditioning tokens, the fixed prompt, and the preceding ground-truth caption tokens. Let

$$Y = (y_1, \dots, y_N)$$

denote the target caption. The model is trained by minimizing the autoregressive cross-entropy objective,

$$\mathcal{L}_{CE} = -\sum_{t=1}^N \log P(y_t | y_{<t}, T_{updated}).$$

To ensure that optimization focuses only on caption generation, the visual conditioning tokens and the prompt tokens are excluded from the language-modeling loss through masking. Consequently, gradients are computed only from the target-caption tokens, while the visual conditioning modules are optimized through backpropagation across the frozen language model.

During inference, the ground-truth caption embeddings are removed, and only the updated visual conditioning tokens together with the fixed textual prompt are provided to the frozen Phi-3 Mini decoder,

$$X_{infer} = [T_{updated}; E_{prompt}].$$

The decoder then generates the caption autoregressively until the end-of-sequence token is produced or the predefined decoding termination criterion is satisfied. The complete caption-generation and inference procedure for the evaluated configurations is provided in Algorithm A4 in Appendix B.

4. Experimental Setup

4.1. Datasets

The proposed framework was primarily evaluated on the MS COCO 2014 Karpathy split [5,6], a commonly used benchmark setting for image captioning. The split contains 82,783 training images, 5000 validation images, and 5000 test images, with five human-annotated reference captions available for each image. The training split was used for model optimization, the validation split was used for checkpoint and decoding-configuration selection, and the test split was reserved exclusively for final evaluation.

An additional cross-dataset evaluation was conducted on Flickr30k [7] to examine the transfer of the COCO-trained visual conditioning model to a different captioning dataset. The evaluation used the Flickr30k test split containing 1000 images. The checkpoint achieving the highest validation CIDEr score on the MS COCO validation split was directly evaluated on Flickr30k without any additional fine-tuning or parameter updates. Accordingly, this experiment is reported as a zero-shot cross-dataset evaluation, rather than as an in-domain Flickr30k result. The datasets used in this study are summarized in Table 1.

Table 1. Summary of the datasets used in this study.

Dataset	Train	Validation	Test	Usage
MS COCO 2014 (Karpathy split)	82,783	5000	5000	Supervised training, validation, and in-domain evaluation
Flickr30k	–	–	1000	Zero-shot Generalization

Unless otherwise stated, the main quantitative results reported in this paper refer to the MS COCO Karpathy test split. The Flickr30k results are presented separately to analyze cross-dataset generalization, rather than in-domain captioning performance.

4.2. Experimental Settings

Experimental environment: All experiments were implemented in Python 3.12.13 using PyTorch 2.11.0 and torchvision 0.26.0 with CUDA 12.8. Training and evaluation were conducted on a single NVIDIA A100-SXM4 GPU with 80 GB of memory. The visual encoder used ResNet-50 initialized with the ImageNet-1K V2 pretrained weights. The Phi-3 Mini 3.8 B language model was loaded using 4-bit NF4 quantization with FP16 computation and remained frozen throughout all experiments. The trainable visual conditioning modules were maintained in FP32, and their outputs were cast to the input embedding dtype of the frozen language model before decoding.

Training configuration: Training was performed in two supervised stages. In the first stage, the global visual conditioning projector was trained for five epochs using AdamW with a learning rate of 1×10^{-4} , a weight decay of 0.01, and gradient clipping with a maximum norm of 1.0. The micro-batch size was 2, and gradient accumulation over 16 steps resulted in an effective batch size of 32. The projector dropout probability was 0.1, and the visual conditioning length was fixed to 10 tokens. One of the five reference captions associated with each training image was used in each epoch according to a fixed cyclic schedule. The checkpoint achieving the highest proxy CIDEr score was selected and fixed as the global anchor for the subsequent controlled architecture experiments. In the second stage, the selected global projector was frozen and used as the global anchor. Only the gated local residual conditioning module was optimized for five epochs. AdamW was used with a learning rate of 5×10^{-5} , a weight decay of 0.01, a batch size of 32, and a maximum gradient norm of 1.0. The local hidden dimension was 512, the number of cross-attention heads was 8, and dropout was set to 0.1. The scalar gate logit was initialized to -4.0 , while the residual projection weights and biases were initialized to zero. The checkpoint achieving the highest validation CIDEr score was selected as the final supervised (cross-entropy, XE) model. Unless otherwise stated, all quantitative results reported in this paper correspond to this supervised model. The principal training configurations of the two supervised stages are summarized in Table 2.

Inference configuration: Unless otherwise specified, the same decoding configuration was used for the MS COCO evaluation and the Flickr30k zero-shot evaluation. Captions were generated using beam search with four beams, a maximum of 13 newly generated tokens, a minimum of 5 newly generated tokens, a length penalty of 0.7, and a no-repeat 3-gram constraintstage 1 (Global Conditioning). This decoding configuration was selected using the MS COCO validation split and fixed before final test evaluation. Experiments varying the visual conditioning length and decoding length are presented separately in the ablation study. The checkpoint and decoding length were determined according to the highest validation CIDEr score, following the standard model-selection protocol adopted throughout this study. Qualitative evaluation was conducted separately using a maximum generation length of 20 tokens to reduce truncation artifacts and better reflect human-readable caption quality.

SCST configuration: Self-Critical Sequence Training (SCST) was conducted as a post-hoc extension rather than as part of the main model. The selected supervised local-fusion model was used for initialization. The ResNet-50 encoder, the Phi-3 Mini language model, and the global visual conditioning projector remained frozen, while only the gated local residual conditioning module was updated. SCST was performed using a randomly selected subset of 4096 MS COCO training images, comprising 1024 mini-batches (512 optimizer updates). The learning rate was set to 1×10^{-6} . The reward was defined as the per-sample difference between the CIDEr scores of a sampled caption and its corresponding greedy-decoded caption. Sampling employed a temperature of 0.8 and top- p of 0.9, with a maximum rollout length of 13 tokens. To stabilize policy optimization, a

cross-entropy regularization term with a weight of 0.05 was added to the policy-gradient objective. The resulting SCST model was evaluated under the same fixed-decoding test protocol as the supervised model and is reported separately as a post-hoc extension.

Table 2. Summary of the training configurations for the global and local conditioning stages.

Setting	Stage 1 (Global Conditioning)	Stage 2 (Local Conditioning)
Trainable module	Global visual conditioning projector	Gated local residual conditioning module
Frozen modules	ResNet-50, Phi-3 Mini	ResNet-50, Phi-3 Mini, Global visual conditioning projector
Epochs	5	5
Optimizer	AdamW	AdamW
Learning rate	1×10^{-4}	5×10^{-5}
Weight decay	0.01	0.01
Effective batch size	32 (micro-batch $2 \times$ accumulation 16)	32
Gradient clipping	Max norm = 1.0	Max norm = 1.0
Dropout	0.1	0.1
Visual conditioning length	10 tokens	10 tokens
Additional settings	Stage 1 (Global Conditioning)	Stage 2 (Local Conditioning)
Checkpoint selection	Highest proxy CIDEr	Highest validation CIDEr

4.3. Evaluation Metrics

Automatic Metrics: The primary quantitative evaluation was conducted using the standard image captioning metrics implemented in the pycocoevalcap toolkit after PTB tokenization. The reported metrics include BLEU-1, BLEU-2, BLEU-3, BLEU-4 [23], METEOR [24], ROUGE-L [25], and CIDEr [26]. Unless otherwise specified, all quantitative results were computed on the MS COCO Karpathy test split using the standard evaluation protocol. Among these metrics, CIDEr was adopted as the primary criterion for model selection because it was specifically designed for image captioning by measuring consensus with multiple human reference captions using TF-IDF-weighted n-gram matching. Consequently, CIDEr has become the de facto benchmark metric for checkpoint selection and model comparison in the MS COCO captioning literature. Therefore, the checkpoint achieving the highest validation CIDEr score was selected for subsequent quantitative evaluation unless otherwise stated. The maximum decoding length was also determined using the MS COCO validation split. Candidate decoding lengths were evaluated under an identical beam-search configuration, and a maximum of 13 newly generated tokens yielded the highest validation CIDEr score. Accordingly, all quantitative results reported in this paper were generated using this fixed decoding length. Because the canonical decoding configuration was optimized for CIDEr-based quantitative evaluation, a separate decoding setting was adopted for qualitative analysis, where human readability and sentence completeness are more important than maximizing automatic metric scores. Because automatic evaluation metrics do not necessarily reflect human-perceived caption quality, qualitative evaluation was conducted using a maximum generation length of 20 tokens. This longer decoding configuration reduces truncation artifacts and provides more natural sentence completion, enabling a more reliable assessment of semantic correctness, detail coverage, and overall fluency.

LLM-based Qualitative Evaluation: To complement the automatic metrics, we adopt an LLM-based qualitative evaluation protocol, following the growing body of work showing that strong LLMs can serve as scalable, human-aligned evaluators of open-ended generation [27,28] and, specifically for image captioning, that LLM-based judgment corre-

lates more closely with human quality judgments than n-gram overlap metrics [29]. The evaluation was conducted over all 5000 images in the MS COCO Karpathy test split, employing GPT-5.4-mini with a fixed evaluation prompt and deterministic decoding. For each image, the generated caption was compared with the corresponding five human reference captions. The evaluation assessed four aspects:

- Semantic Correctness, measuring whether the generated caption correctly describes the main objects, actions, and scene reflected in the reference captions.
- Detail Coverage, evaluating how sufficiently the generated caption conveys useful visual information without requiring every detail appearing in the reference captions.
- Fluency and Completeness, assessing grammatical correctness and sentence completeness.
- Reference Compatibility, determining whether the generated caption is supported by, compatible with, or contradicts the reference captions.

The evaluation additionally classified the reference compatibility status, primary issue, truncation severity, and unsupported-content type according to predefined criteria. Based on the combined assessment, each caption was assigned one of four overall quality levels (Excellent, Good, Adequate, or Poor). The complete evaluation prompt and detailed criteria are provided in Appendix D. The principal components of the evaluation protocol are summarized in Table 3.

Table 3. Summary of the LLM-based qualitative evaluation protocol.

Component	Description
Evaluation model	GPT-5.4-mini
Evaluation scope	Generated captions from the MS COCO Karpathy test split
Evaluation input	One generated caption and five human reference captions per image
Evaluator limitation	The evaluator does not observe the original image
Reference assumption	Reference captions are incomplete descriptions and are not treated as exhaustive annotations
Semantic Correctness	Integer score from 1 to 5 for the correctness of the main objects, actions, and scene
Detail Coverage	Integer score from 1 to 5 for the useful amount of caption content
Fluency and Completeness	Integer score from 1 to 5 for grammatical fluency and sentence completeness
Truncation severity	None, Mild, or Severe
Reference compatibility	Supported, Compatible Extra Detail, Plausible Unverified, Weakly Unsupported, Major Unsupported, or Contradiction
Unsupported-content type	Assigned only for Weakly Unsupported, Major Unsupported, or Contradiction
Primary issue	None, Truncation, Missing Detail, Weak Unsupported Detail, Unsupported Object, Unsupported Attribute, Unsupported Action, Unsupported Scene Detail, Contradiction, Generic, Grammar, or Multiple Issues
Final quality	Excellent, Good, Adequate, or Poor
Evaluation prompt	Fixed prompt provided in Appendix D
LLM decoding	Deterministic evaluation with temperature = 0
Output format	Valid JSON object only

4.4. Experimental Configurations

To systematically evaluate the contribution of each component, four experimental configurations were considered in this study.

- Before these formal comparisons, the global visual conditioning projector was trained and the checkpoint with the highest proxy CIDEr was selected as the frozen G0

anchor. B0 denotes the formal Karpathy test evaluation of this global-only architecture, whereas L0, F0, and B1 use the same frozen global anchor for controlled comparison.

- B0 (Global Visual Conditioning): Only the global visual conditioning projector was trained while both pretrained backbones remained frozen. This configuration serves as the baseline for evaluating the contribution of local visual conditioning.
- L0 (QKV-LoRA): Low-Rank Adaptation (LoRA) [3] was applied to the query, key, and value projection layers of the frozen Phi-3 Mini decoder. The number of trainable parameters was chosen to be comparable to the proposed local conditioning module, providing a parameter-efficient control experiment.
- F0 (Direct Local Residual): Local spatial features extracted from the ResNet-50 [1] feature map were directly added to the global visual conditioning tokens without gated modulation.
- B1 (Gated Local Residual): The proposed model employs gated local residual conditioning, where a learnable scalar gate controls the contribution of the local residual while preserving the global visual conditioning. This configuration represents the primary supervised model investigated in this study.

In addition, a post-hoc sequence-level optimization experiment was conducted using Self-Critical Sequence Training (SCST) [4]. SCST was initialized from the selected B1 model while keeping the network architecture unchanged, allowing the effect of sequence-level optimization to be evaluated independently of the proposed visual conditioning architecture. The experimental configurations are summarized in Table 4.

Table 4. Experimental configurations considered in this study.

Configuration	Global Anchor	Local Feature	Additional Adaptation	Role
B0	Frozen G0 anchor	No	None	Formal global-only baseline
L0		No	Phi-3 QKV-LoRA, rank 10	PEFT control
F0		Yes	Direct local residual	Architectural ablation
B1		Yes	Gated local residual	Proposed main model
B2		Yes	B1 followed by SCST	Post-hoc extension

5. Results and Discussion

5.1. Main Results

The primary quantitative evaluation was conducted on the MS COCO 2014 Karpathy test split under the fixed decoding configuration described in Section 4.2. The global visual conditioning checkpoint selected during Stage 1 served as the frozen G0 anchor for the controlled architecture experiments. B0 denotes the formal test evaluation of the global-only configuration, whereas B1 adds the proposed gated local residual conditioning module while preserving the same frozen global anchor. Table 5 compares the two configurations.

Table 5. Main quantitative results on the MS COCO Karpathy test split.

Model	BLEU-1/2/3/4	METEOR	ROUGE-L	CIDEr
B0 (Global Visual Conditioning)	0.676394 /0.496914 /0.356322 /0.253372	0.237369	0.497198	0.838206
B1 (Gated Local Residual)	0.681606 /0.504426 /0.363676 /0.260442			

As shown in Table 5, B1 outperformed B0 across all reported automatic metrics. CIDEr increased from 0.838206 to 0.869255, corresponding to an absolute gain of 0.031049. BLEU-4 increased from 0.253372 to 0.260442, while METEOR and ROUGE-L increased from 0.237369 and 0.497198 to 0.244324 and 0.504276, respectively. The BLEU-1, BLEU-2, and BLEU-3 scores also increased under the same decoding configuration. The observed improvement suggests that the proposed architecture benefits from combining complementary global semantics with local spatial evidence. While the global visual tokens provide stable image-level context, the gated local residual selectively introduces fine-grained information without substantially disturbing the pretrained visual-language alignment. This design enables additional visual evidence to be incorporated while preserving the robustness of the frozen backbones.

These results indicate that incorporating complementary local spatial information effectively enhances caption quality while preserving the frozen ResNet-50 image encoder and the frozen Phi-3 Mini language model. Because the pretrained ResNet-50 and Phi-3 Mini backbones were held fixed in both configurations, the observed difference is associated with the addition of the gated local residual conditioning module under the evaluated setting. The following sections analyze the contribution of the proposed local conditioning strategy in greater detail through architectural ablation, comparison with a parameter-efficient LoRA control, sequence-level optimization using SCST, qualitative LLM-based evaluation, and cross-domain generalization experiments.

5.2. Architectural Ablation Study

Before evaluating the local-conditioning variants, the global visual conditioning model was trained to establish a common anchor for the controlled architecture comparison. The checkpoint selected at epoch 3 achieved a proxy CIDEr score of 0.7795 and was subsequently fixed as the G0 global anchor. The formal MS COCO Karpathy test evaluation of the corresponding global-only configuration is denoted as B0.

The selected G0 checkpoint was subsequently evaluated as the formal B0 baseline. The G0 training run and the formal B0 evaluation therefore refer to the same architecture under different evaluation objectives. B0 relies exclusively on the global visual representation extracted by global average pooling and serves as the reference architecture for the following local-conditioning variants. The proxy score used for G0 selection and the formal B0 test score were obtained under different evaluation scopes and are therefore not directly compared as captioning results. Accordingly, all subsequent architectural comparisons are performed using the formal Karpathy test results. The selection of the frozen global anchor and the corresponding formal baseline is summarized in Table 6.

Table 6. Selection of the frozen global anchor (G0) and the corresponding formal baseline (B0).

Stage	Configuration	Evaluation	CIDEr
G0	Global anchor selection	Proxy CIDEr	0.7795
B0	Formal baseline	Karpathy test	0.8382

The architectural ablation focused on three configurations sharing the same frozen G0 anchor. B0 used only the global visual conditioning tokens. F0 added a local residual directly to the global tokens without gated modulation.

F0 preserves the frozen G0 anchor while introducing a local residual branch. The local residual is generated from the 7×7 spatial feature map through local query cross-attention and is directly added to the global visual prefix without any learnable modulation. Unlike B0, the global projector remains frozen and only the newly introduced local branch is

trainable. The global visual prefix serves as the query, while the local visual tokens are used as keys and values during local query cross-attention.

B1 applied the proposed gated residual update, in which a learnable scalar gate controlled the contribution of the local residual.

B1 extends F0 by introducing a learnable scalar gate that scales the local residual before addition. This enables the model to regulate the contribution of local spatial information while preserving the frozen G0 global anchor. Consequently, the only architectural difference between F0 and B1 is the introduction of the learnable scalar gate, while all remaining components remain identical. Table 7 reports the results obtained under the same fixed decoding configuration.

Table 7. Architectural ablation results on the MS COCO Karpathy test split under the fixed decoding configuration.

Model	Local Feature	Fusion Strategy	BLEU-4	METEOR	ROUGE-L	CIDEr
B0 (Global Visual Conditioning)	Absent	Global prefix only	0.253372	0.237369	0.497198	0.838206
F0 (Direct Local Residual)	Present	Direct local residual	0.260119	0.245827	0.505733	0.861087
B1 (Gated Local Residual)	Present	Gated local residual	0.260442	0.244324	0.504276	0.869255

Listings A1–A3 (Appendix A) show the corresponding block diagrams for B0, F0, and B1, respectively. F0 adds the local residual directly to the global visual prefix, whereas B1 additionally scales the same residual with a learnable sigmoid gate, $\text{sigmoid}(\alpha)$, before addition—the only architectural difference between the two configurations.

Relative to the global-only baseline (B0), both local-conditioning variants improved all four reported automatic metrics. F0 increased CIDEr from 0.838206 to 0.861087, an absolute gain of 0.022881, indicating that the spatial feature map provided information not contained in the pooled global representation alone. B1 further increased CIDEr to 0.869255, corresponding to gains of 0.031049 over B0 and 0.008168 over F0. These improvements demonstrate that incorporating local spatial information is beneficial even when the pretrained visual encoder and language model remain completely frozen. The largest improvement was consistently observed in CIDEr, which served as the primary optimization and checkpoint-selection metric throughout this study.

The results across the individual metrics were not uniform. B1 achieved the highest BLEU-4 and CIDEr scores, whereas F0 obtained slightly higher METEOR and ROUGE-L scores. Thus, the gate did not improve every metric simultaneously. Nevertheless, because CIDEr was defined in advance as the primary checkpoint-selection and captioning-performance criterion, B1 was retained as the main architectural variant.

The learned scalar gate of B1 was approximately 0.01947 at the selected checkpoint. This value confirms that the model maintained a relatively small scalar multiplier on the local residual. However, the gate value alone does not fully determine the effective magnitude of the local update because the scale of the residual representation also contributes to the final token modification. It is therefore interpreted as evidence of a gated residual formulation, rather than as a direct quantitative measure of the local branch's semantic contribution. Instead, the gate value should be interpreted jointly with the learned residual representation, since both determine the effective local correction applied to the global visual prefix. The learned gate remained small throughout training, indicating that

the local residual acted as a complementary correction rather than replacing the global visual representation.

These findings indicate that simply adding local spatial information is beneficial, but regulating its contribution further improves the balance between global semantic consistency and local visual detail. The gated residual therefore functions as a lightweight mechanism for preventing excessive local corrections while still exploiting complementary spatial evidence.

Overall, the ablation results show that preserving the frozen global anchor and introducing local information as a gated residual provides a more effective visual conditioning strategy than relying on the global representation alone under the evaluated setting. Among the two local-fusion variants, gated fusion achieved the highest CIDEr score and was therefore selected as the principal supervised architecture. These findings support the proposed design principle that preserving a stable global visual anchor while introducing lightweight gated local corrections enables complementary use of local spatial information without replacing the global visual representation.

5.3. Sensitivity Analysis and Configuration Selection

5.3.1. Effect of Visual Prefix Length

To examine the influence of the visual conditioning capacity, we performed a reduced-update screening study by varying the visual prefix length while keeping all remaining training settings unchanged. Because this experiment was intended only for model selection under a common computational budget, the comparison was conducted using proxy CIDEr rather than full validation evaluation.

The prefix-length screening and G0 anchor selection were conducted in separate runs. Table 8 reports three-epoch reduced-update screening results used only to select the prefix length, whereas Table 6 reports the best checkpoint from the full five-epoch Stage 1 training run with the fixed cyclic-caption schedule. Consequently, the two proxy CIDEr values are not directly comparable.

Table 8. Effect of visual prefix length under the reduced-update screening protocol.

Prefix Length	Epoch 1	Epoch 2	Epoch 3	Best Proxy CIDEr
8	0.600762	0.673065	0.720006	0.720006
10	0.624270	0.710732	0.746270	0.746270
12	0.614885	0.713431	0.734474	0.734474

As shown in Table 8, increasing the visual prefix length from 8 to 10 improved the best proxy CIDEr from 0.7200 to 0.7463. Extending the prefix length to 12 did not yield an additional improvement, resulting instead in a lower best proxy CIDEr of 0.7345. These results suggest that increasing the number of visual conditioning tokens beyond 10 did not improve visual-language alignment under the common screening budget.

Moreover, increasing the prefix length also enlarges the trainable projection layers. Therefore, prefix length 10 was adopted throughout the main experiments as a practical balance between captioning performance and model complexity.

5.3.2. Decoding Strategy Selection and Length Sensitivity

The canonical decoding configuration was selected solely based on the validation split. Beam search with four beams, a maximum generation length of 13 tokens, and a length penalty of 0.7 achieved the highest validation CIDEr and was therefore adopted for all subsequent quantitative evaluations. The validation results used to select the canonical decoding configuration are summarized in Table 9.

Table 9. Selection of the canonical decoding configuration based on the MS COCO validation split.

Selection	Beam	Max New Tokens	Length Penalty	Validation CIDEr	Remarks
Initial setting	3	18	0.8	0.763991	Initial evaluation
Improved	5	14	0.8	0.833659	Shorter generation length
Selected	4	13	0.7	0.846599	Highest validation CIDEr

The effect of increasing the maximum generation length on the MS COCO test split is reported in Table 10.

Table 10. Effect of longer decoding on the MS COCO test split.

Max New Tokens	Beam	Length Penalty	CIDEr	Purpose
13 (Canonical)	4	0.7	0.869255	Main quantitative evaluation
20	4	0.7	0.800745	Truncation sensitivity
24	4	0.7	0.792170	Truncation sensitivity

Although the canonical decoding configuration achieved the highest validation CIDEr, the subsequent LLM-based qualitative evaluation revealed that the short decoding length occasionally produced truncated captions. To examine this limitation, additional decoding experiments were conducted on the test split using longer generation lengths. Increasing the maximum generation length to 20 and 24 tokens improved sentence completeness in some cases but reduced CIDEr. Therefore, all quantitative comparisons reported in this paper use the canonical 13-token decoding configuration, whereas the longer decoding settings are presented only as a truncation sensitivity analysis.

5.4. Comparison with LoRA Control

To investigate whether the observed performance improvement originates from the proposed visual conditioning strategy rather than from decoder-side parameter-efficient adaptation, we compared the proposed model with a decoder-side QKV-LoRA control. All models shared the same frozen ResNet-50 encoder, frozen Phi-3 Mini decoder, identical training protocol, and the same G0 global anchor. The only difference was the location of adaptation.

As shown in Table 11, the decoder-side LoRA control produced only a modest improvement over the global-only baseline, increasing CIDEr from 0.838206 to 0.843889. In contrast, the proposed B1 model achieved a CIDEr score of 0.869255, consistently outperforming the LoRA control under the same experimental setting.

Table 11. Comparison with the decoder-side LoRA control on the MS COCO Karpathy test split.

Model	Adaptation Target	BLEU-4	METEOR	ROUGE-L	CIDEr
B0	Global	0.253372	0.237369	0.497198	0.838206
L0	visual projector				
(QKV-LoRA)	Phi-3 decoder	0.255651	0.244468	0.501148	0.843889
B1	(QKV-LoRA)				
(Proposed)	Gated local	0.260442	0.244324	0.504276	0.869255
	visual conditioning				

These findings suggest that, under the evaluated frozen-backbone setting, adapting the visual conditioning pathway is more effective than adapting the decoder through QKV-LoRA. This observation supports the proposed design choice of improving the visual interface between the frozen vision encoder and the frozen language model while

preserving the pretrained language representations. The computational implications of these adaptation strategies are analyzed in the following section.

5.5. Computational Analysis

Table 12 summarizes the computational characteristics measured under identical cached-feature inference settings. During the initial training stage, the global visual-prefix projector updated 67.14 M parameters. After the global projector was frozen, the proposed local residual branch introduced only 3.95 M additional trainable parameters during the subsequent training stage.

Table 12. Stage-specific computational characteristics under cached-feature inference.

Model	Params Updated	Module Size (MB)	Cond. GFLOPs	Full-Gen GFLOPs	Lat. p50 (ms)	Lat. p90 (ms)
B0	67.142 M	256.12	0.1342	984.2276	1174.78	1207.00
B1	3.951 M	15.07	0.3560	984.4494	1171.85	1182.67
B2	3.951 M	15.07	0.3560	984.4494	1171.27	1206.32

Parameters Updated During Training refer only to the parameters optimized at each training stage. The frozen global projector is reused during B1 and B2 training.

Although the local residual branch increased the conditioning computation from 0.1342 GFLOPs to 0.3560 GFLOPs, the total autoregressive generation cost changed only from 984.23 to 984.45 GFLOPs because computation remained dominated by beam-search decoding in the frozen Phi-3 Mini language model. The additional conditioning cost therefore accounted for only approximately 0.023% of the total generation computation.

Consistent with this observation, end-to-end inference latency remained essentially unchanged across all variants. Under batch-1 inference with four-beam decoding, the median latency was approximately 1.17 s per image on an NVIDIA A100-SXM4-80 GB. The measured latency includes visual conditioning and autoregressive caption generation but excludes image preprocessing and ResNet-50 feature extraction. FlashAttention was not enabled during the measurements.

SCST modified only the sequence-level training objective and did not introduce any additional inference-time parameters or computational overhead. Consequently, the computational characteristics of B2 remained effectively identical to those of B1.

The static memory footprint of each model variant is summarized in Table 13.

Table 13. Static memory footprint of each model variant.

Deployment Configuration	Static Loaded Memory
B0	2.305 GB
B1	2.320 GB
B2	2.320 GB

Static memory corresponds to loaded model weights and buffers only, excluding CUDA allocator reservation, KV cache, and other temporary tensors.

Overall, the proposed gated local residual branch improved CIDEr from 0.8382 to 0.8693 while introducing only a small additional conditioning cost. These results indicate that the proposed visual conditioning strategy achieves a favorable trade-off between captioning performance and computational efficiency under the evaluated frozen-backbone setting.

5.6. Effect of SCST

To examine whether the proposed architecture could benefit from sequence-level optimization, we conducted an additional Self-Critical Sequence Training (SCST) experiment using the selected B1 model as initialization. As described in Section 4.2, SCST was

treated as a post-hoc optimization strategy rather than as part of the proposed architecture. During SCST, only the gated local residual conditioning module was updated, while the frozen ResNet-50 encoder, frozen Phi-3 Mini decoder, and frozen G0 global projector remained unchanged.

As shown in Table 14, SCST produced consistent but modest improvements across all reported automatic metrics. CIDEr increased from 0.869255 to 0.870424, corresponding to an absolute gain of 0.001169. Similar small improvements were observed for BLEU-4, METEOR, and ROUGE-L.

Table 14. Effect of SCST on the proposed gated local residual model.

Model	BLEU-4	METEOR	ROUGE-L	CIDEr
B1 (XE-trained)	0.260442	0.244324	0.504276	0.869255
B2 (+SCST)	0.261062	0.244758	0.505061	0.870424

Although the improvement was small, the results indicate that the proposed visual conditioning architecture remains compatible with sequence-level reinforcement learning after supervised training. Because SCST modifies only the optimization objective without changing the inference architecture, the computational characteristics reported in Section 5.5 remain unchanged.

Overall, SCST provided a marginal additional performance gain without introducing any inference-time parameters or computational overhead. Since the primary contribution of this work is the proposed visual conditioning architecture, SCST is reported as an auxiliary optimization experiment rather than as a core component of the proposed method.

5.7. LLM-Based Qualitative Analysis

Automatic captioning metrics primarily quantify lexical or semantic agreement with human references and may not fully capture sentence completeness, fluency, or the qualitative characteristics of individual captions. To complement the automatic evaluation, we conducted an LLM-assisted qualitative assessment using GPT-5.4-mini under the reference-compatible protocol described in Section 4.3.

For each image, the evaluator received one generated caption and the corresponding five human reference captions. Evaluation was performed deterministically with a fixed prompt and a temperature of 0. The evaluator assessed semantic correctness, detail coverage, fluency and completeness, and reference compatibility. Because human reference captions are incomplete descriptions of an image, plausible details were not classified as errors solely because they were absent from the references. Unsupported content was assigned only when the generated caption contradicted or substantially exceeded the information reasonably supported by the reference set.

5.7.1. Paired Analysis of Decoding Length

The canonical Max-13 decoding configuration was selected to maximize validation CIDEr and was used for all main quantitative comparisons. However, preliminary inspection indicated that this relatively short decoding budget frequently produced incomplete captions. We therefore first compared Max-13 with an extended Max-20 configuration on a subset of 500 images from the MS COCO Karpathy test split. The same trained B1 model and the same images were used for both configurations, allowing the effect of decoding length to be examined through paired comparison.

As shown in Table 15, increasing the maximum decoding length substantially improved sentence completeness and overall qualitative quality. Severe truncation decreased from 60.8% under Max-13 to 1.4% under Max-20, while the proportion of captions rated

Good or Excellent increased from 19.4% to 59.2%. Fluency and completeness exhibited the largest mean-score improvement, increasing from 2.638 to 4.922.

Table 15. Reference-compatible paired evaluation of decoding length on 500 MS COCO Karpathy test images.

Evaluation Dimension	Metric	Max-13	Max-20	Difference
Overall quality	Good + Excellent	19.4% (97)	59.2% (296)	+39.8%p
	Poor	58.8% (294)	23.8% (119)	−35.0%p
Truncation	Severe truncation	60.8% (304)	1.4% (7)	−59.4%p
Mean scores	Semantic correctness	3.352	3.588	+0.236
	Detail coverage	2.434	2.926	+0.492
	Fluency and completeness	2.638	4.922	+2.284
Pairwise comparison	Max-13 preferred	6.6% (33)	—	—
	Max-20 preferred	—	73.2% (366)	—
	Tie	20.2% (101)	—	—
Recovery analysis	Longer decoding fixed the identified issue	—	61.8% (309)	—
Reference compatibility	Supported or compatible	57.6% (288)	52.8% (264)	−4.8%p
	Major unsupported	24.6% (123)	23.6% (118)	−1.0%p

Pairwise evaluation preferred Max-20 in 73.2% of the image pairs, whereas Max-13 was preferred in only 6.6%. In addition, longer decoding resolved the identified issue without introducing a new major error in 61.8% of the evaluated cases. These results provide empirical support for using Max-20 in the subsequent qualitative analysis.

However, the effect of longer decoding was not uniform across all evaluation dimensions. The proportion of supported or compatible captions decreased slightly from 57.6% to 52.8%, while the proportion of major unsupported captions remained similar. Thus, Max-20 primarily improved fluency, completeness, and detail coverage rather than fully resolving semantic-grounding errors.

Figure 4 visually summarizes the principal differences between the two decoding settings, including overall quality, truncation severity, and pairwise preference.

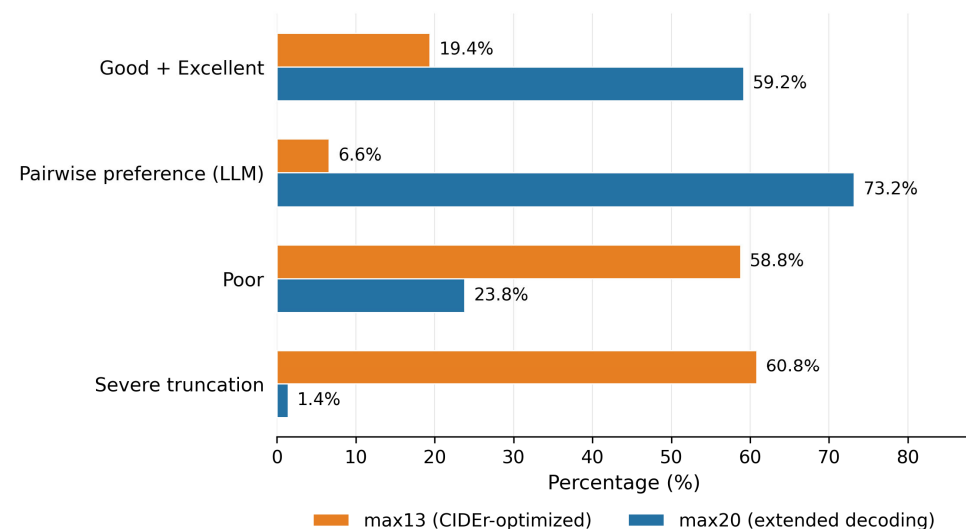


Figure 4. Qualitative outcomes of Max-13 and Max-20 decoding on 500 paired MS COCO test images.

Although Max-20 produced substantially more complete and readable captions, its CIDEr score was lower than that of Max-13, as reported in Section 5.3.2. Therefore, Max-13 was retained as the canonical CIDEr-optimized configuration for standardized quantitative evaluation, whereas Max-20 was adopted for the full LLM-assisted qualitative analysis.

5.7.2. Full-Scale Qualitative Evaluation with Max-20

Based on the paired decoding analysis, the full-scale qualitative evaluation was conducted using the Max-20 configuration over all 5000 images in the MS COCO Karpathy test split. Table 16 summarizes the overall quality, truncation, reference-compatibility, and mean-score results.

Table 16. Full-scale LLM-assisted qualitative evaluation of B1 with Max-20 on the 5000-image MS COCO Karpathy test split.

Evaluation Dimension	Category or Metric	Count	Percentage or Score
Overall quality distribution	Excellent	567	11.34%
	Good	1400	28.00%
	Adequate	1680	33.60%
	Poor	1353	27.06%
Aggregated quality indicator	Good + Excellent	1967	39.34%
	Adequate or better	3647	72.94%
Truncation severity	None	4908	98.16%
	Mild	5	0.10%
	Severe	87	1.74%
Reference compatibility	Supported	822	16.44%
	Compatible extra detail	1225	24.50%
	Plausible unverified	37	0.74%
	Weakly unsupported	1639	32.78%
	Major unsupported	1007	20.14%
	Contradiction	270	5.40%
Aggregated compatibility indicator	Supported or compatible	2084	41.68%
Mean qualitative scores	Semantic correctness	—	3.284/5
	Detail coverage	—	2.829/5
	Fluency and completeness	—	4.927/5

Each evaluation dimension represents a separate distribution over the same 5000 captions and sums independently to 100%. “Good + Excellent,” “Adequate or better,” and “Supported or compatible” are aggregated indicators and should not be added to the corresponding individual categories.

Under the Max-20 configuration, 39.34% of captions were rated Good or Excellent, and 72.94% were rated Adequate or better. Severe truncation occurred in only 1.74% of the outputs, and the mean fluency-and-completeness score reached 4.927. These findings confirm that the longer decoding configuration largely resolved the sentence-incompleteness problem observed under Max-13.

Nevertheless, unsupported semantic content remained the principal qualitative limitation. Only 41.68% of the captions were classified as supported or compatible, whereas 32.78% were weakly unsupported, 20.14% were majorly unsupported, and 5.40% contradicted the references. The most frequent primary issues were unsupported objects and missing details, indicating that increasing decoding length improves surface-level completeness but does not independently guarantee stronger visual grounding.

5.7.3. Representative Reference Selection and Qualitative Examples

To illustrate the aggregate findings, four representative cases were selected: stable generation, truncation recovery, unsupported object generation, and remaining truncation. Each panel displays the source image, one representative human reference caption, the Max-13 output, the Max-20 output, and a concise assessment.

Because five human references are available for each MS COCO image, the displayed reference was selected independently of both generated outputs. The deterministic selection protocol is summarized in Table 17.

Table 17. Representative reference-caption selection protocol for qualitative visualization.

Component	Description
Reference candidates	Five human captions associated with each selected MS COCO Karpathy image
Annotation source	Official Karpathy annotation (dataset_coco.json) or the corresponding MS COCO annotation
Representative-reference method	Human-reference medoid
Medoid criterion	Human caption with the highest average pairwise similarity to the remaining reference captions
Output independence	The representative reference was selected without using either the Max-13 or Max-20 output
Determinism	The same candidate reference set always produces the same representative caption
Purpose	To prevent manual reference selection that might favor either decoding configuration

Representative examples were selected according to the reference-selection protocol summarized in Table 17. For each image, the representative reference caption was determined independently of the model outputs using the human-reference medoid among the five COCO annotations.

Figure 5 presents the corresponding representative examples.

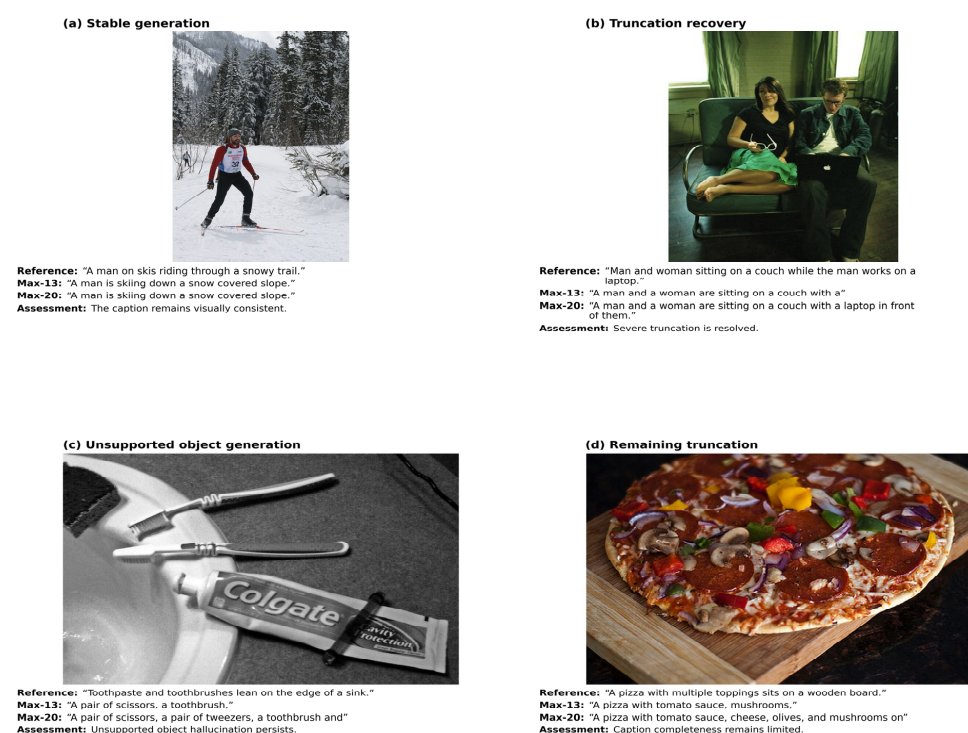


Figure 5. Representative qualitative examples under Max-13 and Max-20 decoding.

Figure 5 demonstrates that the effect of extended decoding depends on the underlying error type. In panel (a), both settings generate the same complete and visually consistent caption. In panel (b), Max-20 resolves the severe truncation produced by Max-13 and restores the reference-supported laptop description. In contrast, panel (c) shows that extending the decoding length does not correct unsupported object generation and can introduce additional unsupported content. Panel (d) further illustrates that a longer decoding budget does not guarantee complete sentence generation in every case.

The metadata and qualitative labels associated with the four representative cases are summarized in Table 18.

Table 18. Representative example summary.

Panel	Case Type	Main Observation
(a)	Stable generation	Both decoding settings generate an accurate caption.
(b)	Truncation recovery	Max-20 resolves severe truncation.
(c)	Unsupported object	Unsupported object persists after longer decoding.
(d)	Remaining truncation	Caption remains incomplete.

Overall, the qualitative analysis reveals a clear distinction between automatic benchmark optimization and human-oriented caption quality. The CIDEr-optimized Max-13 configuration remains appropriate for standardized quantitative comparison, whereas Max-20 substantially improves sentence completeness, fluency, and qualitative preference. However, the limited change in reference compatibility and the representative failure cases demonstrate that additional decoding capacity alone does not resolve unsupported semantic content. Decoding length primarily affects the completeness and readability of generated captions, while semantic grounding remains dependent on the quality of the learned visual-to-language conditioning.

5.8. Cross-Domain Generalization

To examine the generalization ability of the proposed visual conditioning model beyond the training distribution, we evaluated the selected B1 model directly on the Flickr30k test split without any additional fine-tuning or parameter updates. This experiment therefore represents a strict COCO-to-Flickr zero-shot transfer setting, formalized as Algorithm A5 (Appendix B).

As shown in Table 19, the proposed model exhibited a substantial performance decrease under COCO-to-Flickr zero-shot transfer. CIDEr decreased from 0.8693 on the MS COCO Karpathy test split to 0.2931 on Flickr30k, accompanied by consistent reductions in BLEU-4, METEOR, and ROUGE-L.

Table 19. Cross-domain zero-shot evaluation from MS COCO to Flickr30k.

Dataset	Evaluation	BLEU-4	METEOR	ROUGE-L	CIDEr
MS COCO Karpathy test	In-domain	0.260442	0.244324	0.504276	0.869255
Flickr30k test	Zero-shot	0.149339	0.162456	0.399274	0.293084

These results indicate that the visual-to-language alignment learned from MS COCO does not fully transfer to the Flickr30k caption distribution. The observed degradation suggests that the learned global visual conditioning remains sensitive to dataset-specific captioning styles and linguistic distributions.

The observed performance degradation motivates future investigation of lightweight dataset adaptation strategies, such as adapting the global visual conditioning module while

preserving the frozen vision encoder and language model. Such experiments are left for future work and are beyond the scope of the present study.

5.9. Contextual Comparison with Representative Image Captioning Models

To provide broader benchmark context, Table 20 compares the proposed B1 model with representative image captioning systems evaluated on the MS COCO Karpathy test split. Because a single parameter count does not fully characterize model capacity or training scope, the table separately reports total model scale, paired multimodal pretraining before COCO adaptation, and COCO-stage trainable parameters or update scope. Only single-model results reported without SCST or CIDEr-based reinforcement learning are included.

Table 20. Contextual comparison with representative image captioning models on the MS COCO Karpathy test split.

Method	Total Model Scale	Paired Multimodal Pretraining Before COCO	Other Pretraining Data	COCO-Stage Trainable Parameters/ Update Scope	BLEU-4	METEOR	CIDEr
ClipCap [22]	N/R	≈400 M image–text pairs (via CLIP [30])	—	43 M	33.5	27.5	113.1
I-Tuning [31]	N/R		—	14 M	34.8	28.3	116.7
SmallCap [32]	218 M		Retrieval Datastore (COCO captions)	7 M	37.0	27.9	119.7
OSCAR -Large [33]	N/R	6.5 M image–text pairs	Region Features / object tags	345 M	37.4	30.7	127.8
SimVLM -Huge [34]	632 M	≈1.8 B image–text pairs	C4 text	≈1.4 B	40.6	33.7	143.3
CoCa [35]	2.1 B		≈3 B labeled images	N/R	40.9	33.9	143.6
BEiT-3 [36]	1.9 B	21 M image–text pairs	14 M images + 160 GB text	Broad downstream updating	44.1	32.4	147.6
Flamingo -80 B [15]	80 B	2.112 B image–text + 27 M video–text pairs	M3W: 43 M Interleaved webpages	10.6 B; LM And vision encoder frozen	–	–	138.1
BLIP-2 ViT-g +FlanT5-XL [13]	4.1 B	129 M Pretraining images	Synthetic captions also used	1.1 B; Q-Former + image encoder updated, LLM frozen	42.4	–	144.5
Proposed (B1)	≈3.9 B	No dedicated paired multimodal pretraining Before COCO adaptation	ImageNet vision +text only LM pretraining	71.09 M; both backbones frozen	26.0	24.4	86.9

All results are single-model results on the MS COCO Karpathy test split without SCST or CIDEr-based reinforcement learning. Total model scale and COCO-stage trainable parameters follow the values reported in the cited publications and may reflect different

accounting conventions. N/R denotes a value not reliably reported, and “–” denotes an unreported metric.

The proposed method obtains lower absolute benchmark scores than the comparison systems. However, the resource and training conditions differ substantially. Most of its approximately 3.9 B total parameters reside in the frozen Phi-3 Mini decoder, while 71.09 M parameters are learned across the two COCO adaptation stages, and no dedicated paired multimodal pretraining is used beforehand. The comparison systems range from CLIP-based lightweight adaptation to large-scale multimodal pretraining and broader downstream updating. Table 20 is therefore intended to provide benchmark and resource context rather than a controlled ranking; the implications of these differences are discussed in Section 5.10.

5.10. Discussion

The experimental results support the proposed design principle of preserving a stable global visual representation while introducing lightweight local residual corrections. Rather than replacing the global visual conditioning pathway, the proposed architecture treats local spatial information as a complementary residual signal. The architectural ablation demonstrated that this residual formulation consistently outperformed the global-only baseline, while the gated variant further improved CIDEr over direct residual addition under the same training and decoding conditions.

The comparison with the decoder-side LoRA control further suggests that, under the evaluated frozen-backbone setting, allocating lightweight adaptation capacity to the visual-conditioning pathway may be more effective than applying the tested QKV-LoRA configuration within the language model. Because all models shared the same frozen ResNet-50 encoder, frozen Phi-3 Mini decoder, and optimization protocol, the observed differences are associated with the location and form of the added adaptation modules under this experimental setting.

Table 20 places these controlled findings in the broader context of representative captioning systems. Total model scale, prior multimodal pretraining, and downstream trainable parameters represent distinct forms of capacity. Although the proposed system contains approximately 3.9 B parameters in total and is therefore larger than several comparison models in raw parameter count, most of this capacity resides in a frozen text-only language model without prior vision–language alignment. A total of 71.09 M parameters are learned across the two COCO adaptation stages. By contrast, the comparison systems inherit or perform paired multimodal pretraining ranging from millions to billions of examples and generally employ broader downstream adaptation. Accordingly, total parameter count alone does not provide a sufficient basis for interpreting the observed performance differences.

The reported parameter values also require careful interpretation. The trainable-parameter counts for OSCAR, SimVLM, Flamingo, and BLIP-2 are reported together in the BLIP-2 captioning comparison table [13], which provides a common literature anchor but does not guarantee an identical accounting convention across the original studies. For example, the original SimVLM publication reports a 632 M-parameter Huge architecture, whereas the BLIP-2 comparison table reports approximately 1.4 B trainable parameters for the corresponding captioning result. Similarly, the BLIP-2 comparison reports 10.6 B trainable parameters for Flamingo, whereas Flamingo’s architectural breakdown implies approximately 10.2 B parameters in its gated cross-attention/dense layers and Perceiver Resampler. These values are retained transparently rather than reconciled through unsupported component-wise estimation. SmallCap’s total model scale of 218 M parameters is explicitly reported in its original publication [32], whereas corresponding complete-model

figures for ClipCap and I-Tuning are not explicitly reported and are therefore marked as N/R rather than estimated. For BEiT-3, the COCO captioning configuration uses a layer-wise learning-rate decay of 1.0 [36], which is consistent with broad downstream updating rather than adaptation restricted to a small task-specific module.

The comparison also distinguishes methods that share broadly related frozen-backbone principles. ClipCap, I-Tuning, and SmallCap use frozen CLIP visual encoders that inherit pretraining on approximately 400 million image–text pairs [30]. BLIP-2 and Flamingo retain frozen language models but adapt substantially larger visual-conditioning components, with BLIP-2 updating its Q-Former and image encoder during COCO captioning and Flamingo training gated cross-attention layers and a Perceiver Resampler. The proposed method instead keeps both backbones frozen, performs no dedicated multimodal pretraining before COCO adaptation, and learns only its two-stage visual-conditioning interface. These settings are not directly controlled; consequently, Table 20 should be interpreted as broad benchmark context rather than as a parameter-efficiency ranking. The primary controlled evidence for the proposed architecture therefore remains the B0-to-B1 comparison under identical backbones, data, and optimization conditions, where the gated local residual module improves CIDEr from 0.8382 to 0.8693 with 3.95 M additional stage-specific parameters.

From a computational perspective, the proposed local residual branch introduced only a small additional computational cost relative to the frozen backbone. Although the conditioning computation increased slightly, the total generation cost remained almost unchanged because autoregressive decoding dominated the overall inference workload. Similarly, median inference latency remained approximately constant across all variants. These observations indicate that the added local-conditioning branch improves captioning performance with limited change in overall inference cost under the evaluated configuration.

The qualitative evaluation provides additional insight into the behavior of the proposed model beyond conventional automatic metrics. While the canonical decoding configuration was selected according to validation CIDEr, the LLM-assisted qualitative analysis showed that longer decoding substantially reduced severe truncation and improved caption fluency and completeness. At the same time, the reference-compatible evaluation indicated that unsupported or weakly supported semantic content remained in a noticeable portion of the generated captions. This suggests that increasing decoding length primarily improves linguistic completeness rather than fully resolving semantic grounding errors.

The cross-domain evaluation on Flickr30k further highlights the current limitations of the proposed approach. Although the model could be evaluated without target-domain fine-tuning, performance decreased considerably under the COCO-to-Flickr transfer setting. This observation indicates that the learned visual-to-language alignment remains dependent on dataset-specific caption distributions and annotation styles. Consequently, stronger domain adaptation or domain-invariant visual conditioning may be required for robust cross-dataset caption generation.

Overall, the results indicate that gated local residual conditioning is a viable approach for incorporating local visual evidence under frozen vision and language backbones. At the same time, the remaining challenges in semantic grounding and cross-domain generalization motivate further investigation.

6. Conclusions

This study investigated whether local spatial information can improve image captioning when both the visual encoder and language model remain frozen. To address this question, we introduced a global-to-local visual conditioning framework in which a stable

global representation is complemented by a gated local residual branch without modifying either pretrained backbone.

The principal finding is that, among the evaluated conditioning strategies, local visual evidence was more effective when introduced as a regulated residual correction than through direct residual addition. Under the evaluated configuration, this design improved caption generation over global-only conditioning and direct local residual fusion. The comparison with the tested QKV-LoRA control further suggests that, within this frozen-backbone setting, allocating limited adaptation capacity to the visual-conditioning interface may be more effective than applying the evaluated decoder-side adaptation. This interpretation is restricted to the architectures and adaptation settings examined in this study.

The experiments also reveal an important distinction between linguistic completeness and visual grounding. Extending the decoding budget improved fluency and reduced incomplete captions, but it did not consistently eliminate unsupported semantic content. Likewise, the decline observed under cross-dataset evaluation indicates that the learned visual-to-language interface remains dependent on the captioning distribution used for adaptation. Together, these findings show that lightweight visual conditioning can improve the use of frozen pretrained components, while reliable semantic grounding and domain transfer remain unresolved challenges.

The present conclusions are limited to the frozen ResNet-50 and Phi-3 Mini configuration examined in this study. Future work should determine whether the same global-to-local conditioning principle remains effective across different visual encoders, language models, and captioning datasets and should investigate conditioning mechanisms that provide stronger grounding and greater robustness to domain shift.

Author Contributions: Conceptualization, S.L.; Methodology, S.L.; Software, S.L.; Validation, S.L.; Formal analysis, S.L.; Investigation, S.L.; Writing—original draft preparation, S.L.; Writing—review and editing, D.H.; Supervision, D.H. All authors have read and agreed to the published version of the manuscript.

Funding: The publication of this article (Article Processing Charge) was supported by a grant (RS-2025-02220682) from the Technology Development Program for Strengthening Resilience Against Urban Wind Disasters, funded by the Ministry of the Interior and Safety (MOIS, Republic of Korea).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The MS COCO 2014 Karpathy split used in this study is publicly available from the MS COCO website. The Flickr30k dataset used for the supplementary cross-dataset evaluation is publicly available from the Flickr30k project website under its respective usage terms. The zero-shot evaluation was conducted on the standard Flickr30k test split containing 1000 images. The Phi-3 Mini 3.8 B base model is publicly available from Microsoft's model repository under the MIT License. The proposed architecture, training procedure, experimental settings, and evaluation protocols are described in sufficient detail throughout this paper to facilitate independent reproduction of the reported experiments.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Architectural Summary Listings

Listing A1. Architecture of the B0 global visual-prefix baseline (derived from the selected G0 anchor).

```

Input image
  ↓
Frozen ResNet-50
  ↓
Global average pooling
  ↓
Global visual feature [2048]
  ↓
Trainable global visual-prefix projector
  ↓
Global visual prefix [10 × 3072]
  ↓
Frozen Phi-3 Mini
  ↓
Caption

```

Listing A2. Architecture of F0 (direct local residual conditioning). The local residual is added directly to the global visual prefix without gated modulation.

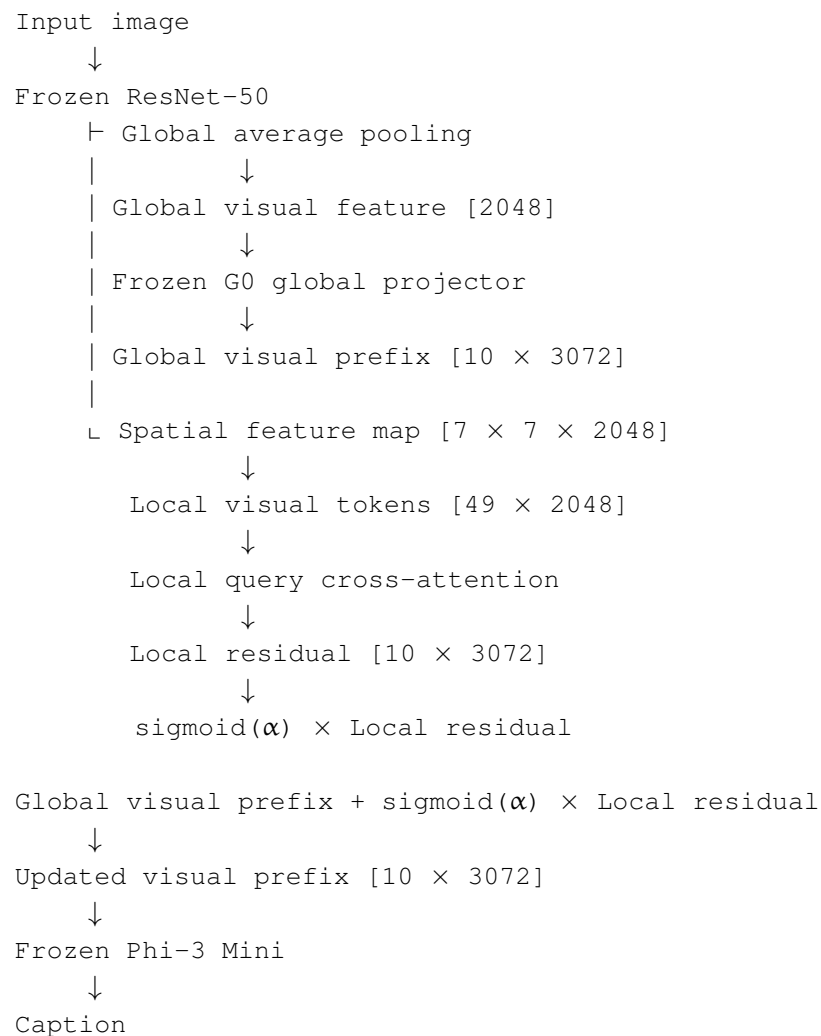
```

Input image
  ↓
Frozen ResNet-50
  ├ Global average pooling
  │   ↓
  │   Global visual feature [2048]
  │   ↓
  │   Frozen G0 global projector
  │   ↓
  │   Global visual prefix [10 × 3072]
  │
  └ Spatial feature map [7 × 7 × 2048]
      ↓
      Local visual tokens [49 × 2048]
      ↓
      Local query cross-attention
      ↓
      Local residual [10 × 3072]

Global visual prefix + Local residual
  ↓
Updated visual prefix [10 × 3072]
  ↓
Frozen Phi-3 Mini
  ↓
Caption

```

Listing A3. Architecture of B1 (gated local residual conditioning, proposed). Identical to F0 except that the local residual is scaled by a learnable sigmoid gate, $\text{sigmoid}(\alpha)$, before addition—the only architectural difference between F0 and B1.



Appendix B. Pseudocode for Training and Inference

Algorithm A1. Global Visual-Prefix Anchor Training

Input:

```

Training image-caption pairs D_COCO
Frozen visual encoder E_v = ResNet-50
Frozen language model L = Phi-3 Mini
Trainable global visual-prefix projector P_g
Number of visual-prefix tokens L_p = 10
Fixed textual prompt
  
```

Output:

```

Selected G0 anchor checkpoint  $\theta_{G0}$ 
  
```

Procedure:

1. Initialize the trainable global projector P_g .
2. Freeze all parameters of ResNet-50 and Phi-3 Mini.
3. For each training epoch e :

Algorithm A1. Cont.

- 3.1. For each image-caption pair (I, y) in D_{COCO} :
 - a. Extract the visual feature:

$$\mathbf{h} = E_v(I)$$
 - b. Apply global average pooling:

$$\mathbf{g} = \text{GAP}(\mathbf{h}) \in \mathbb{R}^{2048}$$
 - c. Project the global feature into visual-prefix tokens:

$$\mathbf{V}_g = P_g(\mathbf{g}) \in \mathbb{R}^{(L_p \times d_{\text{LM}})}$$
 - d. Prepend $\mathbf{V}_g$ to the embeddings of the fixed textual prompt.
 - e. Append the target-caption embeddings and compute the teacher-forcing captioning loss:

$$L_{\text{CE}} = \text{CrossEntropy}(L(\mathbf{V}_g, \text{prompt}, y_{<t}), y_t)$$
 - f. Update only the parameters of P_g .
- 3.2. Evaluate the checkpoint using proxy CIDEr.
4. Select the checkpoint with the highest proxy CIDEr as the G_0 anchor.

Return:

 θ_{G_0} **Algorithm A2. Direct Local Residual Conditioning**

Input:

Image-caption pairs D_{COCO}
 Frozen G_0 global anchor θ_{G_0}
 Frozen ResNet-50 visual encoder E_v
 Frozen Phi-3 Mini language model L
 Trainable local feature projector f_{local}
 Trainable learnable query tokens $\mathbf{Q}$
 Trainable local residual projection f_{proj}
 Spatial feature map size $7 \times 7 \times 2048$
 Visual-prefix length $L_p = 10$
 Fixed textual prompt

Output:

F0 DirectLocal checkpoint

Procedure:

1. Load the selected G_0 global-prefix anchor.
2. Freeze ResNet-50, Phi-3 Mini, and the G_0 global projector.
3. Initialize f_{local} , $\mathbf{Q}$, and f_{proj} .
4. For each image-caption pair (I, y) :
 - 4.1. Extract the ResNet-50 visual features.

Algorithm A2. Cont.

- 4.2. Obtain the global feature by global average pooling:
 $\mathbf{g} \in \mathbb{R}^{2048}$
- 4.3. Generate the frozen global visual prefix:
 $\mathbf{V_g} \in \mathbb{R}^{(L_p \times d_{LM})}$
- 4.4. Extract the spatial feature map:
 $\mathbf{S} \in \mathbb{R}^{(7 \times 7 \times 2048)}$
- 4.5. Flatten and project the spatial feature map:
 $\mathbf{T_l} = f_{\text{local}}(\mathbf{S}) \in \mathbb{R}^{(49 \times 512)}$
- 4.6. Apply local query cross-attention:
 $\mathbf{A} = \text{CrossAttn}(\mathbf{Q}, \mathbf{T_l}, \mathbf{T_l})$
- 4.7. Project the attention output into the language-model embedding space:
 $\mathbf{R} = f_{\text{proj}}(\mathbf{A}) \in \mathbb{R}^{(L_p \times d_{LM})}$
- 4.8. Directly add the local residual to the global prefix:
 $\mathbf{V} = \mathbf{V_g} + \mathbf{R}$
- 4.9. Condition frozen Phi-3 Mini on $\mathbf{V}$ and the fixed textual prompt, and compute the cross-entropy loss.
- 4.10. Update only f_{local} , $\mathbf{Q}$, and f_{proj} .

Return:

F0 checkpoint

Algorithm A3. Gated Local Residual Conditioning

Input:

Image-caption pairs D_{COCO}
 Frozen G0 global anchor θ_{G0}
 Frozen ResNet-50 visual encoder E_v
 Frozen Phi-3 Mini language model L
 Trainable local feature projector f_{local}
 Trainable learnable query tokens $\mathbf{Q}$
 Trainable local residual projection f_{proj}
 Trainable scalar gate α
 Fixed textual prompt

Output:

B1 GatedLocal checkpoint

Procedure:

1. Load the selected G0 global-prefix anchor.
2. Freeze ResNet-50, Phi-3 Mini, and the G0 global projector.
3. Initialize f_{local} and $\mathbf{Q}$.
4. Initialize the weights and biases of f_{proj} to zero.
5. Initialize the scalar gate logit α to -4.0 .

Algorithm A3. Cont.

6. For each image-caption pair (I, y) :
 - 6.1. Extract the global image feature:

$$\mathbf{g} = \text{GAP}(\mathbf{E}_v(I)) \in \mathbb{R}^{2048}$$
 - 6.2. Generate the frozen global visual prefix:

$$\mathbf{V}_g \in \mathbb{R}^{(L_p \times d_{LM})}$$
 - 6.3. Extract the spatial visual feature map:

$$\mathbf{S} \in \mathbb{R}^{(7 \times 7 \times 2048)}$$
 - 6.4. Flatten and project the spatial feature map:

$$\mathbf{T}_l = f_{\text{local}}(\mathbf{S}) \in \mathbb{R}^{(49 \times 512)}$$
 - 6.5. Apply local query cross-attention:

$$\mathbf{A} = \text{CrossAttn}(\mathbf{Q}, \mathbf{T}_l, \mathbf{T}_l)$$
 - 6.6. Compute the local residual:

$$\mathbf{R} = f_{\text{proj}}(\mathbf{A}) \in \mathbb{R}^{(L_p \times d_{LM})}$$
 - 6.7. Apply the learnable gate:

$$\mathbf{R}_{\text{gated}} = \text{sigmoid}(\alpha) \times \mathbf{R}$$
 - 6.8. Construct the updated visual prefix:

$$\mathbf{V} = \mathbf{V}_g + \mathbf{R}_{\text{gated}}$$
 - 6.9. Condition frozen Phi-3 Mini on $\mathbf{V}$ and the fixed textual prompt, and compute the captioning loss.
 - 6.10. Update only f_{local} , $\mathbf{Q}$, f_{proj} , and α .

Return:

B1 checkpoint

Algorithm A4. Caption Generation and Inference

Input:

Test image I
 Frozen ResNet-50 visual encoder $\mathbf{E}_v$
 Frozen Phi-3 Mini language model L
 Trained visual-conditioning modules
 Fixed textual prompt
 Fixed decoding configuration

Output:

Generated caption c

Procedure:

1. Extract visual features from I using frozen ResNet-50.
2. Compute the global visual prefix $\mathbf{V}_g$.
3. Construct the visual-conditioning tokens according to the selected configuration:

Algorithm A4. Cont.

```

    If using B0:
         $\mathbf{V} = \mathbf{V\_g}$ 

    Else if using F0:
        a. Compute the local residual  $\mathbf{R}$  from the projected
           spatial
           features using learnable query cross-attention.
        b. Set:
             $\mathbf{V} = \mathbf{V\_g} + \mathbf{R}$ 

    Else if using B1:
        a. Compute the local residual  $\mathbf{R}$  from the projected
           spatial
           features using learnable query cross-attention.
        b. Set:
             $\mathbf{V} = \mathbf{V\_g} + \text{sigmoid}(\alpha) \times \mathbf{R}$ 

4. Provide the fixed textual prompt:
   "Describe the image in one concise sentence."

5. Generate a caption from frozen Phi-3 Mini conditioned on  $\mathbf{V}$  and
   the fixed textual prompt.

6. Apply the fixed decoding configuration:
   beam search,
   length penalty,
   minimum and maximum numbers of newly generated tokens,
   and no-repeat n-gram constraint.

7. Return the generated caption.

Return:
c

```

Algorithm A5. Zero-Shot Flickr30k Transfer Evaluation**Input:**

```

COCO-trained B1 checkpoint
Flickr30k test images
Flickr30k reference captions
Fixed textual prompt
Fixed decoding configuration

```

Output:

```

Zero-shot Flickr30k evaluation scores

```

Procedure:

1. Load the COCO-trained B1 GatedLocal checkpoint.
 2. Do not update any model parameters.
 3. Do not fine-tune the model on Flickr30k.
 4. For each Flickr30k test image:
 - 4.1. Extract visual features using frozen ResNet-50.
 - 4.2. Construct the gated local visual-conditioning tokens.
-

Algorithm A5. Cont.

4.3. Generate a caption using frozen Phi-3 Mini with the same fixed textual prompt and decoding configuration.

5. Compare the generated captions with the Flickr30k reference captions.

6. Report BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L, and CIDEr.

Return:

Zero-shot transfer metrics

Appendix C. Hyperparameter and Reproducibility Details

All model, training, decoding, and evaluation hyperparameters are reported in the main text. Specifically, model architecture details are summarized in Section 3, training settings in Section 4.2 (Table 2), and evaluation settings and decoding in Tables 3 and 9, respectively. No additional hyperparameters beyond those reported in the main text were used.

Appendix D. LLM-Based Qualitative Evaluation Prompt**System Prompt**

You are an expert evaluator for image captioning outputs.

You will receive:

- five human reference captions for one COCO image
- one machine-generated caption for the same image

Important limitations:

You do NOT see the image.

The reference captions are incomplete natural-language descriptions, not exhaustive image annotations.

Therefore, absence from the references is NOT automatically an error.

Evaluate the generated caption in a reference-compatible way:

- Penalize content only when it is contradicted by the references, changes the main object/action/scene, or is very unlikely given all references.
- Do NOT mark ordinary compatible visual details as Major_Unsupported merely because they are absent from the references.
- Use Major_Unsupported only for major object/action/scene errors.
- Use Compatible_Extra_Detail or Plausible_Unverified for details that are not in references but do not conflict with them.

Definitions:

1. Truncation severity

- None: caption is syntactically complete and interpretable.
- Mild: caption is slightly incomplete or awkward, but the main meaning is still usable.
- Severe: caption clearly ends mid-phrase, for example after "with a", "and a", "of the", "in a", "holding a", or otherwise cannot be read as a complete caption.

2. Reference compatibility status

- Supported: the main caption content is directly supported by one or more references.
- Compatible_Extra_Detail: the caption includes details not explicitly mentioned in references, but they are ordinary visual details and do not conflict with the references.
- Plausible_Unverified: the caption includes details that cannot be verified from the references but are plausible for the scene.
- Weakly_Unsupported: the caption includes a questionable detail, but the main object/action/scene remains compatible.
- Major_Unsupported: the caption introduces a major object, action, or scene that is not supported and is unlikely given the references.
- Contradiction: the caption clearly conflicts with the references.

3. Unsupported content type

Use this only when the status is Weakly_Unsupported, Major_Unsupported, or Contradiction.
Otherwise use None.

4. Primary issue

Choose the single most important issue:

- None
- Truncation
- Missing_Detail
- Weak_Unsupported_Detail
- Unsupported_Object
- Unsupported_Attribute
- Unsupported_Action
- Unsupported_Scene_Detail
- Contradiction
- Generic
- Grammar
- Multiple_Issues

5. Final quality

- Excellent: complete, fluent, and the main content is supported or compatible with references.
- Good: mostly correct and complete, with only minor compatible or plausible extra details.
- Adequate: usable but generic, mildly incomplete, or contains weakly unsupported detail.
- Poor: severe truncation, contradiction, wrong main object/action/scene, or major unsupported content.

Scoring:

Give integer scores from 1 to 5.

- semantic_correctness: correctness of main objects/actions/scene against the references
- detail_coverage: useful amount of caption content, without requiring every reference detail
- fluency_completeness: grammatical fluency and sentence completeness

Important scoring guidance:

- A caption can be Good even if it has Compatible_Extra_Detail.
- A caption can be Adequate or Good even if it is somewhat generic.
- A caption should not be Poor only because it is generic.
- A caption should not be Poor only because it contains a detail absent from references.
- Severe truncation should strongly reduce fluency_completeness and final_quality.
- Major_Unsupported should be reserved for major errors, not minor extra details.
- If the caption captures the main object/action/scene correctly, prefer Supported, Compatible_Extra_Detail, or Plausible_Unverified over Major_Unsupported.

Return ONLY valid JSON.

Do not include markdown.

User Prompt Template

Image ID: {image_id}

File name: {file_name}

Reference captions:

1. {ref_1}
2. {ref_2}
3. {ref_3}
4. {ref_4}
5. {ref_5}

Generated caption:

{predicted_caption}

Evaluate the generated caption.

Return JSON with this exact schema:

```
{
  "image_id": {image_id},
  "truncation_severity": "None | Mild | Severe",
  "reference_compatibility_status": "Supported | Compatible_Extra_Detail
  | Plausible_Unverified | Weakly_Unsupported | Major_Unsupported |
  Contradiction",
  "unsupported_content_type": "None | Object | Attribute | Action |
  Scene_Detail | Spatial | Multiple",
  "primary_issue": "None | Truncation | Missing_Detail |
  Weak_Unsupported_Detail | Unsupported_Object | Unsupported_Attribute |
  Unsupported_Action | Unsupported_Scene_Detail | Contradiction |
  Generic | Grammar | Multiple_Issues",
  "semantic_correctness": 1,
  "detail_coverage": 1,
  "fluency_completeness": 1,
  "final_quality": "Excellent | Good | Adequate | Poor",
  "brief_reason": "short explanation"
}
```

References

1. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016*; IEEE: New York, NY, USA, 2016; pp. 770–778. [\[CrossRef\]](#)
2. Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A.A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, H.; Behl, H.; et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv* **2024**, arXiv:2404.14219.
3. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022*.
4. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-Critical Sequence Training for Image Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017*; IEEE: New York, NY, USA, 2017; pp. 7008–7024. [\[CrossRef\]](#)
5. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Computer Vision—ECCV 2014, Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Springer: Cham, Switzerland, 2014; pp. 740–755. [\[CrossRef\]](#)
6. 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 (CVPR), Boston, MA, USA, 7–12 June 2015*; IEEE: New York, NY, USA, 2015; pp. 3128–3137. [\[CrossRef\]](#)
7. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [\[CrossRef\]](#)
8. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: A Neural Image Caption Generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015*; IEEE: New York, NY, USA, 2015; pp. 3156–3164. [\[CrossRef\]](#)
9. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015*; PMLR: Cambridge, MA, USA, 2015; Volume 37, pp. 2048–2057.
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30, Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
11. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-Memory Transformer for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020*; IEEE: New York, NY, USA, 2020; pp. 10578–10587.
12. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping Language–Image Pre-Training for Unified Vision–Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML), Baltimore, MD, USA, 17–23 July 2022*; PMLR: Cambridge, MA, USA, 2022; Volume 162, pp. 12888–12900.
13. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language–Image Pre-Training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning (ICML), Honolulu, HI, USA, 23–29 July 2023*; PMLR: Cambridge, MA, USA, 2023; Volume 202, pp. 19730–19742.
14. Dai, W.; Li, J.; Li, D.; Tiong, A.M.H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; Hoi, S. InstructBLIP: Towards General-Purpose Vision–Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems 36, Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NIPS 2023), New Orleans, LA, USA, 10–16 December 2023*; Curran Associates, Inc.: Red Hook, NY, USA, 2023.
15. Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems 35, Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022*; Curran Associates, Inc.: Red Hook, NY, USA, 2022.
16. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems 36, Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NIPS 2023), New Orleans, LA, USA, 10–16 December 2023*; Curran Associates, Inc.: Red Hook, NY, USA, 2023.
17. Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; Wei, F. Kosmos-2: Grounding Multimodal Large Language Models to the World. In *Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024*.
18. Tsimpoukelli, M.; Menick, J.L.; Cabi, S.; Eslami, S.M.A.; Vinyals, O.; Hill, F. Multimodal Few-Shot Learning with Frozen Language Models. In *Advances in Neural Information Processing Systems 34, Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NIPS 2021), Virtual, 6–14 December 2021*; Curran Associates, Inc.: Red Hook, NY, USA, 2021.
19. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; de Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019*; PMLR: Cambridge, MA, USA, 2019; Volume 97, pp. 2790–2799.

20. Li, X.L.; Liang, P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Online, 1–6 August 2021; Association for Computational Linguistics: Kerrville, TX, USA, 2021; pp. 4582–4597. [\[CrossRef\]](#)
21. Lester, B.; Al-Rfou, R.; Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, 7–11 November 2021; Association for Computational Linguistics: Kerrville, TX, USA, 2021; pp. 3045–3059. [\[CrossRef\]](#)
22. Mokady, R.; Hertz, A.; Bermano, A.H. ClipCap: CLIP Prefix for Image Captioning. *arXiv* **2021**, arXiv:2111.09734.
23. 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 (ACL)*, Philadelphia, PA, USA, 7–12 July 2002; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 311–318. [\[CrossRef\]](#)
24. 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*, Ann Arbor, MI, USA, 29 June 2005; Association for Computational Linguistics: Kerrville, TX, USA, 2005; pp. 65–72.
25. Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out, Proceedings of the ACL-04 Workshop*, Barcelona, Spain, 21–26 July 2004; Association for Computational Linguistics: Kerrville, TX, USA, 2004; pp. 74–81.
26. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. CIDEr: Consensus-Based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 7–12 June 2015; IEEE: New York, NY, USA, 2015; pp. 4566–4575. [\[CrossRef\]](#)
27. Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36, Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS 2023)*, New Orleans, LA, USA, 10–16 December 2023; Curran Associates, Inc.: Red Hook, NY, USA, 2023.
28. Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; Zhu, C. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, 6–10 December 2023; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 2511–2522. [\[CrossRef\]](#)
29. Chan, D.M.; Petryk, S.; Gonzalez, J.E.; Darrell, T.; Canny, J. CLAIR: Evaluating Image Captions with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, 6–10 December 2023; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 13638–13646. [\[CrossRef\]](#)
30. 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 *Proceedings of the 38th International Conference on Machine Learning (ICML)*, Virtual, 18–24 July 2021; PMLR: Cambridge, MA, USA, 2021; Volume 139, pp. 8748–8763.
31. Luo, Z.; Hu, Z.; Xi, Y.; Zhang, R.; Ma, J. I-Tuning: Tuning Frozen Language Models with Image for Lightweight Image Captioning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 4–10 June 2023; pp. 1–5. [\[CrossRef\]](#)
32. 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 (CVPR)*, Vancouver, BC, Canada, 17–24 June 2023; IEEE: New York, NY, USA, 2023; pp. 2840–2849.
33. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *Computer Vision—ECCV 2020, 16th European Conference*, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 121–137. [\[CrossRef\]](#)
34. Wang, Z.; Yu, J.; Yu, A.W.; Dai, Z.; Tsvetkov, Y.; Cao, Y. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Virtual, 25–29 April 2022.
35. Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; Wu, Y. CoCa: Contrastive Captioners are Image-Text Foundation Models. *arXiv* **2022**, arXiv:2205.01917.
36. Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O.K.; Singhal, S.; Som, S.; et al. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 17–24 June 2023; IEEE: New York, NY, USA, 2023; pp. 19175–19186. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.