CoDe: Synergizing Contrastive Discriminative Retrieval and Confidence-Aware Fusion for Zero-Shot Image Captioning

Yuzhe Lu[1], Zhongjie Zhu[1,†], Zhijing Yu[1], Di Ge[1] and Renwei Tu[1]

[1]The Key Laboratory of Ningbo Industrial Vision and Industrial Intelligence, Zhejiang Wanli University, Ningbo, China
2023882010@zwu.edu.cn, zhuzhongjie@zwu.edu.cn, yuzhijing9265@163.com, gedizxc@163.com, turenwei@zwu.edu.cn
[†]Corresponding author: zhuzhongjie@zwu.edu.cn

***Abstract*—Zero-Shot Image Captioning (ZSIC) serves as a pivotal bridge between visual perception and language generation without relying on paired training data. However, existing retrieval-augmented methods are impeded by embedding space anisotropy and static fusion mechanisms. The former introduces hard negatives that are visually similar yet semantically contradictory, whereas the latter lacks awareness of generation uncertainty, leading to severe semantic misalignment and object hallucination. To address these challenges, we propose CoDe: Synergizing Contrastive Discriminative Retrieval and Confidence-Aware Fusion for Zero-Shot Image Captioning. First, the Contrastive Discriminative Retrieval (CDR) module constructs a discriminative semantic manifold to filter topological noise. By employing contrastive calibration, CDR eliminates geometric outliers to ensure that external memory precisely anchors to the core visual content. Second, the Confidence-Aware Fusion (CAF) module incorporates a dynamic gating unit driven by an entropy-based uncertainty metric. This mechanism adaptively modulates the contribution of visual, linguistic, and memory signals, effectively suppressing hallucinations while preserving linguistic fluency. Extensive experiments on MS COCO, Flickr30k, and NoCaps datasets demonstrate that CoDe consistently outperforms existing methods across multiple evaluation metrics, particularly exhibiting superior robustness in cross-domain scenarios.**

***Index Terms*—Zero-Shot Image Captioning, Discriminative Retrieval, Contrastive Learning, Hallucination.**

I. Introduction

Image Captioning aims to automatically generate natural language descriptions that accurately reflect the semantic content of images, serving as a pivotal task connecting computer vision and natural language processing. While advanced visual perception has made significant strides across diverse domains [1], [2], translating these rich visual semantics into natural language remains a formidable challenge. Traditional supervised learning methods [3]–[5] rely on large-scale, high-quality image-text pair datasets to train encoder-decoder models. However, this heavy reliance on paired annotated data constitutes a significant annotation bottleneck, making it difficult for models to scale to data-scarce open-domain scenarios. Recently, with the advent of large-scale vision-language pre-trained models such as CLIP [6], [7], Zero-Shot Image Captioning (ZSIC) has emerged. It aims to leverage the aligned cross-modal knowledge of pre-trained models to achieve image caption generation without training on paired data.

Existing ZSIC methods can be primarily categorized into two major paradigms: Training-Free Methods and Text-Only Training Methods. Existing training-free methods [8], [9] focus on directly bridging pre-trained VLMs with Large Language Models (LLMs) during inference. By updating the context cache via gradient backpropagation or guiding sampling with CLIP scores, these approaches offer high flexibility. However, they are often plagued by substantial latency and the risk of semantic drift due to the forced manipulation of the language generation process.

To circumvent these computational bottlenecks, Text-Only Training methods [10], [11] introduce a lightweight paradigm that fine-tunes the decoder solely on large-scale unpaired text corpora. These approaches bridge the cross-modal gap by projecting visual features into the text space, often incorporating external memory banks to enhance descriptive details and supplement concept retrieval. While this strategy effectively mitigates knowledge forgetting and boosts inference speed, current retrieval-dependent solutions still face two critical challenges.

First, due to the anisotropy and cone effect in the CLIP embedding space, the distributions of images and texts are not perfectly aligned. Existing methods relying on naive nearest neighbor retrieval often retrieve hard negatives, which refer to samples that are close in Euclidean distance but semantically mismatched. This retrieval strategy neglects the local manifold structure of the embedding space, resulting in retrieval results heavily contaminated with noise. Second, the fusion mechanism lacks confidence awareness. Current models tend to blindly fuse retrieved information without measuring the uncertainty of the retrieval results. When retrieval quality degrades due to the modality gap, the model is unable to adaptively suppress the injection of erroneous information, consequently leading to severe hallucinations.

To tackle these challenges, we introduce CoDe: Synergizing Contrastive Discriminative Retrieval and Confidence-Aware Fusion for Zero-Shot Image Captioning. First, targeting the issue of hard negatives in retrieval, we design the Contrastive Discriminative Retrieval (CDR) module. This module con-

structs a discriminative semantic space during the offline phase and utilizes a contrastive calibration mechanism during the online phase to explicitly suppress noise located at the decision boundaries, thereby eliminating distractors that are geometrically proximate but semantically outlying. Second, addressing the static blindness in the fusion process, we propose the Confidence-Aware Fusion (CAF) module. By quantifying the confidence of external memory via a dynamic gating mechanism, CAF adaptively adjusts fusion weights, thereby effectively suppressing hallucinations while fully leveraging external knowledge. Our contributions include:

- We propose the Contrastive Discriminative Retrieval (CDR) module. By analyzing the neighborhood consistency between the image and candidate texts to construct a semantic confusion penalty term, this module identifies and eliminates hard negatives that are highly similar to the query image in the visual embedding space yet semantically contradictory, thereby improving the semantic accuracy of retrieval results.
- We design the Confidence-Aware Fusion (CAF) module. CAF introduces a dynamic gating mechanism based on distributional statistics to monitor the uncertainty of the language model, visual consistency, and memory guidance in real time. Through token-level adaptive calibration, CAF suppresses hallucinations caused by misleading guidance while ensuring linguistic fluency.
- We conduct extensive comparative experiments on benchmark datasets such as MS COCO. The results indicate that our method outperforms existing baselines across multiple mainstream evaluation metrics and exhibits stronger robustness in cross-domain scenarios, validating the effectiveness of the proposed model in Zero-Shot Image Captioning tasks.

II. Related Work

Zero-Shot Image Captioning (ZS-IC) aims to break the reliance on large-scale paired annotated data, achieving cross-modal generation by leveraging only pre-trained Vision-Language Models (e.g., CLIP) and uni-modal data. Based on whether the parameters of the language decoder are updated on text-only corpora, existing methods are primarily categorized into Training-Free methods and Text-Only Training methods.

A. Training-Free Zero-Shot Image Captioning

The core paradigm of training-free methods involves aligning frozen pre-trained models via inference-time optimization. As a pioneer, ZeroCap [8] employed gradient-based guidance to minimize CLIP matching loss. However, this approach incurs high computational overhead and risks disrupting syntactic structures. To address efficiency, MAGIC [9] introduced a gradient-free re-ranking strategy to modulate output distributions, while ConZIC [12] explored non-autoregressive generation via Gibbs sampling to enable iterative refinement.

Regarding content perception, Patch Matters [13] adopted a patch-level divide-and-conquer strategy to enhance local detail capture, yet it struggles with global semantic coherence due to the lack of external knowledge guidance. Similarly, TP-Cap [14] attempted to replace retrieval with trigger-enhanced generation. Although it activates LLM capabilities via learned vectors, the complete detachment from external memory leaves the model lacking stable knowledge anchors when handling long-tail concepts, limiting generation accuracy in complex scenarios.

B. Text-Only Training Zero-Shot Image Captioning

Text-Only Training (ToT) Methods approaches fine-tune decoders on text corpora via CLIP. Early works (CapDec [15], DeCap [10]) mitigated feature distribution shifts through noise injection or memory projection, while Mac-Cap [16] improved local feature alignment via sub-region aggregation. However, these methods rely primarily on statistical alignment, lacking explicit modeling of semantic concepts. To enhance entity grounding, ViECap [17] introduced CLIP-based hard prompts but remains constrained by fixed vocabularies. Addressing this, MeaCap [11] and EntroCap [18] integrated external memory with "retrieve-filter-generate" paradigms and entropy-based screening. Yet, these pure text filtering mechanisms neglect global sentence topology, limiting complex logical generation. Recent paradigms like Patch-ioner [19] (patch-centric) and Diffusion Bridge [20] (generative alignment) explored granular and manifold alignment. Despite geometric precision, they struggle with global semantic planning and lack uncertainty awareness, leading to trade-offs between specificity and factual accuracy.

Our Approach. Different from prior arts, CoDe proposes a universal framework compatible with both Training-Free and ToT paradigms. We introduce Contrastive Discriminative Retrieval (CDR) to rectify high-dimensional semantic misalignment and Confidence-Aware Fusion (CAF) for dynamic adaptive weighting. This synergy breaks the static limitations of traditional retrieval augmentation, effectively mitigating hallucinations without requiring paired training data.

III. Methodology

The overall architecture of the proposed CoDe network is illustrated in Fig. 1. Initially, the input image is encoded into visual embeddings, serving as the query for the Contrastive Discriminative Retrieval (CDR) module. By integrating pre-computed cluster centers with a hard negative mining mechanism, CDR efficiently filters topological outliers to precisely retrieve semantically consistent candidates from the external memory bank. Subsequently, these retrieved texts are parsed into Subject-Verb-Object triplets and refined into high-confidence key concepts. These concepts serve as explicit semantic anchors, forming a multi-modal input stream alongside the original visual features that is fed into the CBART [21] backbone. At the core of the generator lies the Confidence-Aware Fusion (CAF) module. Driven by matching scores from Sentence-BERT and CLIP, a dynamic weight calibration unit adaptively modulates the fusion ratios of visual, linguistic, and memory signals, thereby achieving superior feature synergy and cross-modal alignment. Finally, the decoder selects the

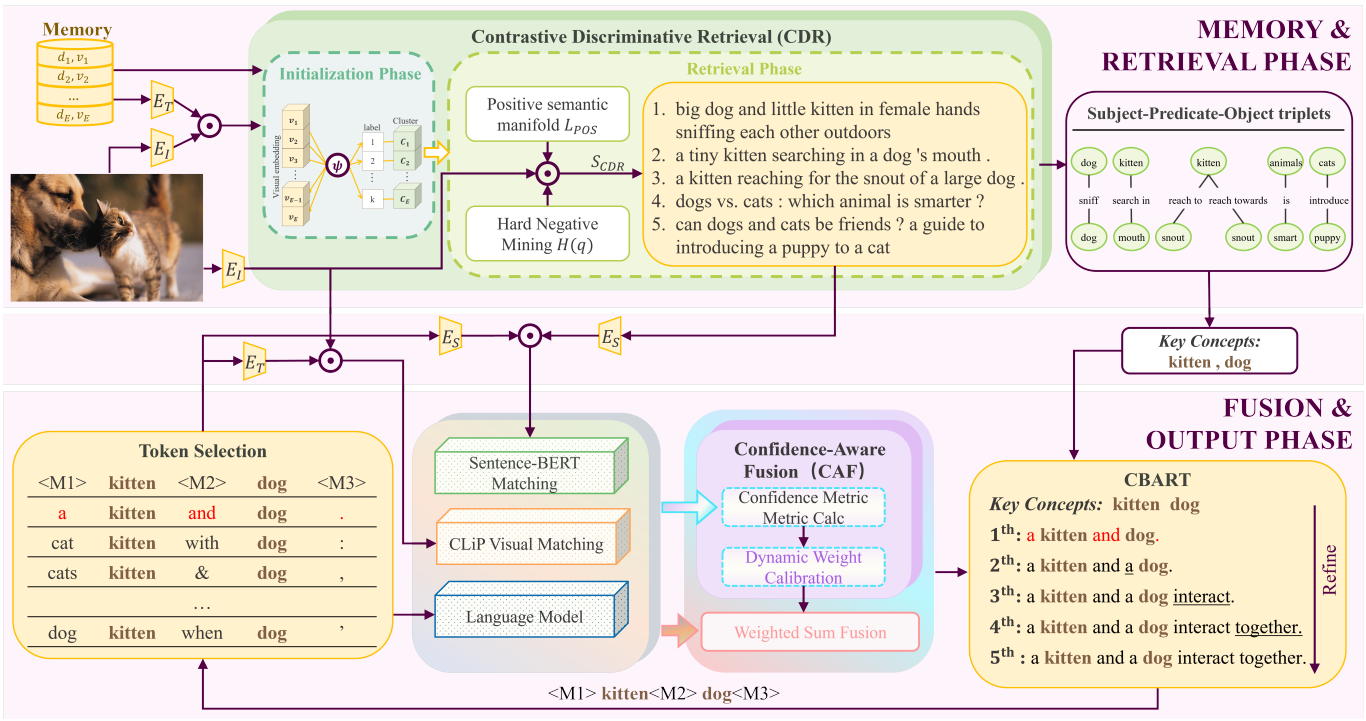

Fig. 1. Overview of the proposed CoDe framework. The pipeline consists of two phases: *i)* Memory Retrieval Phase: Given an input image, the CDR module utilizes positive semantic manifolds and Hard Negative Mining $\mathcal{H}(q)$ to retrieve matching descriptions (Sec. 3.1), which are transformed into triplets to extract Key Concepts (Sec. 3.2). *ii)* Fusion & Output Phase: Based on these concepts, the CAF module calculates confidence scores for visual (E_I), semantic (E_S), and language flows. Via Dynamic Weight Calibration, it adaptively fuses multi-modal signals to guide the CBART backbone in generating descriptions through Iterative Refinement (Sec. 3.3). E_I, E_T, and E_S denote the CLIP visual, CLIP text, and Sentence-BERT encoders, respectively; $\odot$ denotes similarity calculation.

optimal token based on the fused probability distribution and employs an iterative refinement strategy to progressively generate and optimize the descriptive text sequence.

A. Contrastive Discriminative Retrieval

Conventional retrieval-based ZSIC approaches predominantly rely on nearest neighbor search using pre-trained global features (e.g., CLIP). However, we identify a critical structural deficiency in this point-to-point cosine similarity mechanism: the high-dimensional embedding space suffers from visual-semantic misalignment, where visually proximal samples may be semantically contradictory. Such retrieval lacks global structural awareness, easily recalling hard negatives that induce factual hallucinations in downstream generation. To mitigate this, we propose the Contrastive Discriminative Retrieval (CDR) module. CDR formalizes the retrieval process into two distinct phases: an offline construction of a discriminative semantic space, and an online contrastive calibration via hard negative mining.

Discriminative Semantic Space Construction. To capture the latent structure of the external knowledge base, we first establish a semantic index during the offline initialization phase. We define the external memory bank as $\mathcal{M} = \{(d_i, v_i)\}_{i=1}^{E}$, where E represents the scale of the memory bank, d_i denotes the i-th text description serving as the retrieval target for the generative model, and v_i is the corresponding normalized visual embedding extracted by the pre-trained CLIP text encoder.

To transform the unstructured $\mathcal{M}$ into an ordered semantic topology, we employ the MiniBatchKMeans algorithm to perform unsupervised partitioning on the visual embedding set $\{v_i\}_{i=1}^{E}$. Our goal is to find a set of optimal semantic clusters $\mathcal{C} = \{C_1, \ldots, C_K\}$ to minimize the intra-cluster variance:

$$\min_{\{\mu_j\}_{j=1}^{K}} \mathcal{J}_{cluster} = \sum_{j=1}^{K} \sum_{v_i \in C_j} \|v_i - \mu_j\|_2^2 \tag{1}$$

where K is the pre-defined number of semantic clusters, representing the granularity of the semantic space; μ_j denotes the centroid vector of the j-th semantic cluster C_j; and $\mathcal{J}_{cluster}$ is the target inertia value for optimization. We construct a topological mapping operator Φ:

$$\Phi(v) = \underset{j \in \{1,\ldots,K\}}{\operatorname{argmin}} \|v - \mu_j\|_2 \tag{2}$$

The function $\Phi(v)$ maps any continuous visual embedding to a discrete semantic label $l \in \{1, \ldots, K\}$, *i.e.*, the Semantic Cluster Index. This offline indexing mechanism defines clear decision boundaries for the memory bank, serving as the structural basis for subsequently distinguishing between homogeneous positives and heterogeneous negatives.

Hard Negative Mining and Calibration. In the Retrieval Phase, given a query image I, we first extract its visual

embedding $q = E_I(I)$. CDR then performs a preliminary retrieval, recalling the Top-$3N$ candidate samples based on cosine similarity to form the set $\mathcal{P}_{init}$. Utilizing the mapping operator Φ, we lock the set of positive semantic manifolds to which these high-confidence samples belong:

$$\mathcal{L}_{pos} = \{\Phi(v) \mid v \in \mathcal{P}_{init}\} \tag{3}$$

To enhance the discriminative power of retrieval, we explicitly mine hard negatives that easily confuse the model. We define the hard negative set $\mathcal{H}(q)$ as samples that are visually extremely similar but semantically belong to mutually exclusive regions:

$$\mathcal{H}(q) = \underset{v \in \mathcal{M}, \Phi(v) \notin \mathcal{L}_{pos}}{\text{Top-M}} \quad q^\top v \tag{4}$$

where the constraint $\Phi(v) \notin \mathcal{L}_{pos}$ strictly requires negative samples to originate from non-target clusters; M is the budget size for negative mining; and $q^\top v$ denotes the cosine similarity between the query image and the memory item. This step explicitly captures the visual traps that lead the model to generate hallucinations.

To suppress visual traps while preserving true semantic associations, we design a calibration function inspired by the InfoNCE loss to re-rank candidate samples. Specifically, if a candidate sample is too close to a hard negative in the embedding space, it indicates that it carries confusing semantic components and should be down-weighted. For any initial candidate $v^+ \in \mathcal{P}_{init}$, its calibrated score S_{CDR} is defined as:

$$S_{CDR} = q^\top v^+ - \tau \cdot \psi(v^+, \mathcal{H}) \tag{5}$$

where S_{CDR} is the final Contrastive Score, and τ is a penalty weight hyperparameter controlling the intensity of the semantic distinctiveness constraint. $\psi(v^+, \mathcal{H})$ is our proposed semantic confusion penalty term, which explicitly models the degree of semantic overlap between the candidate sample and the hard negatives:

$$\psi(v^+, \mathcal{H}) = \max_{v^- \in \mathcal{H}(q)} {v^+}^\top v^- \tag{6}$$

Intuitively, $\psi(v^+, \mathcal{H})$ measures the semantic affinity between the candidate v^+ and its most confusing negative sample. When v^+ is highly similar to a certain hard negative v^-, the penalty term ψ increases significantly, thereby reducing the final score of that candidate. Finally, we re-rank the candidate set based on the calibrated scores and select the top-N samples:

$$Mem = \underset{v \in \mathcal{P}_{init}}{\text{Top-N}} \left(S_{CDR}\right) \tag{7}$$

where Mem denotes the final output set of memory items, which will be fed into the subsequent concept extraction and text generation modules. Through this process, CDR achieves a paradigm shift from geometric nearest neighbor search to semantic discriminative retrieval.

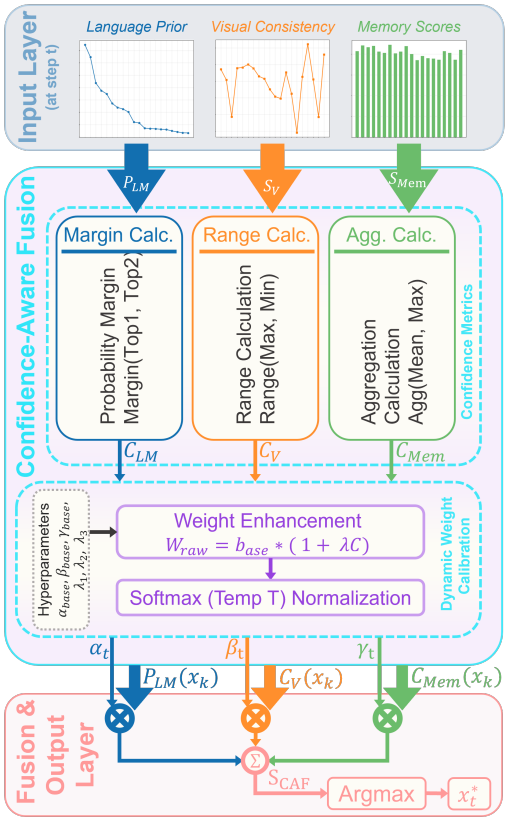

Fig. 2. Schematic of the Confidence-Aware Fusion (CAF) module. The pipeline proceeds in two stages: *i)* Metric Calculation: Derives confidence metrics (C_{LM}, C_V, C_{Mem}) via statistical aggregation of language, visual, and memory signals. *ii)* Dynamic Weight Calibration: Generates adaptive weights ($\alpha_t, \beta_t, \gamma_t$) via Softmax to integrate multi-source distributions for the final token prediction x_t^*.

B. Concept Extraction and Generation Architecture

To bridge the gap between the retrieved unstructured text and the generation model, we adopt the "retrieve-filter-generate" paradigm from MeaCap. After obtaining the memory set Mem with high semantic consistency via the CDR module, we need to extract key visual concepts from it to guide the subsequent generation.

Specifically, we utilize an off-the-shelf Textual Scene Graph Parser to parse each sentence in Mem into (Subject-Predicate-Object) triplets. Through statistical clustering and frequency filtering, we extract the top-ω Key Concepts, denoted as Key_ω, that are most relevant to the image. These concepts not only preserve the core semantics of the retrieved text but also effectively eliminate the interference of redundant modifiers.

In the generation phase, we employ the pre-trained CBART as the backbone network. Unlike traditional autoregressive models, CBART is capable of refining sentences in parallel through a BERT-like Masked Token Prediction task, given the constraint tokens (i.e., the extracted concepts Key_ω). In the t-th iteration, the model predicts the token probability distribution P_{LM} for the masked positions based on the current

context $x_{<t}$:

$$P_{LM}(x_k) = P(x_k \mid x_{<t}; \theta_{LM}) \tag{8}$$

where x_k represents the candidate token, and θ_{LM} represents the frozen language model parameters.

Realizing that relying solely on linguistic priors leads to hallucinations, Zeng *et al.* introduced visual consistency (CLIP) and memory similarity (Sentence-BERT) as supplementary signals:

$$S_V(x_k) = \frac{E_I(I)^\top E_T(x_k)}{\|E_I(I)\|\|E_T(x_k)\|} \tag{9}$$

$$S_M(x_k) = \frac{\mu_m^\top E_S(x_k)}{\|\mu_m\|\|E_S(x_k)\|} \tag{10}$$

where $E_I(\cdot)$ and $E_T(\cdot)$ are the image and text encoders of CLIP, respectively; $E_S(\cdot)$ denotes the Sentence-BERT encoder; I is the query image; and μ_m is the average feature vector of the Top-N retrieval results. Thus, for each candidate token, we obtain basic signals from three dimensions: P_{LM} (Grammaticality), S_V (Visualness), and S_M (Knowledge).

C. Confidence-Aware Fusion

Conventional static weighting for language (P_{LM}), visual (S_V), and memory (S_M) signals is suboptimal, as it neglects temporal dynamics where specific tokens (e.g., function words vs. entities) demand distinct modal reliance. To address this rigidity, we propose the Confidence-Aware Fusion (CAF) module (Fig. 2). By quantifying real-time confidence via predictive distributions, CAF dynamically calibrates fusion ratios, enabling adaptive modulation that aligns with the varying reliability of each modality.

Multimodal Scoring and Confidence Metrics. To achieve dynamic calibration of heterogeneous signals, we derive confidence metrics based on the statistical attributes of the base scores on the candidate set $\mathcal{V}_t$ at each step t.

First, for the language prior, we quantify the model's certainty via the "sharpness" of its distribution P_{LM}. Using the Probability Margin, we compute C_{LM} as the normalized difference between the top-1 and top-2 probabilities:

$$C_{LM} = \left[\frac{P_{LM}(x_{(1)}) - P_{LM}(x_{(2)})}{P_{LM}(x_{(1)}) + \epsilon} \right]_0^1 \tag{11}$$

where $x_{(1)}$ and $x_{(2)}$ denote the top-2 candidates in $\mathcal{V}_t$, and ϵ is a smoothing term to prevent division by zero. A high C_{LM} implies strong internal certainty, indicating the language model should dominate to minimize external interference.

Simultaneously, regarding visual consistency, we measure the Discriminative Power of visual signals. A large score dispersion implies high distinguishability, whereas a flat distribution suggests ambiguity. Thus, we define C_V as the Range of base scores:

$$C_V = \left[\eta \cdot \left(\max_{x \in \mathcal{V}_t} S_V(x) - \min_{x \in \mathcal{V}_t} S_V(x) \right) \right]_0^1 \tag{12}$$

where η is a scaling factor. This ensures visual weights are boosted only when they clearly distinguish candidates, avoiding noise during abstract concept generation.

Finally, for memory guidance, we assess reliability by combining Overall Relevance (mean) and Peak Matching (max). Consequently, we define the memory confidence C_M as the aggregation of the mean and maximum values of the memory similarity scores S_M:

$$C_M = \left[\frac{1}{2} \left(\frac{1}{|\mathcal{V}_t|} \sum_{x \in \mathcal{V}_t} S_M(x) + \max_{x \in \mathcal{V}_t} S_M(x) \right) \right]_0^1 \tag{13}$$

This metric adaptively enhances guidance when external knowledge aligns with the context, while automatically reducing dependency on low-relevance retrieval to prevent erroneous knowledge injection.

Dynamic Weight Calibration. Upon obtaining the real-time confidence for each modality, CAF introduces a two-stage gating mechanism to dynamically adjust the fusion ratios. Our goal is to adaptively scale the weights based on the uncertainty state of the current generation step while preserving the baseline model priors. First, we augment the base weights using the confidence metrics to derive the raw weight vector $\mathbf{w}_{raw}$:

$$\mathbf{w}_{raw} = \begin{cases} \alpha_{base}(1 + \lambda_1 C_{LM}) \\ \beta_{base}(1 + \lambda_2 C_V) \\ \gamma_{base}(1 + \lambda_3 C_M) \end{cases} \tag{14}$$

where $\alpha_{base}, \beta_{base}, \gamma_{base}$ are preset hyperparameters representing the global importance priors of each modality, and $\lambda_1, \lambda_2, \lambda_3$ are modulation factors controlling the gain magnitude of confidence on weights. This design ensures that even if a modality's base weight is low, it can still dominate the current generation step provided its confidence is sufficiently high.

To map the modulated weights to a valid probability distribution and control the "sharpness" of the fusion strategy, we subsequently employ a Softmax function with a temperature coefficient to compute the final dynamic weights $(\alpha_t, \beta_t, \gamma_t)$:

$$[\alpha_t, \beta_t, \gamma_t] = \mathrm{Softmax}(\mathbf{w}_{raw}/T) \tag{15}$$

where T is a temperature hyperparameter regulating the model's sensitivity to high-confidence signals. Finally, at generation step t, we perform a weighted sum of the normalized scores from the three heterogeneous sources to obtain the final fusion score S_{CAF}:

$$S_{CAF}(x_k) = \sum \begin{cases} \alpha_t \cdot P_{LM}(x_k) \\ \beta_t \cdot S_V(x_k) \\ \gamma_t \cdot S_M(x_k) \end{cases} \tag{16}$$

The model selects the optimal token x_t^* by maximizing S_{CAF} within $\mathcal{V}_t$. This mechanism transforms generation from rigid static interpolation into an uncertainty-driven adaptive process. Specifically, CAF prioritizes linguistic fluency during low-entropy states while pivoting to visual-memory guidance

TABLE I
ZERO-SHOT RESULTS ON MS COCO AND NOCAPS. TRAINING AND MEMORY DENOTE LM FINE-TUNING AND EXTERNAL MEMORY. NOCAPS REPORTS CIDER ON IN/NEAR/OUT SPLITS. BOLD INDICATES BEST RESULTS. TF/TOT: TRAINING-FREE/TEXT-ONLY TRAINING.

Models	Text Corpus		MS COCO				NoCaps (CIDEr)			
	Training	Memory	B@4	M	C	S	In	Near	Out	Overall
ZeroCap [8]	×	×	2.6	11.5	14.6	5.5	13.3	14.9	19.7	16.6
Tewel *et al.* [30]	×	×	2.2	12.7	17.2	7.3	13.7	15.8	18.3	16.9
ConZIC [12]	×	×	1.3	11.2	13.3	5.0	15.4	16.0	20.3	17.5
CLIPRe [31]	×	CC3M	4.6	13.3	25.6	9.2	23.3	26.8	36.5	28.2
MeaCap$_{TF}$ [11]	×	CC3M	4.5	14.1	26.0	9.4	23.0	27.4	33.9	29.7
MeaCap$_{ToT}$ [11]	CC3M	CC3M	7.8	15.2	38.8	10.7	30.8	36.6	45.3	39.8
CoDe$_{TF}$ (Ours)	×	CC3M	5.1	15.1	30.8	10.3	23.7	29.0	36.3	31.7
CoDe$_{ToT}$ (Ours)	CC3M	CC3M	**9.1**	**16.2**	**42.2**	**11.2**	**31.1**	**36.9**	**48.8**	**40.9**

TABLE II
IN-DOMAIN AND CROSS-DOMAIN UNPAIRED CAPTIONING RESULTS. TOT MODELS ARE FINE-TUNED ON COCO TEXT. BOLD DENOTES BEST. TF/TOT: TRAINING-FREE/TEXT-ONLY TRAINING.

Methods	MSCOCO				MSCOCO → Flickr30k			
	B@4	M	C	S	B@4	M	C	S
MAGIC [9]	12.9	17.4	49.3	11.3	6.2	12.2	17.5	5.9
CLIPRe [31]	12.4	20.4	53.4	14.8	9.8	16.7	30.1	10.3
MeaCap [11]	18.2	22.9	73.7	17.1	11.7	17.3	34.3	11.1
CoDe$_{TF}$ (Ours)	8.6	19.2	48.8	14.6	7.1	14.7	21.5	10.4
CoDe$_{ToT}$ (Ours)	**19.5**	**23.3**	**79.1**	**17.8**	**12.7**	**17.5**	**34.5**	**11.2**

TABLE III
ABLATION STUDIES ON THE MS COCO DATASET. WE EVALUATE THE EFFECTIVENESS OF CDR AND CAF MODULES UNDER TRAINING-FREE (TF) AND TEXT-ONLY TRAINING (TOT) SETTINGS, RESPECTIVELY. BOLD NUMBERS DENOTE THE BEST PERFORMANCE.

Methods	CDR	CAF	MSCOCO			
			B@4	M	C	S
CoDe$_{TF}$	×	✓	4.7	14.3	28.9	9.8
	✓	×	4.9	14.4	28.2	9.9
	✓	✓	**5.1**	**15.1**	**30.8**	**10.3**
CoDe$_{ToT}$	×	✓	8.3	15.9	40.9	10.9
	✓	×	8.0	15.8	40.3	10.8
	✓	✓	**9.1**	**16.2**	**42.2**	**11.2**

for high-entropy entities, effectively mitigating hallucinations without compromising syntactic coherence.

IV. EXPERIMENTS

A. Experimental Settings

Datasets and Metrics. We conduct experiments on MS COCO [22] and Flickr30k [23] using standard Karpathy splits, and NoCaps [24] to evaluate open-domain generalization. Following MeaCap, we use the text-only CC3M [25] corpus for memory bank construction and optional text-only fine-tuning. We report standard captioning metrics: B@4, M, C, and S (BLEU@4 [26], METEOR [27], CIDEr [28], and SPICE [29]).

Implementation Details. CoDe builds upon frozen foundation models: CLIP-ViT-B/32 for cross-modal similarity, Sentence-BERT (all-MiniLM-L6-v2) for retrieval, and CBART-large as the generative backbone. To ensure robust generalization, all key hyperparameters were determined through empirical validation. For the CDR module, we set semantic clusters $K = 50$ to provide optimal discriminative power without semantic fragmentation. During inference, we retrieve top-$3N$ ($N = 5$) candidates for calibration and retain top-N for concept extraction, a balance empirically chosen to maximize semantic diversity while mitigating topological noise. For the CAF module, we configure the sequence length range as $[10, 20]$, refinement steps as 10, and repetition penalty as 2.0. The candidate pool size is $|\mathcal{V}_t| = 200$. Base weights are initialized to $\alpha_{base} = 0.1$, $\beta_{base} = 0.4$, $\gamma_{base} = 0.2$ with a dynamic temperature $T = 0.3$. We evaluate two variants: CoDe$_{TF}$ (Training-Free, memory only) and CoDe$_{ToT}$ (Text-Only Training, fine-tuned on CC3M text).

B. Comparative Experiments

To validate the effectiveness of CoDe under different training paradigms, we quantitatively compare the proposed method with existing Training-Free methods (ZeroCap [8], Tewel *et al.* [30], ConZIC [12], CLIPRe [31], MeaCap$_{TF}$ [11]) and Text-Only Training methods (MeaCap$_{ToT}$) on the MS COCO and NoCaps datasets. Table I presents the detailed comparison results on the MS COCO Karpathy test split and the NoCaps validation set. In the Training-Free setting, CoDe$_{TF}$ significantly outperforms MeaCap$_{TF}$ in terms of CIDEr and SPICE scores. This improvement in semantic consistency is attributed to the CDR module, which eliminates hard negatives by constructing a discriminative space. This effectively alleviates visual-semantic misalignment and provides precise anchors for generation.In the Text-Only Training setting, CoDe$_{ToT}$ further surpasses MeaCap$_{ToT}$ comprehensively. This gain stems from the CAF module, which dynamically regulates fusion weights based on generation uncertainty. By rectifying factual biases inherent in static strategies, CAF achieves precise control over long-sentence descriptions.In terms of cross-domain generalization, CoDe$_{ToT}$ achieves a CIDEr score of 48.8 on the NoCaps out-of-domain test set. This indicates that when encountering unseen concepts, the synergy between the discriminative retrieval of CDR and the

Meacap: A domestic cat sitting on a lap with the laptop.
CoDe: A baby elephant in the wild with other elephants.
GT: An adult elephant and baby elephant loving on one another.

Meacap: A dog holding a frisbee is walking away from a hand in the grass.
CoDe: A dog holding a frisbee in its mouth.
GT: A dog holding a yellow frisbee in it's mouth.

Meacap: The annual fireworks display raises money for a local church.
CoDe: The annual fireworks display over late gothic revival structure and the clock tower.
GT: A night sky view of a tower and fireworks behind it.

Meacap: A domestic cat covered in the winter snow.
CoDe: A cute pretty cat hiding under a blanket.
GT: A cat is peeping out from under a blanket.

Meacap: A giant panda with his new mother.
CoDe: A giant panda with her cute baby.
GT: A mother and baby panda are sitting together.

Fig. 3. Qualitative examples of image captioning results from MeaCap and CoDe, along with corresponding human-annotated Ground Truth (GT).

adaptive correction of CAF ensures high descriptive accuracy and robustness in open-world scenarios.

To evaluate the model's adaptability to unpaired data and cross-domain scenarios, we fine-tune the model solely on the MSCOCO text corpus and evaluate it on both in-domain and Flickr30k cross-domain test sets (Table II).In the in-domain unpaired setting, CoDe$_{ToT}$ demonstrates a significant advantage, achieving a CIDEr score of 79.1, which substantially surpasses the strongest baseline, MeaCap (73.7). This improvement in semantic fidelity is attributed to the CAF module. By employing dynamic weight calibration, CAF effectively mitigates the knowledge forgetting induced by text-only training, maximizing visual consistency while preserving linguistic fluency. In cross-domain transfer scenarios, CoDe$_{ToT}$ consistently outperforms all comparative methods. This superior robustness stems from the discriminative nature of the CDR module. The structured semantic prototypes constructed by CDR extract domain-invariant visual concepts, ensuring the retrieval of accurate semantic anchors even when confronting novel distributions.

C. Ablation Studies

To systematically evaluate the contributions of key components within the CoDe framework, we conduct comprehensive ablation studies (see Table III). These experiments are designed to investigate the individual contributions and synergistic effects of Contrastive Discriminative Retrieval (CDR) and Confidence-Aware Fusion (CAF), encompassing both the Training-Free (CoDe$_{TF}$) and Text-Only Training (CoDe$_{ToT}$) settings.

The results validate the criticality and significant synergy of both modules. First, regarding CDR, reverting to nearest neighbor retrieval causes consistent degradation, exemplified by a CIDEr drop from 30.8 to 28.9 in the CoDe$_{TF}$ setting. This confirms that topological calibration is essential to filter semantically conflicting hard negatives. Second, regarding CAF, substituting it with static fusion impairs performance, with the SPICE score of CoDe$_{ToT}$ declining from 11.2 to 10.8. This indicates that static weights fail to adapt to dynamic uncertainty, thereby inducing hallucinations. Finally, the full model CoDe$_{ToT}$ achieves an optimal CIDEr score of 42.2, underscoring their complementarity, where CDR ensures input precision by purifying semantic anchors, while CAF guarantees decoding robustness via adaptive gating.

D. Visualization Analysis

To intuitively assess generation quality, Fig. 3 presents qualitative comparisons between the baseline MeaCap and our CoDe on the MSCOCO test set. The visual examples clearly demonstrate CoDe's significant advantages in suppressing object hallucinations and capturing fine-grained semantics.

Observations reveal that the baseline suffers from severe visual-semantic misalignment and hallucinations. For instance, as shown in the fourth column, MeaCap misinterprets a white blanket as "winter snow" due to embedding space anisotropy, where texture similarity misguides retrieval. In contrast, CoDe accurately identifies "hiding under a blanket," validating that CDR effectively filters visually similar yet semantically contradictory hard negatives via discriminative manifold construction. Moreover, the baseline exhibits blind reasoning driven by language priors, hallucinating non-existent details like "raising money" or "children riding" in the first and third columns. Conversely, CoDe captures faithful details such as "gothic revival" by leveraging CAF. By dynamically monitoring uncertainty, CAF suppresses hallucinations during modal conflicts, forcing the generation trajectory to align strictly with visual facts.

V. Conclusion

We present CoDe, a framework synergizing Contrastive Discriminative Retrieval and Confidence-Aware Fusion to address visual-semantic misalignment and static fusion limitations in Zero-Shot Image Captioning. Specifically, the Contrastive Discriminative Retrieval (CDR) module employs discriminative manifold calibration to mitigate embedding space anisotropy. By filtering geometrically proximal yet semantically contradictory topological noise, CDR ensures precise semantic anchoring of external memory. Complementarily, the Confidence-Aware Fusion (CAF) module utilizes an entropy-based dynamic gating unit to adaptively modulate multi-modal weights.

This mechanism effectively resolves uncertainty-agnostic hallucinations, enforcing faithfulness to visual facts while preserving linguistic fluency. Supporting both Training-Free and Text-Only Training paradigms, CoDe achieves efficient inference without paired annotations. Extensive experiments on MS COCO, Flickr30k, and NoCaps confirm its superiority over state-of-the-art methods. Future work will extend this framework to spatiotemporal tasks like video captioning and optimize retrieval efficiency for large-scale applications.

VI. Acknowledgment

This work was supported in part by the Key Program of Zhejiang Province Natural Science Foundation (LZ24F010004), Zhejiang Provincial "Pioneer, Leader+ X" Science and Technology Program, Ningbo Municipal "Yongjiang Science and Technology Innovation 2035" Major Application Demonstration Program(2024Z023), Ningbo Municipal "Yongjiang Science and Technology Innovation 2035" Key R&D Program (2024Z122, 2024Z295, 2025Z040) ,Ningbo Municipal Major Project of Science and Technology Innovation 2025 (2022Z076)

References

[1] Z. Zhu, R. Zhang, Y. Bai, Y. Wang, and J. Sun, "Bilateral cross enhancement with self-attention compensation for semantic segmentation of point clouds," *Journal of Image and Graphics*, vol. 29, no. 8, pp. 2388–2398, 2024, doi: 10.11834/jig.230430.

[2] Z. Zhu, L. Zhang, P. Li, R. Tu, Y. Bai, and Y. Wang, "Text image super-resolution via structure perception enhancement and cross modal fusion," *Journal of Image and Graphics*, vol. 30, no. 5, pp. 1364–1376, 2025, doi: 10.11834/jig.240559.

[3] W. Jiang, T. Xu, H. Li, Z. Li, Y. Fang, and Z. Tang, "What happens in the surroundings: A benchmark for 360° image captioning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2025, pp. 1–8.

[4] G. Yang, D. Wu, J. An, S. Xiao, and Z. Lin, "Beyond similarity: Aligning retrieval with generation for enhanced image captioning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2025, pp. 1–8.

[5] A. Izhar, N. Japar, N. Idris, and T. Dang, "MicarVLMoE: A modern gated cross-aligned vision-language mixture of experts model for medical image captioning and report generation," *arXiv preprint arXiv:2504.20343*, 2025.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. N. Am. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol. (NAACL HLT)*, vol. 1, 2019, pp. 4171–4186.

[7] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[8] Y. Tewel, Y. Shalev, I. Schwartz, and L. Wolf, "Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 17918–17928.

[9] S. Yixuan, L. Tian, L. Yahui, L. Fangyu, Y. Dani, W. Yan, K. Lingpeng, and C. Nigel, "Language models can see: Plugging visual controls in text generation," *Computing Research Repository*, 2023.

[10] W. Li, L. Zhu, L. Wen, and Y. Yang, "Decap: Decoding clip latents for zero-shot captioning via text-only training," *arXiv preprint arXiv:2303.03032*, 2023.

[11] Z. Zeng, Y. Xie, H. Zhang, C. Chen, Z. Wang, and B. Chen, "MeaCap: Memory-augmented zero-shot image captioning," *Computing Research Repository*, 2024.

[12] Z. Zeng, H. Zhang, R. Lu, D. Wang, B. Chen, and Z. Wang, "Conzic: Controllable zero-shot image captioning by sampling-based polishing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 23465–23476.

[13] R. Peng, H. He, Y. Wei, Y. Wen, and D. Hu, "Patch matters: Training-free fine-grained image caption enhancement via local perception," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 3963–3973.

[14] R. Zhang, L. Wang, Y. He, T. Pan, Z. Yu, and Y. Li, "Tpcap: Unlocking zero-shot image captioning with trigger-augmented and multi-modal purification modules," *arXiv preprint arXiv:2502.11024*, 2025.

[15] D. Nukrai, R. Mokady, and A. Globerson, "Text-only training for image captioning using noise-injected clip," *arXiv preprint arXiv:2211.00575*, 2022.

[16] L. Qiu, S. Ning, and X. He, "Mining fine-grained image-text alignment for zero-shot captioning via text-only training," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 5, Mar. 2024, pp. 4605–4613.

[17] J. Fei, T. Wang, J. Zhang, Z. He, C. Wang, and F. Zheng, "Transferable decoding with visual entities for zero-shot image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 3136–3146.

[18] J. Yan, Y. Xie, S. Zou, Y. Wei, and X. Luan, "EntroCap: Zero-shot image captioning with entropy-based retrieval," *Neurocomputing*, vol. 611, p. 128666, 2025.

[19] L. Bianchi, G. Pacini, F. Carrara, N. Messina, G. Amato, and F. Falchi, "One patch to caption them all: A unified zero-shot captioning framework," *arXiv preprint arXiv:2510.02898*, 2025.

[20] J. R. Lee, Y. Shin, G. Son, and D. Hwang, "Diffusion bridge: Leveraging diffusion model to reduce the modality gap between text and vision for zero-shot image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 4050–4059.

[21] X. He, "Parallel refinements for lexically constrained text generation with bart," *arXiv preprint arXiv:2109.12487*, 2021.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," *Lecture Notes in Computer Science*, vol. 8693, pp. 740–755, 2014.

[23] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, 2014.

[24] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson, "Nocaps: Novel object captioning at scale," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 8947–8956.

[25] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. 56th Annu. Meet. Assoc. Comput. Linguistics (Vol. 1: Long Papers)*, Jul. 2018, pp. 2556–2565.

[26] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meet. Assoc. Comput. Linguistics*, Jul. 2002, pp. 311–318.

[27] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. and/or Summarization*, Jun. 2005, pp. 65–72.

[28] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 4566–4575.

[29] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2016, pp. 382–398.

[30] Y. Tewel, Y. Shalev, R. Nadler, I. Schwartz, and L. Wolf, "Zero-shot video captioning with evolving pseudo-tokens," *arXiv preprint arXiv:2207.11100*, 2022.

[31] W. Li, L. Zhu, L. Wen, and Y. Yang, "Decap: Decoding clip latents for zero-shot captioning via text-only training," *arXiv preprint arXiv:2303.03032*, 2023.