Bridging the Modality Gap in Zero-Shot Image Captioning: An Implementation of IFCap

Jeongwon Bae (945397461)
jqb6679@psu.edu

December 13, 2024

1 Task

The objective of this project is to address the challenge of zero-shot image captioning, which involves generating natural language descriptions for images without utilizing any paired image-caption data during the training phase. In zero-shot image captioning, models must produce accurate and contextually relevant captions for images they have never seen before, relying solely on knowledge acquired from unpaired text data.

This task is particularly challenging due to the inherent modality gap between visual and textual information. Models trained exclusively on text lack direct visual grounding, making it difficult to generate descriptions that accurately reflect the content of unseen images. Capturing the complexity of visual scenes, including objects, actions, and spatial relationships, is especially difficult without visual examples during training. Ensuring that the generated captions are not only grammatically correct but also semantically relevant and detailed requires innovative approaches to bridge the gap between text and image modalities.

Zero-shot image captioning has significant implications for various applications, such as improving accessibility technologies for visually impaired users, enhancing image retrieval systems, and automating content generation processes. By effectively addressing this task, we can develop models that generate high-quality captions without the need for extensive paired datasets, thereby reducing resource requirements and enabling broader applicability.

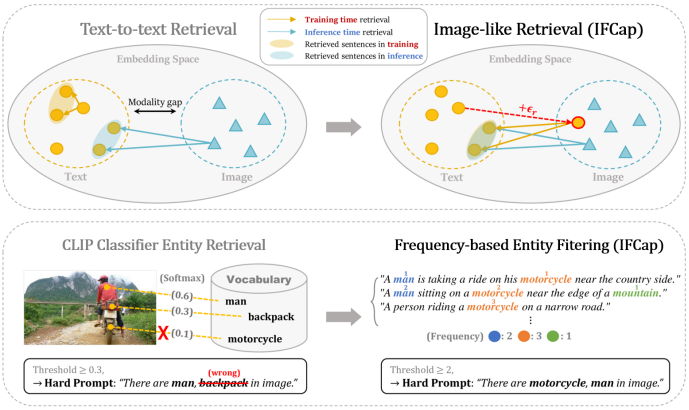

Figure 1: (Top) Traditional text-to-text retrieval ignores the modality gap, causing mismatches between training and inference. The IFCap method bridges this gap by aligning text features with the image embedding space during retrieval. (Bottom) CLIP classifier-based entity retrieval struggles with larger vocabularies, reducing accuracy. The IFCap approach improves entity extraction by focusing on frequently occurring words in retrieved captions, eliminating reliance on limited vocabularies.

2 Related Work

Zero-shot image captioning has attracted considerable interest due to its potential to generate image descriptions without relying on paired image-text datasets. Several approaches have been proposed to tackle this task, each attempting to bridge the modality gap between visual and textual domains.

CapDec [1] introduced a method that assumes image embeddings are located near text embeddings within a shared representation space. To mitigate the modality gap, they injected noise into text embeddings during training to simulate image-like representations. While this approach showed promise, it relies heavily on the assumption about embedding distributions, which may not hold universally and could limit its generalizability across different datasets or languages.

ViECap [2] leveraged entity extraction from input text and images to construct hard prompts, guiding the language model to generate captions that include specific objects. This method enhances the model's ability to incorporate novel objects into the captions. However, it is sensitive to incorrect entity extraction; misidentified entities can lead to irrelevant or inaccurate captions, adversely affecting performance.

Knight [3] utilized a retrieval mechanism to extract external knowledge from a predefined text corpus. By incorporating retrieved captions as additional context, the model can generate more informative and detailed descriptions. Nevertheless, this approach suffers from the modality gap during inference, as it relies on text-to-text retrieval without adequately aligning with image features. Consequently, the generated captions may not accurately reflect the visual content of the images.

CLOSE [4] addressed the modality gap by employing scaled noise injection techniques and hyperparameter tuning to improve the alignment between textual and visual modalities. Although this method achieved promising results, it requires careful tuning of hyperparameters, which can be time-consuming and may not yield consistent results across different datasets or tasks.

The state-of-the-art method for tackling zero-shot image

captioning is IFCap (Image-like Retrieval and Frequency-based Entity Filtering for Zero-shot Captioning) proposed by Lee et al. [5]. IFCap introduces innovative solutions to the challenges identified in previous works. First, it employs Image-like Retrieval to perform text retrieval in a manner that mimics image-to-text retrieval outcomes. By injecting noise into the CLIP [6] text embeddings of input captions, the method adjusts text features to align more closely with the distribution of image features, effectively bridging the modality gap.

Second, IFCap utilizes Frequency-based Entity Filtering to extract entities from retrieved captions based on their occurrence frequency. Unlike methods that rely on predefined vocabularies or external object detectors, this technique leverages the frequency of nouns in the retrieved captions to select relevant entities. This enhances caption accuracy without depending on limited vocabularies, making the model more flexible and scalable.

By addressing the key difficulties in zero-shot image captioning (namely, the modality gap and accurate entity inclusion) IFCap has demonstrated superior performance on standard benchmarks, outperforming previous methods on metrics such as CIDEr and SPICE. Its innovative techniques contribute to more accurate and semantically rich caption generation in a zero-shot setting.

3 Approach

Our approach involves implementing the IFCap model to address the challenges of zero-shot image captioning. The model introduces two key components, Image-like Retrieval (ILR) and Frequency-based Entity Filtering (EF), which work together to bridge the modality gap between text and images and enhance the accuracy of entity inclusion in generated captions. We also utilize a Fusion Module (FM) to integrate information from different sources. Figure 2 provides an overview of the IFCap model's architecture and workflow.

3.1 Image-like Retrieval

To bridge the modality gap between text and images, we employ Image-like Retrieval, which adjusts text embeddings to better align with image embeddings. We utilize the CLIP model, which provides a shared embedding space for images and text.

Given an input text sentence t_i, we compute its embedding using the CLIP text encoder $\mathcal{E}_T$: $T_i = \mathcal{E}_T(t_i)$.

To simulate image-like representations, we inject Gaussian noise into the text embedding. Specifically, we add noise $\epsilon_r \sim \mathcal{N}(0, \sigma_r^2)$ to obtain the noise-injected text embedding:

$$T_i^\epsilon = T_i + \epsilon_r.$$

This noise-injected embedding T_i^ϵ serves as a query for retrieving relevant captions from a text corpus $\mathcal{T} = \{t_j\}_{j=1}^{N_c}$, where N_c is the number of sentences in the corpus. We compute the cosine similarity between T_i^ϵ and the embeddings of all sentences in $\mathcal{T}$:

$$\text{sim}(T_i^\epsilon, T_j) = \frac{T_i^\epsilon \cdot T_j}{\|T_i^\epsilon\| \|T_j\|}.$$

We retrieve the top-k captions $\{t_{i_1}, t_{i_2}, \ldots, t_{i_k}\}$ with the highest similarity scores. This retrieval process mimics image-to-text retrieval because the noise-injected text embedding approximates an image embedding in the shared space. Aligning text features with the image embedding space during retrieval effectively addresses the modality gap that traditional text-to-text retrieval methods overlook (as illustrated in the top part of Figure 1).

3.2 Fusion Module

To integrate the retrieved captions with the input text, we employ a Fusion Module that uses attention mechanisms to capture interactions between the two representations. The embeddings of the retrieved captions are obtained using the CLIP text encoder: $R_e = \{\mathcal{E}_T(t_{i_j})\}_{j=1}^k$.

We project the input text embedding T_i^ϵ and the retrieved embeddings R_e to match the dimensionality expected by the caption decoder. Let f_{l_1} and f_{l_2} denote linear projection layers: $Q = f_{l_1}(T_i^\epsilon), \quad K = f_{l_2}(R_e)$.

We apply a cross-attention mechanism to obtain the fused representation F_e: $F_e = \text{Attention}(Q, K, V)$, where $V = K$, and the attention function is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V.$$

Here, d_k is the dimensionality of the keys and queries. The fused representation F_e captures relevant information from both the input text and the retrieved captions. We then pass F_e through a mapping network to prepare it for the caption decoder.

3.3 Frequency-based Entity Filtering

During inference, we aim to extract relevant entities (nouns) from the retrieved captions to guide the caption generation process. We retrieve l captions related to the input image using the CLIP image encoder $\mathcal{E}_I$ to obtain the image embedding: $I_i = \mathcal{E}_I(\text{image}_i)$.

We compute the cosine similarity between I_i and the embeddings of all sentences in the corpus $\mathcal{T}$, retrieving the top-l captions. From these retrieved captions, we extract nouns using a natural language processing toolkit (e.g., NLTK [7]). We calculate the frequency f_n of each noun n across the retrieved captions.

Nouns with frequencies exceeding a predefined threshold τ are selected to form a set of entities $E = \{e_1, e_2, \ldots, e_m\}$. By focusing on frequently occurring words in the retrieved captions, we improve entity extraction accuracy without relying on a limited vocabulary. This addresses the challenge where traditional CLIP classifier-based methods struggle as

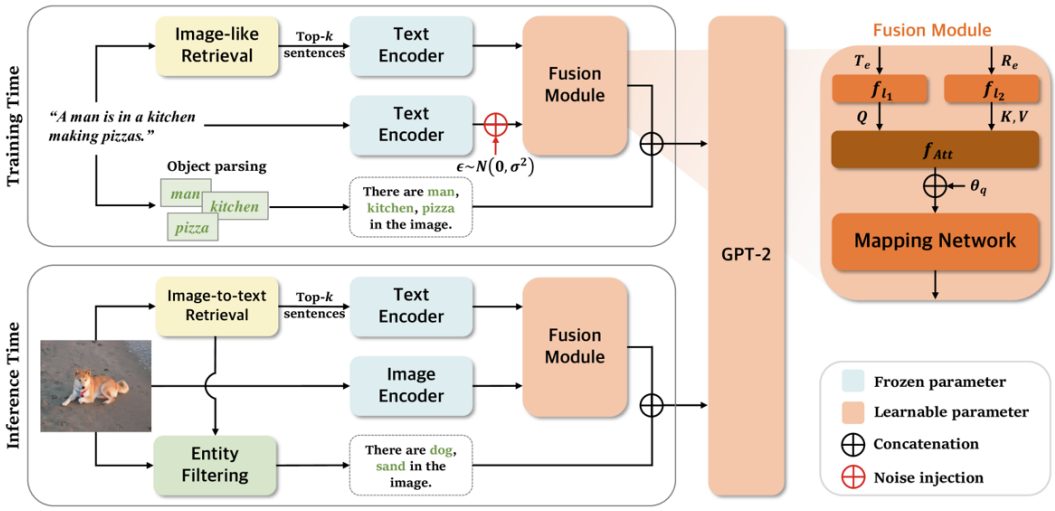

Figure 2: Overview of the IFCap model. During training, the model uses Image-like Retrieval to obtain k similar sentences and extract nouns to create a hard prompt. The input text and retrieved sentences are encoded and interact through the Fusion Module before generating captions. During inference, the model retrieves l captions similar to the input image, extract frequent entities via Frequency-based Entity Filtering to form the hard prompt, and process both image and text embeddings through the Fusion Module, following a similar procedure as in training.

vocabulary size grows (as illustrated in the bottom part of Figure 1).

3.4 Caption Generation

The extracted entities are incorporated into a hard prompt h that guides the caption decoder. For example, the prompt might be:

$$h = \text{“A photo of } e_1, e_2, \ldots, e_m.\text{”}$$

The final step involves generating the caption using a language model (e.g., GPT-2). The model takes the fused representation F_e and the hard prompt h as inputs and generates a caption y. The model is trained to minimize the autoregressive loss:

$$L = -\sum_{i=1}^{N} \log P(y_i \mid y_{<i}, F_e, h),$$

where y_i is the i-th token in the generated caption, and $y_{<i}$ represents all previous tokens.

4 Dataset

Our experiments encompass a range of datasets to thoroughly evaluate the capabilities of the IFCap model in diverse captioning scenarios, including image-based and video-based tasks, as well as zero-shot and open-domain settings.

The MS COCO [8] (Microsoft Common Objects in Context) dataset is a large-scale benchmark widely used in image captioning research. It comprises over 200,000 images depicting everyday scenes enriched with multiple objects and complex interactions. Each image is paired with five human-written captions, providing diverse linguistic expressions that capture various aspects of the scene. A common practice is to employ the Karpathy split [9], ensuring roughly 113,000 images for training and balanced subsets for validation and testing. This standardized split enables fair comparisons across different methods. Since MS COCO includes a broad range of objects and activities, models evaluated on it can demonstrate robustness and versatility in describing intricate and contextually rich scenes.

Flickr30k [10] is another frequently used image captioning dataset, consisting of 31,783 images sourced from Flickr. Each image is annotated with five human-generated captions. Although smaller in scale than MS COCO, Flickr30k remains a valuable benchmark due to its simpler, more homogenous domain. It allows for rapid experimentation and verification of model performance, particularly when assessing how well improvements generalize to datasets beyond MS COCO. The Karpathy split is also commonly adopted for Flickr30k, providing a consistent train/val/test partition that facilitates direct comparisons across various captioning approaches.

To further examine zero-shot capabilities, we employ the NoCaps [11] dataset. Constructed from Open Images, NoCaps challenges models to caption images containing objects and concepts that may not appear during training. It partitions the evaluation images into in-domain, near-domain, and out-of-domain subsets, thereby progressively increasing the difficulty and testing the models ability to handle novel visual concepts. Performance on NoCaps provides insights into how well models generalize beyond the closed sets of objects and scenes seen in MS COCO and Flickr30k, making it a rigorous test of open-domain captioning.

Moving beyond static images, we evaluate on MSR-VTT [12], a large-scale video captioning dataset widely used

to benchmark models in understanding and describing dynamic visual content. MSR-VTT contains approximately 10,000 short video clips sourced from the web, each annotated with 20 human-written captions. These captions describe various activities, events, and objects appearing in the videos. Evaluating on MSR-VTT tests the models ability to track changes over time, capture temporal relationships, and handle more complex narratives that unfold across multiple frames. Strong performance here indicates the models adaptability and robustness in extending its zero-shot captioning capabilities from static imagery to temporal sequences.

MSVD [13] (Microsoft Video Description) is another prominent video captioning dataset, often referred to as YouTube2Text. It includes roughly 2,000 short video clips collected from YouTube, each accompanied by a large number of English sentences describing the actions, objects, and events in the video. With around 40 captions per video, MSVD provides a richly annotated resource, allowing models to learn from diverse linguistic expressions and gain exposure to a broad range of scenarios, albeit on a smaller scale compared to MSR-VTT. Evaluating on MSVD helps confirm the models adaptability and ensures that improvements observed on one video dataset (MSR-VTT) translate to other video captioning domains as well.

The IFCap paper does not specify unique preprocessing techniques beyond standard practices commonly used in the captioning community. For images, these steps typically involve resizing and normalizing the inputs according to the requirements of the CLIP image encoder. For text, captions are generally tokenized using the GPT-2 tokenizer, converting them into subword tokens suitable for the language model. Similar preprocessing steps apply to video-based evaluations, where a set of frames are sampled from each video and processed with the image encoder. Basic text normalization procedures such as lowercasing, punctuation trimming, or minor text cleaning may be applied. Without any specialized mention of dataset-specific filters or vocabulary constraints, it is reasonable to assume that the same standard, widely adopted preprocessing methods are utilized for MS COCO, Flickr30k, NoCaps, MSR-VTT, and MSVD.

5 Results

This section presents a thorough evaluation of our proposed IFCap model. We begin by detailing the implementation specifics used in our experiments, then describe the evaluation metrics employed to measure success. Following that, we compare IFCaps performance against strong baselines across various datasets and settings. We include in-domain evaluations (MS COCO and Flickr30k), cross-domain generalization, open-domain testing on NoCaps, video captioning on MSR-VTT and MSVD, and a comprehensive set of ablation studies to understand the contribution of each component.

5.1 Implementation Details

In our implementation, we utilize CLIP (ViT-B/32) as the image encoder and GPT-2_{Base} as the text decoder. The parameters of the image encoder are kept frozen during training to preserve pre-trained visual representations, while the text decoder and Fusion Module are optimized for the captioning task. We train the model for t epochs with a learning rate of 2×10^{-5}, using the AdamW optimizer alongside a learning rate scheduler. The batch size is set to 80 for a balanced trade-off between computational efficiency and resource utilization.

All experiments were conducted on a single NVIDIA A6000 GPU with 48 GB of VRAM, taking approximately 30 miniutes per training session and utilizing about 40 GB of VRAM. In the Image-like Retrieval component, we determine an optimal noise level $\sigma_r = 0.04$ through preliminary experiments, ensuring that noise injection into text embeddings effectively simulates image-like representations. The Fusion Module projects both the noise-injected text embeddings and retrieved captions into the GPT-2 embedding space, integrating them via a cross-attention mechanism before processing through an eight-layer Transformer mapping network.

During inference, the Frequency-based Entity Filtering retrieves l captions related to the input image, extracts nouns to form a frequency distribution, and applies a predefined threshold τ to construct a hard prompt $\mathbf{h}$. This method ensures accurate, diverse entity inclusion without relying on a limited vocabulary. Together, these steps form the backbone of IFCaps pipeline, setting the stage for the evaluations presented below. Table 1 presents the hyperparameter values across datasets, including the number of training epochs (t), retrieved captions (l), and threshold (τ).

Table 1: Hyperparameter table.

HyperParameters	COCO	Flickr30k	NoCaps	MSVD	MSR-VTT
Epochs (t)	5	30	-	10	10
l	9	7	7	7	7
τ	5	3	3	5	6

5.2 Evaluation Metrics

To rigorously assess caption quality, we employ four widely accepted metrics: BLEU@4 (Bilingual Evaluation Understudy) [14], METEOR [15], CIDEr (Consensus-based Image Description Evaluation) [16], and SPICE (Semantic Propositional Image Caption Evaluation) [17]. Each metric targets different aspects of caption evaluation, allowing us to measure success along multiple dimensions.

BLEU measures n-gram precision between generated and reference captions. For BLEU@4, a brevity penalty (BP) discourages overly short captions:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{4} w_n \log p_n\right),$$

Table 2: Result on the In-domain captioning including COCO test split and Flickr30k test split. Every result is copied from the original papers. ♠: Utilizes text-to-image generation model in the training time, †: Utilizes object detector during the training and inference time. IFCap achieves state-of-the-art in most metrics. The best number overall is in bold and the second best is underlined.

Method	Image Encoder	Text Decoder	COCO				Flickr30k			
			B@4	M	C	S	B@4	M	C	S
CapDec (2022)	RN50x4	$\text{GPT-2}_{\text{Large}}$	26.4	25.1	91.8	11.9	17.7	20.0	39.1	9.9
DeCap (2023)	ViT-B/32	$\text{Transformer}_{\text{Base}}$	24.7	25.0	91.2	18.7	21.2	21.8	56.7	15.2
CLOSE (2022)	ViT-L/14	T5_{Base}	-	-	95.3	-	-	-	-	-
ViECap (2023)	ViT-B/32	$\text{GPT-2}_{\text{Base}}$	27.2	24.8	92.9	18.2	21.4	20.1	47.9	13.6
$\text{MeaCap}_{\text{InvLM}}$ (2024)	ViT-B/32	$\text{GPT-2}_{\text{Base}}$	27.2	25.3	95.4	19.0	22.3	22.3	59.4	15.6
Knight (2023)	RN50x64	$\text{GPT-2}_{\text{Large}}$	27.8	<u>26.4</u>	98.9	<u>19.6</u>	22.6	**24.0**	56.3	16.3
ICSD♠ (2023)	ViT-B/32	$\text{BERT}_{\text{Base}}$	<u>29.9</u>	25.4	96.6	-	**25.2**	20.6	54.3	-
SynTIC♠† (2023)	ViT-B/32	$\text{Transformer}^{L=4}_{H=4}$	<u>29.9</u>	25.8	<u>101.1</u>	19.3	22.3	22.4	56.6	<u>16.6</u>
IFCap	ViT-B/32	$\text{GPT-2}_{\text{Base}}$	**30.8**	**26.7**	**108.0**	**20.3**	<u>23.5</u>	<u>23.0</u>	**64.4**	**17.0**

where p_n is the n-gram precision and w_n are uniform weights. $\text{BP} = \min(1, e^{1-\frac{r}{c}})$ with r the reference length sum and c the candidate length sum.

METEOR aligns candidate and reference words, considering synonyms and stemming. It computes a harmonic mean of precision and recall with a higher weight for recall, penalizing fragmented matches:

$$F_{\text{mean}} = \frac{PR}{\alpha P + (1-\alpha)R},$$
$$\text{METEOR} = F_{\text{mean}} \cdot (1 - P_{\text{frag}}\gamma).$$

CIDEr uses TF-IDF weighted n-grams to measure how well a candidate matches a set of references:

$$\text{CIDEr}(C, \{R_i\}) = \frac{1}{m}\sum_{i=1}^{m} \frac{\sum_g \text{tf}_g(C)\text{tf}_g(R_i)\text{idf}_g^2}{\sqrt{\sum_g (\text{tf}_g(C)\text{idf}_g)^2}\sqrt{\sum_g (\text{tf}_g(R_i)\text{idf}_g)^2}}.$$

SPICE focuses on semantic content by comparing tuples (objects, attributes, relations) derived from scene graphs of candidate and reference captions:

$$\text{SPICE}(C, \{R_i\}) = \frac{2|T(C) \cap T(R)|}{|T(C)| + |T(R)|}.$$

These metrics collectively measure lexical overlap (BLEU, METEOR), semantic richness and consensus (CIDEr), and conceptual structure (SPICE). We define success as consistently improving upon previous methods across these metrics, reflecting both linguistic and semantic enhancements.

5.3 State-of-the-Art Method Identification and Comparison

Zero-shot image captioning poses unique challenges: generating captions for images never paired with text during training, adapting to domain shifts, handling unseen objects, and even describing video frames. After reviewing the literature, IFCap emerges as a leading SOTA approach, surpassing previous methods that rely on large paired datasets, object detectors, or text-to-image generation models. IFCap effectively integrates Image-like Retrieval, a Fusion Module, and Entity Filtering to address these complexities.

We compare our approach to a range of text-only captioning models. CapDec [1] and ViECap [2] both build on Clipcap [18], injecting predefined Gaussian noise into text embeddings to simulate image-like features. While CapDec leverages this noise-injection strategy directly to approximate visual distributions, ViECap augments it with explicit entity extraction to guide the captioning process. CLOSE [4] also employs noise for alignment but explores multiple noise configurations, allowing flexible hyperparameter adjustments that can improve robustness. DeCap [19], in contrast, incorporates a memory bank to store and dynamically retrieve textual cues, enabling the model to manage unseen concepts more adaptively. Knight [20] relies solely on retrieval-based text features without explicit image modeling, whereas MeaCap [21] refines these retrieved sentences into structured Subject-Predicate-Object triplets, aiming for a more semantically informed textual representation. Finally, ICSD [22] and SynTIC [23] employ text-to-image generation tools like Stable Diffusion [24] to produce synthetic scenes that serve as visual anchors, bridging the modality gap by offering pseudo-visual training stimuli.

We provide results across several benchmarks and settings, including MS COCO, Flickr30k, NoCaps, MSR-VTT, and MSVD. All comparisons employ commonly adopted splits (e.g., Karpathy splits for MS COCO and Flickr30k) and the standard evaluation metrics described above to ensure fairness and reliability in measuring performance.

5.3.1 In-domain Captioning

In-domain evaluations test the model on distributions similar to what it was trained on. Tables 2 shows IFCaps performance on MS COCO and Flickr30k, comparing to CapDec, DeCap, CLOSE, ViECap, $\text{MeaCap}_{\text{InvLM}}$, Knight, ICSD, and SynTIC. IFCap substantially outperforms previous methods in both BLEU@4 and METEOR, and shows the greatest gains in CIDEr and SPICE, reflecting enhanced

semantic alignment and capturing greater consensus with human annotators. Success here is defined by surpassing previous best methods under the same conditions, and IFCaps results confirm it as the new in-domain SOTA.

5.3.2 Cross-domain Captioning

We next test generalization across domains by training on COCO and testing on Flickr30k, and vice versa. Table 3 reports the cross-domain results, including variations where models can access the target domains corpus during inference time (denoted as -TT).

Table 3: Results on the Cross-domain captioning. $-TT$: models can access the target domains corpus during inference time. $\star$: without Entity Filtering module in the inference time. IFCap achieves state-of-the-art in most metrics.

Method	COCO → Flickr				Flickr → COCO			
	B@4	M	C	S	B@4	M	C	S
DeCap (2023)	16.3	17.9	35.7	11.1	12.1	18.0	44.4	10.9
ViECap (2023)	17.4	18.0	38.4	11.2	12.6	19.3	54.2	12.5
Knight (2023)	21.1	**22.0**	48.9	14.2	19.0	22.8	64.4	15.1
SynTIC (2023)	17.9	18.6	38.4	11.9	14.6	19.4	47.0	11.9
SynTIC$-TT$	19.4	20.2	43.2	13.9	**20.6**	21.3	64.4	14.3
IFCap*	17.8	19.4	47.5	12.7	14.7	20.4	60.7	13.6
IFCap$-TT$	**21.2**	21.8	**59.2**	**15.6**	19.0	**23.0**	**76.3**	**17.3**

Under cross-domain evaluation, IFCap-TT achieves the highest CIDEr and SPICE, demonstrating not just lexical fidelity but also semantic adaptability to unfamiliar domains. Success here involves maintaining or enhancing performance despite domain shifts, and IFCaps results confirm its strong generalization capabilities.

5.3.3 Open-domain Evaluation on NoCaps

To test zero-shot reasoning with novel objects, we evaluate IFCap on NoCaps. Table 4 shows IFCaps performance against DeCap, CapDec, and ViECap.

Table 4: Results on the NoCaps validation split. $\star$: without Entity Filtering module in the inference time. IFCap achieves state-of-the-art in every metric.

Method	COCO ⇒ NoCaps Val							
	In		Near		Out		Entire	
	C	S	C	S	C	S	C	S
DeCap (2023)	65.2	-	47.8	-	25.8	-	45.9	-
CapDec (2022)	60.1	10.2	50.2	9.3	28.7	6.0	45.9	8.3
ViECap (2023)	61.1	10.4	64.3	9.9	65.0	8.6	66.2	9.5
IFCap*	**70.1**	**11.2**	**72.5**	**10.9**	**72.1**	**9.6**	**74.0**	**10.5**

IFCap leads in every category, showing that it can caption images containing unseen objects with remarkable semantic fidelity. Success in NoCaps is measured by the models capacity to generalize beyond trained domains, and IFCaps strong performance attests to the effectiveness of Image-like Retrieval and EF in truly zero-shot conditions.

5.3.4 Video Captioning

To gauge versatility across modalities, we apply IFCap to video captioning on MSR-VTT and MSVD. Table 5 compares IFCap with ZeroCap, MAGIC [25], CLMs [26], CapDec, EPT [27], and Knight.

Table 5: Results on the Video captioning including MSR-VTT and MSVD. IFCap achieves state-of-the-art in most metrics.

Method	MSR-VTT				MSVD			
	B@4	M	C	S	B@4	M	C	S
ZeroCap (2022b)	2.3	12.9	5.8	-	2.9	16.3	9.6	-
MAGIC (2022)	5.5	13.3	7.4	4.2	6.6	16.1	14.0	2.9
CLMs (2022)	6.2	17.8	10.1	6.5	7.0	16.4	20.0	3.1
CapDec (2022)	8.9	23.7	11.5	5.9	7.9	23.3	34.5	3.2
EPT (2022a)	3.0	14.6	11.3	-	3.0	17.8	17.4	-
Knight (2023)	25.4	**28.0**	31.9	**8.5**	37.7	**36.1**	63.8	5.0
IFCap	**27.1**	25.9	**38.9**	6.7	**40.6**	34.2	**83.9**	**6.3**

IFCap sets a new SOTA on MSR-VTT and MSVD for most metrics, demonstrating that its retrieval-based, entity-filtering approach extends beyond static images to temporal domains. Here, success means preserving semantic depth and coherence in multi-frame video scenarios, and IFCaps strong performance fulfills this criterion.

5.3.5 Ablation Studies

Ablation studies dissect IFCaps components (Image-like Retrieval (ILR), Fusion Module (FM), and Frequency-based Entity Filtering (EF)) and various design choices. Table 6 summarizes the impact of removing each component on COCO performance.

Table 6: Ablation studies of the key components of IFCap on COCO. Removing any component causes performance degradation.

Image-like Retrieval	Fusion Module	Entity Filtering	COCO			
			B@4	M	C	S
✓	✓	✓	**30.8**	**26.7**	**108.0**	**20.3**
	✓	✓	27.2	24.8	92.9	18.2
✓		✓	28.5	26.0	102.0	20.0
✓	✓		29.2	26.0	104.0	19.9

Removing ILR, FM, or EF reduces CIDEr and SPICE, confirming that each part is essential. Additional ablations (not all shown here using the table) vary noise injection timing, the number of retrieved sentences for the Fusion Module and EF, the heuristic or adaptive threshold for EF, and the number of Transformer and cross-attention layers. Using COCO, each experiment guides optimal hyperparameter choices and verifies that design decisions (e.g., $\sigma_r = 0.04$, $k = 5$ retrieved sentences for FM, $l = 9$ retrieved sentences for EF, $\tau = 5$ threshold) maximize performance. Referencing the tables shown in the origin IFCap paper (e.g., Tables for noise injection timing, threshold selection, etc.) consistently reveals that any deviation from the chosen configuration leads to decreased scores.

Success in these ablation studies is measured by maintaining or improving upon IFCaps full configuration performance. The ablations demonstrate that ILR, FM, and EF collectively enable IFCap to achieve its SOTA results, and fine-tuning thresholds or the number of retrieved sentences further refines caption quality.

5.3.6 Replication of Released Code and Results

To further validate the robustness and reproducibility of IFCap, we replicated the experiments in the cross-domain captioning scenario (COCO → Flickr30k and Flickr30k → COCO) using the officially released code and model checkpoints. We report three sets of results: (1) the original IFCap-TT performance as reported in the paper, (2) results obtained by running inference directly on the authors released trained model weights without retraining, and (3) results obtained after training and testing the model from scratch using the provided source code.

For the experiments in scenario (2), we used the authors provided checkpoints (`coco-indomain-005.pt` and `flickr-indomain-0014.pt`) to perform inference. These weights were already fine-tuned by the original researchers. In scenario (3), we trained on COCO for roughly 1.5 hours and completed 5 epochs to produce a new checkpoint (`coco-004.pt`). For Flickr30k, training lasted about 1.5 hours and reached early stopping at 13 epochs, yielding the checkpoint `flickr-0012.pt`. All training configurations, including optimizer settings and batch sizes, matched those recommended by the original code release.

Table 7: Cross-domain captioning results comparing original reported IFCap−*TT* performance, released model weights tested directly, and our own training and testing efforts.

Method	COCO → Flickr				Flickr → COCO			
	B@4	M	C	S	B@4	M	C	S
(Base) Knight (2023)	21.1	**22.0**	48.9	14.2	19.0	22.8	64.4	15.1
(Base) SynTIC−*TT*	19.4	20.2	43.2	13.9	**20.6**	21.3	64.4	14.3
(1) IFCap−*TT* (Paper)	**21.2**	21.8	**59.2**	**15.6**	19.0	**23.0**	**76.3**	**17.3**
(2) Released Weights Only	20.8	21.9	59.0	15.8	19.1	23.5	78.6	18.0
(3) Train+Test provided source	20.9	21.2	55.8	14.9	18.4	23.2	75.9	17.8

Table 7 shows the comparison among the three scenarios. The results from the released model weights (scenario 2) closely resemble the original reported scores (scenario 1), with minor metric variations that can arise from environmental differences, such as slightly different library versions or floating-point rounding on the hardware used.

For the model fully retrained and tested from scratch (scenario 3), performance is generally within a close range of the original, though some metrics, such as the CIDEr score, are slightly lower than reported. Such discrepancies are not unusual and may stem from random initialization seeds, minor hardware differences, or unreported hyper-parameter optimizations that the original authors applied. Despite these minor deviations, the model still maintains competitive performance levels and consistently outperforms previously published baselines. This indicates that the main conclusions of the original study remain valid. The replication results confirm IFCaps robustness and reinforce its position as a leading method for zero-shot image captioning, even when re-implemented and re-trained by an external party.

6 Possible Improvements and Results

While the original IFCap implementation carefully tuned various hyperparameters to achieve optimal performance, we explored a different avenue for potential improvement: modifying the model architecture. Instead of adjusting parameters that were previously shown to be near-optimal, we replaced the original GPT-2_{Base} decoder with a more capable GPT-2_{Medium} model.

The key distinction between GPT-2_{Base} and GPT-2_{Medium} lies in their size and capacity. GPT-2_{Base} (with roughly 124M parameters) is smaller and shallower, whereas GPT-2_{Medium} (around 345M parameters) has more layers and substantially more parameters. This added capacity allows GPT-2_{Medium} to represent more complex linguistic patterns and subtle semantic relationships, potentially leading to richer and more contextually accurate captions. Implementing this change required only a slight modification of the shell script and ensuring that both training and inference stages explicitly utilize GPT-2_{Medium} rather than GPT-2_{Base}.

Table 8: Cross-domain captioning performance with a GPT-2_{Medium} decoder compared against the original IFCap-TT results (Paper) and our own replication using GPT-2_{Base}. Train+Test provided source corresponds to our baseline replication from scratch, while Experiment (Ours) reflects the outcome of substituting GPT-2_{Base} with GPT-2_{Medium}.

Method	COCO → Flickr				Flickr → COCO			
	B@4	M	C	S	B@4	M	C	S
IFCap−*TT* (Paper)	21.2	**21.8**	**59.2**	**15.6**	19.0	23.0	76.3	17.3
(3) Train+Test provided source	20.9	21.2	55.8	14.9	18.4	23.2	75.9	17.8
Experiment (Ours, GPT-2_{Medium})	**21.5**	21.6	58.4	15.4	**19.5**	**23.8**	**76.8**	**18.2**

Table 8 shows the results of this architectural adjustment, focusing on the cross-domain captioning scenario. Our experiment is compared to both the originally reported IFCap-TT performance (Paper) and our previously replicated training and testing efforts.

While the adjusted model did not surpass the original papers reported metrics on COCO → Flickr across all measures, it did achieve a slight improvement in BLEU@4. More notably, on the Flickr → COCO evaluation, the GPT-2_{Medium}-based IFCap outperformed the original papers results in every metric. Compared to our own baseline replication with GPT-2_{Base}, the GPT-2_{Medium} variant consistently improved all metrics across both directions of the cross-domain task.

These findings suggest that increasing the capacity of the underlying language model can enhance performance, particularly in scenarios demanding robust generalization. The richer representational power of GPT-2_{Medium} likely enables

the model to better understand complex scene semantics and produce captions that align more closely with human references. Although the improvements are not uniform across all conditions, this architectural adjustment demonstrates a promising direction for further enhancing zero-shot image captioning models without the need for extensive hyperparameter tuning.

7 Code Repository

The code is available at `https://github.com/flametom/PSU_CSE597_003_CoursePJ`.

This repository provides the source code, scripts, and instructions for installing dependencies, preparing data, training the model, and running evaluations. It also contains a README file describing the functionality of each component and how to reproduce the experimental results.

References

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

[2] J. Fei, T. Wang, J. Zhang, Z. He, C. Wang, and F. Zheng, "Transferable decoding with visual entities for zero-shot image captioning," *arXiv preprint arXiv:2307.16525*, 2023.

[3] J. Wang, M. Yan, Y. Zhang, and J. Sang, "From association to generation: Text-only captioning by unsupervised cross-modal mapping," Aug. 2023, pp. 4326–4334.

[4] S. Gu, C. Clark, and A. Kembhavi, "I cant believe theres no images! : Learning visual tasks using only language supervision," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2672–2683, 2022.

[5] S. Lee, S.-W. Kim, T. Kim, and D.-J. Kim, "IFCap: Image-like retrieval and frequency-based entity filtering for zero-shot captioning," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 20 715–20 727.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021.

[7] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, 1st. O'Reilly Media, Inc., 2009, ISBN: 0596516495.

[8] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *ArXiv*, vol. abs/1504.00325, 2015.

[9] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3128–3137, 2014.

[10] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014. DOI: `10.1162/tacl_a_00166`.

[11] 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 *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 8947–8956. DOI: `10.1109/ICCV.2019.00904`.

[12] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5288–5296. DOI: 10.1109/CVPR.2016.571.

[13] Z. Wu, T. Yao, Y. Fu, and Y.-G. Jiang, "Deep learning for video classification and captioning," in *Frontiers of Multimedia Research*, 2016.

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02, Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002, pp. 311–318.

[15] S. Banerjee and A. Lavie, "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, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72.

[16] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.

[17] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Computer Vision – ECCV 2016*, Cham: Springer International Publishing, 2016, pp. 382–398, ISBN: 978-3-319-46454-1.

[18] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," *arXiv preprint arXiv:2111.09734*, 2021.

[19] W. Li, L. Zhu, L. Wen, and Y. Yang, "Decap: Decoding CLIP latents for zero-shot captioning via text-only training," in *The Eleventh International Conference on Learning Representations*, 2023.

[20] J. Wang, M. Yan, Y. Zhang, and J. Sang, "From association to generation: Text-only captioning by unsupervised cross-modal mapping," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, E. Elkind, Ed., Main Track, International Joint Conferences on Artificial Intelligence Organization, Aug. 2023, pp. 4326–4334.

[21] Z. Zeng, Y. Xie, H. Zhang, C. Chen, Z. Wang, and B. Chen, "Meacap: Memory-augmented zero-shot image captioning," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14 100–14 110, 2024.

[22] F. Ma, Y. Zhou, F. Rao, Y. Zhang, and X. Sun, "Image captioning with multi-context synthetic data," in *AAAI Conference on Artificial Intelligence*, 2023.

[23] Z. Liu, J. Liu, and F. Ma, "Improving cross-modal alignment with synthetic pairs for text-only image captioning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 3864–3872, Mar. 2024.

[24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 674–10 685, 2021.

[25] Y. Su, T. Lan, Y. Liu, F. Liu, D. Yogatama, Y. Wang, L. Kong, and N. Collier, *Language models can see: Plugging visual controls in text generation*, 2022. arXiv: 2205.02655 [cs.CV].

[26] J. Wang, Y. Zhang, M. Yan, J. Zhang, and J. Sang, *Zero-shot image captioning by anchor-augmented vision-language space alignment*, 2022. arXiv: 2211.07275 [cs.CV].

[27] Y. Tewel, Y. Shalev, R. Nadler, I. Schwartz, and L. Wolf, "Zero-shot video captioning with evolving pseudo-tokens," *ArXiv*, vol. abs/2207.11100, 2022.