Negative Entity Suppression for Zero-Shot Captioning with Synthetic Images

Zimao Lu[1], Hui Xu[1,2*], Bing Liu[1,2*], Ke Wang[1,2]

[1]School of Computer Science and Technology/ School of Artificial Intelligence,China University of Mining and Technology, Jiangsu, China

[2]Mine Digitization Engineering Research Center of the Ministry of Education

{luzimao, xuhui, liubing, wangke}@cumt.edu.cn

Abstract

Text-only training provides an attractive approach to address data scarcity challenges in zero-shot image captioning (ZIC), avoiding the expense of collecting paired image-text annotations. However, although these approaches perform well within training domains, they suffer from poor cross-domain generalization, often producing hallucinated content when encountering novel visual environments. Retrieval-based methods attempt to mitigate this limitation by leveraging external knowledge, but they can paradoxically exacerbate hallucination when retrieved captions contain entities irrelevant to the inputs. We introduce the concept of ***negative entities***—objects that appear in generated caption but are absent from the input—and propose Negative Entity Suppression (NES) to tackle this challenge. NES seamlessly integrates three stages: (1) it employs synthetic images to ensure consistent image-to-text retrieval across both training and inference; (2) it filters negative entities from retrieved content to enhance accuracy; and (3) it applies attention-level suppression using identified negative entities to further minimize the impact of hallucination-prone features. Evaluation across multiple benchmarks demonstrates that NES maintains competitive in-domain performance while improving cross-domain transfer and reducing hallucination rates, achieving new state-of-the-art results in ZIC.

Introduction

Image captioning bridges computer vision and natural language processing by generating textual descriptions for visual content using encoder-decoder architectures. While existing approaches achieve impressive results in supervised settings, they rely heavily on large-scale paired text-image annotations, limiting their applicability to out-of-distribution images containing unfamiliar objects. This limitation motivates Zero-shot Image Captioning (ZIC), which aims to generate captions without paired training data, reducing annotation costs and enabling deployment across previously unseen domains. However, as shown in Fig. 1, current ZIC methods face a significant challenge: despite strong in-domain performance, they suffer substantial degradation when applied to cross-domain tasks.

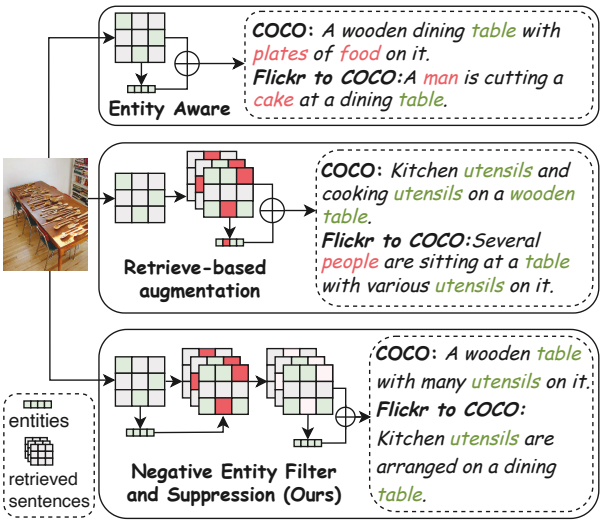

Figure 1: Illustration of cross-domain degradation in ZIC models. COCO: training and inference on COCO dataset; Flickr to COCO: training on Flickr30k and inference on COCO. Correct and incorrect entities are marked in green and red, respectively. Entity-aware models, lacking image-text pairs, tend to generate hallucinated entities with high semantic association to existing entities. Retrieval-based models achieve better in-domain performance but introduce new hallucinations that reduce cross-domain generalization. Our model incorporates identification and suppression modules to effectively mitigate the impact of hallucinated information in retrieved content.

Two interconnected factors contribute to cross-domain performance degradation. First, the **cross-modal gap**: lacking high-quality visual information during training, models fail to develop robust visual priors and instead over-rely on linguistic patterns learned from text. Second, **cross-domain hallucination**: when encountering out-of-domain scenarios, models tend to generate objects that are highly associated with existing content but not actually present in the images. These hallucinated entities typically either share visual similarities with actual image content or co-occur with them in training corpora frequently.

*Corresponding Author

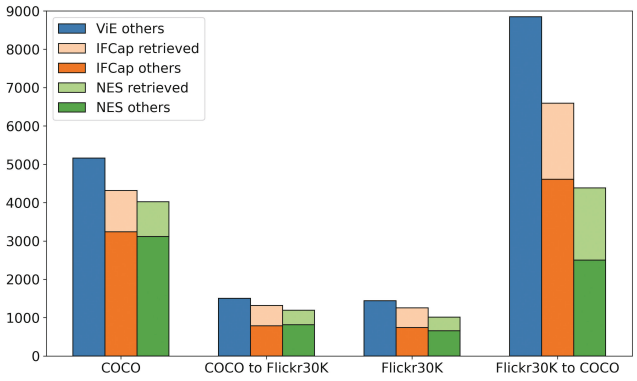

Figure 2: Analysis of hallucination patterns in existing ZIC methods. Hallucinated entities are classified as *retrieved* (originating from retrieved captions) or *others* across four scenarios. The results demonstrate that retrieval-based approaches such as IFCap suffer from the risk of hallucinations, while our NES method effectively reduces both the overall number of hallucinated entities and those specifically induced by retrieval.

Existing approaches have attempted to address these challenges from various perspectives, such as entity extraction(Junjie Fei et al. 2023; Soeun Lee et al. 2024), data synthesis(Liu, Liu, and Ma 2024; Ma et al. 2024) and external knowledge enhancement(Ramos, Elliott, and Martins 2023; Kim et al. 2025). However, most methods treat the modality gap and hallucination issues as separate problems, overlooking their intrinsic connection. This motivates a unified framework that can simultaneously mitigate modality gap and suppress hallucinated entities, thereby improving cross-domain generalization in ZIC.

To investigate the underlying causes of cross-domain performance degradation, we analyze hallucination phenomena in existing models. As illustrated in Figure 2, our analysis reveals that while IFCap achieves a lower overall hallucination rate compared to ViE, approximately 20% of the hallucinated entities are present in the retrieved information(Lu et al. 2025). This finding suggests that identifying and filtering out these entities from the retrieval corpus could potentially improve the model's generation performance. We validate this hypothesis through our proposed NES model, which achieves a 30% reduction in total hallucinations and a 10% decrease in retrieval-induced hallucinations on the Flickr30k-to-COCO task.

To optimize retrieval-assisted generation while reducing hallucinations, we analyze of the relationships between different data sources, encoders, and retrieval accuracy across the COCO dataset using multiple CLIP encoders. As illustrated in Figure 3, our findings demonstrate that synthetic images achieve superior balance across accuracy and recall metrics while reducing hallucination compared to text-based retrieval, and the RN50 model achieves the lowest hallucination rate while maintaining accuracy and recall on par with other models.

Motivated by these findings and existing approaches, we propose the **Negative Entity Suppression (NES)**, a framework that simultaneously addresses modality gap and hallucination challenges through synthetic image-to-text processing and hallucination-aware filtering. Rather than treating these issues separately, NES leverages synthetic images to both enhance text-only training inputs and enable consistent image-to-text retrieval across both training and inference phases. For retrieved entities, we categorize them into positive entities (present in the input) and negative entities (absent from the input), then apply targeted attention-level suppression to reduce the interference of hallucination-prone information in retrieved content.

In summary, our contributions are as follows:

- **Synthetic Image-to-text Retrieval:** We employ synthetic images generated by diffusion models for both input enhancement and retrieval queries, maintaining consistency between training and inference while reducing modality gap inherent in text-only retrieval approaches.
- **Negative Entity Identification and Filtering:** We develop an entity verification framework that distinguishes between visually grounded and ungrounded entities in retrieved content, utilizing ground-truth captions during training and visual-semantic alignment during inference.
- **Attention-level Hallucination Suppression:** We design a targeted attention mechanism that identifies and suppresses negative entity influences within retrieved caption features, minimizing their impact on feature fusion and overall caption quality.
- **Superior Cross-domain Performance:** We demonstrate that NES maintains in-domain performance while significantly improving cross-domain generalization, achieving a 14% improvement in CIDEr and reducing hallucination rates by 54% on the Flickr30k-to-COCO task.

Related Work

Text-only Training Image Captioning

Text-only training approaches aim to enhance the mapping from vision to language in the absence of image–text alignment. A pioneering method, ClipCap (Ron Mokady, Amir Hertz, and Amit H. Bermano 2021), uses a mapping network to transform visual features into semantic prefixes for language models. Building on this, ViECap (Junjie Fei et al. 2023) introduces an entity extraction module that serves as hard prompts for the decoder, mitigating hallucinations caused by overreliance on pretrained knowledge. IFCap (Soeun Lee et al. 2024) incorporates a retrieval module to replace image-derived entities with those from retrieved captions for better accuracy. MERCap (Zeng et al. 2025) constructs an image–entity memory bank, retrieving relevant entities during inference to complement extracted entities for caption generation. Diffusion Bridge (Lee et al. 2025) leverages diffusion models trained exclusively on text embeddings to progressively align vision embeddings with the text embedding distribution through a reverse diffusion process, effectively reducing the modality gap in CLIP for improved ZIC.

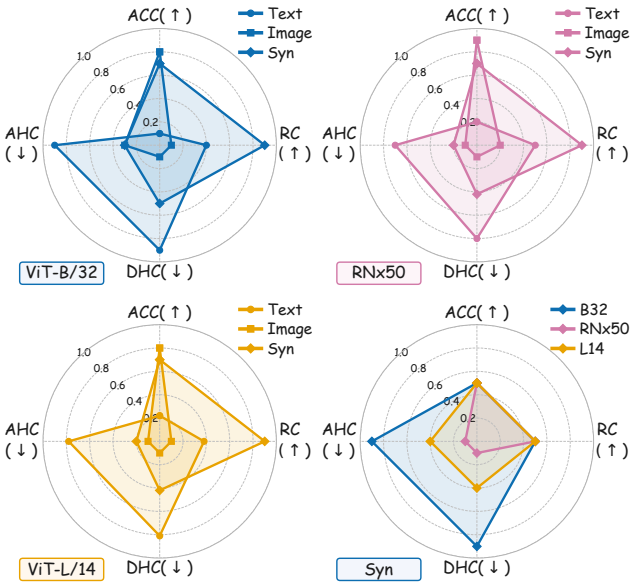

Figure 3: Analysis of retrieval performance across different data sources and CLIP encoders on the COCO dataset (normalized). Higher values indicate better performance for accuracy (ACC) and recall (RC), while lower values indicate better performance for average hallucination count per image (AHC) and deduplicated hallucination count (DHC).

Retrieval-based Image Captioning

Retrieval-based methods in image captioning typically fall into image-to-text and text-to-text retrieval. The former retrieves captions semantically aligned with the input image, while the latter finds similar captions based on a query sentence. These techniques have been widely adopted to enhance captioning quality and transferability. For example, (Ramos, Elliott, and Martins 2023) proposes an end-to-end model that jointly embeds images and retrieved captions, while (Rita Ramos et al. 2023) designs a lightweight model that trains only cross-attention layers for better generalization across domains.

In ZIC, retrieval modules are commonly designed as plug-and-play components for domain adaptation. (Zequn Zeng et al. 2024) employs image-to-text retrieval from a large-scale textual memory bank through semantic similarity matching and introduces a retrieve-then-filter mechanism to extract and refine the most relevant entity prompts for caption generation; (Soeun Lee et al. 2024) employs a dual-phase retrieval strategy that performs text-to-text retrieval during training to align textual features and switches to image-to-text retrieval during inference; (Yang et al. 2023) adopts image-to-image retrieval based on visual feature similarity to retrieve the most visually analogous images from a reference database and leverages their corresponding captions as retrieval-augmented guidance for zero-shot generation.

Hallucination Suppression in Image Captioning

Hallucination has become a prominent issue in image captioning, especially in zero-shot or out-of-distribution settings. Hallucinations occur when the model generates objects, attributes, or relations not present in the image, undermining the model's reliability.

To quantify hallucination, (Anna Rohrbach et al. 2018) proposed the CHAIR metric, which identifies unmatched entities in generated captions. (Petryk et al. 2024) extended this with ALOHa to handle open-vocabulary hallucinations. With those metrics, (Sarkar, Zhang, and Liu 2025) demonstrated that imbalanced cross-modal attention distribution causes models to over-rely on language priors, and proposed suppressing non-visual attention heads. (Huang et al. 2024) found that summary tokens from self-attention mislead generation and proposed OPERA to selectively down-weight these misleading summary tokens during generation. (Leng et al. 2023) identified hallucinated entities by comparing the model's output on perturbed and original images.

Methodology

Text-only training methods face two challenges: the cross-modal gap between textual training data and visual inference inputs, and the cross-domain hallucinations when encountering out-of-domain scenarios. These challenges are particularly pronounced in zero-shot scenarios, where models must generate captions for unseen visual domains without corresponding training examples.

To address these challenges, we employ synthetic images to enable image-to-text retrieval during training and suppress hallucination-prone features using negative entities filtered from retrieved captions. Our framework integrates three components: (1) **Image-to-text Retrieval** for modality consistency, (2) **Negative Entity Filtering** to improve generation accuracy, and (3) **Attention-level Suppression** to reduce hallucination tendency.

Image-to-text Retrieval

Text-to-text retrieval introduces hallucination information that degrades performance, as text-based queries lack visual grounding and result in retrieval of semantically similar but visually irrelevant content. As demonstrated in our analysis mentioned before, synthetic images reduce hallucination rates while maintaining retrieval quality. We therefore use synthetic images as retrieval queries to bridge the modality gap and ensure consistent image-to-text retrieval across training and inference.

Synthetic Image Generation. For ground-truth captions T, we employ a pretrained diffusion model to generate corresponding synthetic images $\tilde{I}$. These synthetic images serve as visual queries for retrieval, enabling visual-semantic alignment capture while reducing dependency on textual priors. Since these captions are not designed as generation prompts and diffusion models may introduce biases, the resulting images may not perfectly capture the original descriptions. We quantify this limitation and address it through quality-based filtering and multimodal fusion.

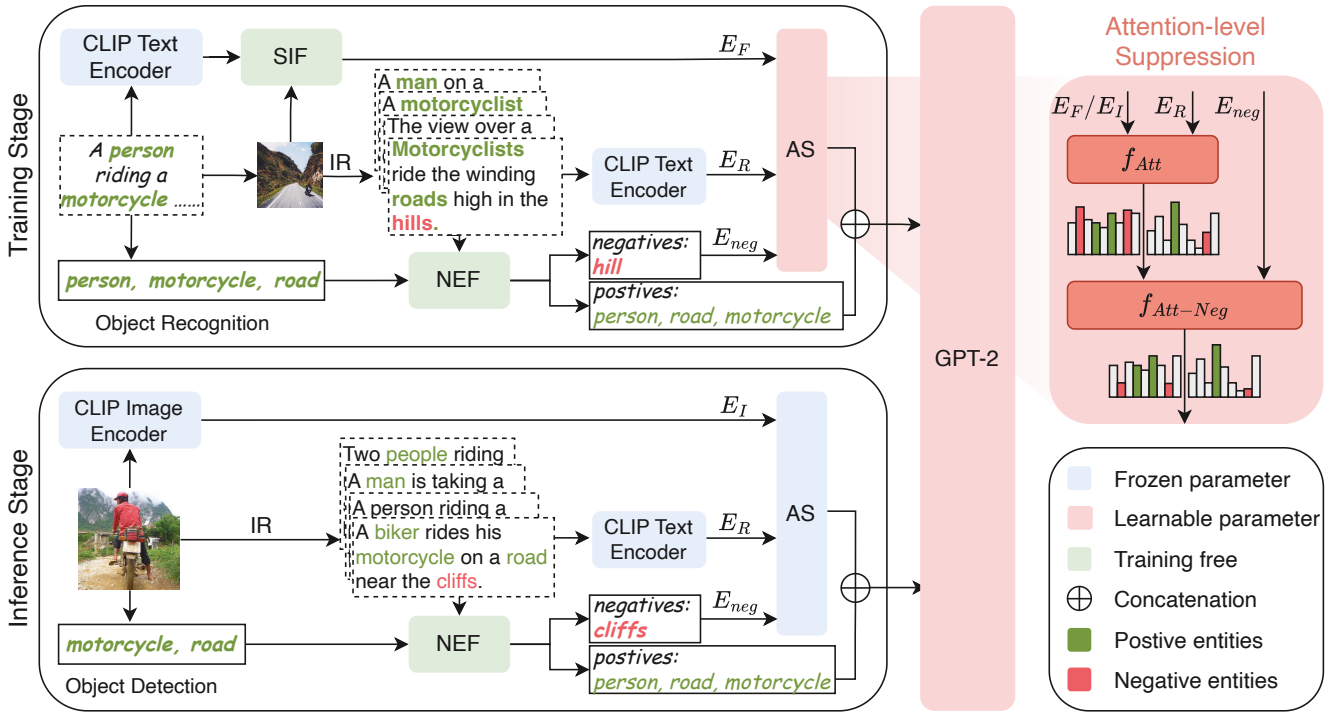

Figure 4: Overall framework of NES. (a) Training phase: The framework generates synthetic images from input text using diffusion models, then retrieves relevant captions via IR (Image-to-text Retrieval). The SIF (Synthetic Image Fusion) module enhances input text features using synthetic images. The NEF (Negative Entity Filtering) module categorizes retrieved entities into positive and negative sets based on input text entities. The AS (Attention-level Suppression) module first fuses retrieved captions with enhanced input features, then suppresses hallucination-prone features using negative entities. Final features are concatenated with positive entities and fed to GPT-2 for generation. (b) Inference phase: The pipeline directly uses input images for retrieval without SIF enhancement. NEF filtering is performed using image-extracted entities with CLIP similarity. The AS module remains consistent with training.

Synthetic Image Filtering and Fusion. To address potential quality issues with synthetic images, we compute synthetic image embedding $\mathbf{E}_{\tilde{I}} = \text{CLIP}_v(\tilde{I})$ for synthetic images $\tilde{I}$ and text embedding $\mathbf{E}_T = \text{CLIP}_t(T)$ for ground-truth captions T, then calculating the CLIPScore as:

$$\text{CLIPScore}(\mathbf{E}_{\tilde{I}}, \mathbf{E}_T) = \frac{\mathbf{E}_{\tilde{I}} \cdot \mathbf{E}_T}{||\mathbf{E}_{\tilde{I}}|| \cdot ||\mathbf{E}_T||} \quad (1)$$

We store CLIPScores for all synthetic image-text pairs and discard those below threshold τ to ensure quality.

To enhance representation quality, we combine synthetic image features with their corresponding text embeddings. Since synthetic images may not capture all semantic nuances while text embeddings lack visual grounding, fusing both modalities creates more robust representations.

Image-to-text Retrieval and Mapping. For retrieval, we use filtered synthetic images during training and input images during inference. We integrate retrieved captions through a multi-step feature fusion process. First, we encode the retrieved captions $\mathbf{E}_R = \text{CLIP}_t(R)$ and adjust their feature dimensions via linear projection layers f_{l1} and f_{l2}. The cross-attention module f_{Att} then fuses the processed retrieval features with the fused synthetic image-text features to obtain $\mathbf{E}_{\text{Attn}}$. Finally, $\mathbf{E}_{\text{Attn}}$ is fed into a trainable mapping network f_{Map} that encodes the semantic representation of the input. This process can be formalized as:

$$\mathbf{E}_{\text{Attn}} = f_{\text{Att}}\big(f_{l1}(\mathbf{E}_{\text{F}}),\ f_{l2}(\mathbf{E}_R)\big) \quad (2)$$

$$\mathbf{E}_{\text{map}} = f_{\text{Map}}(\mathbf{E}_{\text{Attn}}) \quad (3)$$

Negative Entity Filtering

Retrieved captions often contain entities that are semantically inconsistent with the input image, particularly in cross-domain scenarios where domain shift exacerbates this mismatch. Such inconsistent entities can introduce hallucinations that mislead the caption generation process. To address this problem, we propose entity-aware processing that explicitly identifies negative entities.

Entity Identification During Training. During training, we leverage the availability of ground-truth captions to establish supervised entity classification. We extract key entities e_{key} from ground-truth captions and candidate entities e_{can} from retrieved captions using parsing tools. Key entities automatically become positive entities e_{pos} (i.e., $e_{\text{pos}} = e_{\text{key}}$), while candidate entities not appearing in e_{key} are labeled as nega-

Method	Encoder	Decoder	COCO				Flickr30k			
			B@4	M	C	S	B@4	M	C	S
CapDec (2022)	RN50 × 4	GPT-2_{Large}	26.4	25.1	91.8	11.9	17.7	20.0	39.1	9.9
DeCap (2023)	ViT-B/32	Transformer$_{\text{Base}}$	24.7	25.0	91.2	18.7	21.2	21.8	56.7	15.2
ViECap (2023)	ViT-B/32	GPT-2_{Base}	27.2	24.8	92.9	18.2	20.3	20.2	47.8	13.6
SynTIC(2023)	ViT-B/32	Transformer$^{L=4}_{H=4}$	<u>29.9</u>	25.8	101.1	19.3	22.3	<u>22.4</u>	56.6	<u>16.6</u>
ICSD(2023)	ViT-B/32	BERT$_{\text{Base}}$	<u>29.9</u>	25.4	96.6	-	**25.2**	20.6	54.3	-
IFCap (2024)	ViT-B/32	GPT-2_{Base}	**30.8**	<u>26.7</u>	<u>108.0</u>	<u>20.3</u>	23.2	**22.9**	<u>64.4</u>	**17.0**
NES	ViT-B/32	GPT-2_{Base}	**30.8**	**26.8**	**109.9**	**20.6**	<u>24.3</u>	22.1	**66.8**	15.9

Table 1: The results of the in-domain captioning included the COCO test split and the Flickr30k test split.

tive entities $e_{\text{neg}} = e_{\text{can}} \setminus e_{\text{key}}$.

Entity Identification During Inference. During inference, ground-truth captions are unavailable, making it challenging to directly identify positive entities e_{pos}. We design a similarity-based filtering mechanism that leverages visual-semantic alignment for reliable entity selection.

The filtering process operates as follows: We use the ViECap method to extract key entities e_{key} from the input image. From retrieved captions, we extract candidate entities e_{can}. For each candidate entity not in e_{key}, we compute its similarity with the image representation. Entities exceeding threshold τ_{sim} are added to positive entities e_{pos}, while others become negative entities e_{neg}. This process is formulated as:

$$e_{\text{filter}} = e_{\text{can}} \setminus e_{\text{key}} \tag{4}$$

$$e_{\text{pos}} = e_{\text{key}} \cup \{e \in e_{\text{filter}} \mid S_{\text{sim}}(e, E_I) > \tau_{sim}\} \tag{5}$$

$$e_{\text{neg}} = e_{\text{filter}} \setminus e_{\text{pos}} \tag{6}$$

where $S_{\text{sim}}(e, E_I)$ is the cosine similarity function.

Attention-level Hallucination Suppression

To further leverage negative entities to improve the accuracy of generated captions and reduce hallucination tendencies caused by retrieved captions and language priors, we identify and suppress features related to negative entities in the fused features, thereby further enhancing the model's generalization capability.

Hallucination-prone Feature Filtering. To selectively suppress features associated with negative entities, we first encode all negative entities $\mathbf{e}_{\text{neg}}$ using the CLIP text encoder and then apply a cross-attention mechanism f_{neg} to compute attention weights between the negative entity embeddings $\mathbf{E}_{\text{neg}}$ (as queries) and the fused feature representations $\mathbf{E}_{\text{F}}$ (as keys and values). This process identifies which feature dimensions exhibit strong correlated with hallucination-prone entities. Based on the computed attention weights, we employ threshold-based filtering strategies to suppress the contribution of high-correlation features, thereby reducing hallucination tendencies in the final representations. Different filtering strategies are evaluated in the ablation study.

Hallucination-prone Feature Suppression. Based on the attention weights computed in the filtering stage, we identify features that exhibit strong correlations with negative entities. For features whose attention weights exceed threshold τ_{neg}, we apply a suppression factor $\lambda < 1$ to reduce their contribution to the decoder during caption generation. This selective suppression mechanism is formulated as:

$$\mathbf{E}_{\text{A}} = f_{\text{neg}}(\mathbf{E}_{\text{neg}}, \mathbf{E}_{\text{map}}) \tag{7}$$

$$\mathbf{E}_{\text{sup}} = \begin{cases} \lambda \cdot \mathbf{E}^{(h)}_{\text{map}}, & \text{if } \mathbf{E}^{(h)}_{\text{A}} > \tau_{\text{neg}} \\ \mathbf{E}^{(h)}_{\text{map}}, & \text{otherwise} \end{cases} \tag{8}$$

where $\mathbf{E}_{\text{sup}}$ represents the suppressed feature representations, τ_{neg} is the threshold of hallucinations, and λ is a hyperparameter that controls the suppression strength.

Experiments

Experimental Details

Implementation Details. We employ Stable Diffusion v1.5 for synthetic image generation at 512×512 resolution with 20 denoising steps. CLIP-ViT-B/32 serves as both the image and text encoder, while GPT-2_{base} acts as the text decoder. During training, we freeze the encoder parameters and only update the AS Module and text decoder. The model is trained for 5 epochs with a batch size of 80 using the AdamW optimizer with an initial learning rate of 2×10^{-5} and cosine annealing scheduling. For threshold selection, we set synthetic image filtering threshold $\tau = 0.6$, entity similarity threshold $\tau_{sim} = 0.2$, suppression strength λ is set to 0.3. All experiments are conducted on a single NVIDIA RTX 4080 Super GPU with 16GB VRAM, requiring approximately one hour and 12GB of GPU memory for training.

Datasets and Metrics. For in-domain evaluation, we conduct experiments on the MS-COCO (Chen et al. 2015) and Flickr30k (Peter Young et al. 2014) datasets using the Karpathy split (Andrej Karpathy and Li Fei-Fei 2015). For cross-domain evaluation, we assess performance on the NoCaps (Agrawal et al. 2019) validation set, which contains novel object categories not seen during training. We adopt standard image captioning metrics including CIDEr (Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh 2015), BLEU@4 (Kishore Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), and SPICE (Anderson et al. 2016). To evaluate hallucination suppression effectiveness, we report CHAIR (Rohrbach et al. 2019) scores on

COCO dataset, which measure the proportion of objects that are not present in the ground-truth captions. We also compute entity recall rates to assess how well our method captures ground-truth entities in the generated captions.

Baselines. We compare our model with several state-of-the-art text-only image captioning methods. CapDec (David Nukrai, Ron Mokady, and Amir Globerson 2022) and ViECap (Junjie Fei et al. 2023) build upon ClipCap (Ron Mokady, Amir Hertz, and Amit H. Bermano 2021), using Gaussian noise to align text and image features, with ViECap additionally incorporating entity extraction for hallucination mitigation. CLOSE (Sophia Gu, Christopher Clark, and Aniruddha Kembhavi 2023) explores various noise configurations, while DeCap (Wei Li et al. 2023) introduces a memory bank mechanism for improved retrieval. IFCap (Soeun Lee et al. 2024) employs retrieval-based entity filtering with frequency-based selection to replace image-derived entities with retrieved ones for enhanced accuracy. ICSD (Ma et al. 2024) and SynTIC (Liu, Liu, and Ma 2024) represent recent advances that leverage text-to-image generation models like Stable Diffusion (Rombach et al. 2022) to bridge the modality gap between text-only training and visual inference.

In-domain captioning

We evaluate NES on in-domain settings using the COCO and Flickr30k datasets, with results reported in Table 1. Compared with previous state-of-the-art text-only image captioning methods, NES achieves superior performance across all metrics on the COCO dataset, even outperforming models that employ larger architectures. On Flickr30k, NES demonstrates competitive performance on BLEU@4 and METEOR while achieving the highest CIDEr score, highlighting its effectiveness in generating accurate and contextually relevant captions.

Cross-domain captioning

We assess the transferability of NES across diverse domains through cross-domain scenarios between COCO and Flickr30k datasets, as well as the NoCaps validation set, with results presented in Table 2 and Table 3.

In cross-domain settings between COCO and Flickr30k, NES achieves state-of-the-art performance across most metrics in both transfer directions. Particularly noteworthy is the Flickr30k-to-COCO transfer scenario, we provide results for both inference with and without target-domain retrieval texts, where NES achieves substantial improvements with CIDEr scores of 69.9 compared to the best baseline IFCap's 60.7.

Hallucination Suppression Analysis

To evaluate the effectiveness of our negative entity suppression mechanism, we analyze hallucination rates using the CHAIR metric on COCO in-domain and Flickr30k-to-COCO cross-domain scenarios, as shown in Table 4. CHAIR measures the proportion of generated objects that are not present in the ground-truth captions, providing a direct assessment of hallucination suppression capability. Specifically, CHAIR-S (sentence-level) measures the percentage of generated captions containing at least one hallucinated object, while CHAIR-I (instance-level) measures the percentage of hallucinated objects among all generated objects.

NES achieves lower hallucination rates than baseline methods in both evaluation settings. The improvement is more notable in cross-domain scenarios, where baseline methods experience performance drops while NES maintains stable hallucination rates. This confirms that our negative entity suppression approach effectively reduces hallucinated content generation across different domains.

Methods	MSCOCO → Flickr30k				Flickr30k → MSCOCO			
	B@4	M	C	S	B@4	M	C	S
Without target domain's corpus								
DeCap	16.3	17.9	35.7	11.1	12.1	18.0	44.4	10.9
ViECap	17.4	18.0	38.4	11.2	12.9	19.3	54.2	12.5
SynTIC	17.9	18.6	38.4	11.9	14.6	19.4	47.0	11.9
IFCap	17.8	19.4	47.5	12.7	14.7	20.4	60.7	13.6
NES	**18.8**	**20.0**	**52.3**	**13.7**	**17.3**	**20.7**	**69.9**	**15.0**
With target domain's corpus								
SynTIC$-TT$	19.4	20.2	43.2	13.9	20.6	21.3	64.4	14.3
IFCap$-TT$	21.2	**21.8**	59.2	**15.6**	19.0	23.0	76.3	17.3
NES$-TT$	**21.3**	21.4	**59.3**	14.9	**22.9**	**23.8**	**87.5**	**18.6**

Table 2: Cross-Domain Evaluation. X → Y means source domain to target domain. $-TT$: models can access to target domain's corpus during inference time.

Methods	in-domain		near-domain		out-of-domain		Overall	
	C	S	C	S	C	S	C	S
CapDec	60.1	10.2	50.2	9.3	28.7	6.0	45.9	8.3
DeCap	65.2	-	47.8	-	25.8	-	45.9	-
ViECap	61.1	10.4	64.3	9.9	65.0	8.6	66.2	9.5
IFCap	**75.8**	**12.4**	72.3	11.6	60.2	8.9	70.5	10.8
IFCap*	70.1	11.2	72.5	10.9	72.1	9.6	74.0	10.5
NES	74.7	11.7	**75.4**	**11.7**	**72.6**	**10.2**	**76.2**	**11.2**

Table 3: Cross-domain captioning results on the NoCaps validation set. $\star$:without Entity Filtering module in the inference time.

Method	COCO In-domain			Flickr30k→COCO		
	C-S	C-I	Recall	C-S	C-I	Recall
ViECap	16.2	10.8	**43.7**	47.2	27.3	**45.9**
IFcap	9.7	6.6	42.9	22.8	15.1	42.3
NES	***8.1***	***5.6***	43.2	***12.1***	***8.0***	44.9

Table 4: CHAIR scores and entity recall on COCO in-domain and Flickr30k-to-COCO cross-domain scenarios. ***Lower CHAIR*** values indicate better hallucination suppression. **Higher recall** values indicate better coverage of ground-truth entities.

Ablation Study

We conduct extensive experiments to identify the impact of each key component in NES: Synthetic Images Retrieval

(SIR), Synthetic Images Fusion (SIF), Negative Entity Filtering (NEF), and Attention-level Suppression (AS). Given that our primarily aims to improve cross-domain generalization, we conduct ablation studies on the challenging Flickr30k-to-COCO transfer scenario to better demonstrate the effectiveness of each component in cross-domain settings.

SIR	SIF	NEF	AS	B@4	CIDEr	C-S	C-I	Recall
✓	✓	✓	✓	**22.9**	**87.5**	12.1	8.0	44.9
	✓	✓	✓	20.0	79.9	16.4	10.3	44.5
✓		✓	✓	20.7	82.9	12.0	7.9	44.7
✓	✓	✓		21.4	83.9	11.8	7.9	44.1
✓	✓			21.4	83.0	8.6	***6.2***	40.8
✓				21.0	83.3	***8.4***	***6.2***	41.4
	✓			13.8	50.2	21.8	18.6	31.9
		✓	✓	19.1	77.7	23.9	14.3	**45.3**
				19.0	76.3	22.8	15.1	42.3

Table 5: Ablation studies of the key components of NES.

Ablation Settings: To systematically analyze each component's contribution in cross-domain scenarios, we conduct ablation studies on the Flickr30k-to-COCO task. For the Synthetic Image-based Retrieval (SIR) module, we replace synthetic images with original text captions as retrieval queries while maintaining the identical retrieval and processing pipeline. For Synthetic Image Fusion (SIF), we remove the fusion mechanism and use only text features for generation. For Negative Entity Filtering (NEF), we disable entity filtering by skipping the negative entity identification step, allowing all retrieved entities to pass through the pipeline. For Attention-level Suppression (AS), we disable the attention-based suppression mechanism. The results are shown in Table 5.

Synthetic Image-based Retrieval: When using SIR, all metrics except recall show significant improvements, especially hallucinations decrease about 50%. This indicates that synthetic image-based retrieval identifies more semantically relevant captions compared to text-based methods. Notably, SIR alone achieves the lowest hallucination rates, confirming its effectiveness in reducing false content. To further examine SIR's impact, we vary the number of retrieved descriptions. While recall shows an increasing trend as the number of retrieved captions grows, analysis reveals that numerous retrieval captions introduce synonym confusion (e.g., 'tv' vs 'television').

Synthetic Image Fusion: SIF is a simple yet important module designed to combine the richness of visual features with the accuracy of textual descriptions for optimal decoding performance. When using CLIPScore α for fusion, we test both forward fusion ($\alpha \cdot E_{\tilde{I}} + (1-\alpha) \cdot E_T$) and reverse fusion ($(1-\alpha) \cdot E_{\tilde{I}} + \alpha \cdot E_T$). Additionally, we evaluate the effects of different fixed values for fusion weights. The results reveal an interesting trade-off: CLIPScore-based fusion achieves higher captioning scores and recall rates, while fixed values can achieve lower hallucination rates. This suggests that dynamic weighting based on CLIPScore optimizes overall caption quality, whereas fixed weighting provides more stable suppression of hallucinated content.

Negative Entity Filtering: NEF serves as the most critical component of our framework, designed to identify and filter out entities absent from images while preserving generation quality. To evaluate NEF's effectiveness, we calculate the number of hallucinated entities in generated captions and their sources. Hallucinations are categorized as retrieval-sourced if they appear in retrieved entities, or model-sourced otherwise. We compare three filtering strategies: GT (ground truth filtering), CLIP-Det (CLIP detection filtering), and our NEF approach. The results demonstrate that NEF effectively reduces retrieval-sourced hallucinations while maintaining competitive generation quality, showing that our approach strikes an optimal balance between filtering precision and content preservation.

Attention-level Suppression: Although NEF filters out negative entities from the entity list used for final caption generation, the retrieved captions used in feature fusion still contain textual descriptions of these entities. To address this residual issue, AS employs feature-level suppression by identifying and downweighting attention heads that strongly correlate with negative entity embeddings.

We evaluate different suppression strategies for selecting attention heads and varying suppression intensities controlled by hyperparameter λ. Our analysis reveals that $\lambda = 0.3$ achieves the best overall performance (CIDEr: 87.5), yielding 3.6 CIDEr improvement over the configuration without AS (CIDEr: 83.9). This moderate suppression level strikes a favorable balance: stronger suppression ($\lambda < 0.3$) removes excessive features and degrades caption quality, while weaker suppression ($\lambda > 0.3$) retains more negative entity correlations without providing additional performance gains. Notably, AS introduces a trade-off between caption quality and hallucination metrics, as evidenced by a slight increase in CHAIR-S from 11.8 to 12.1 when AS is enabled. This suggests that AS primarily enhances cross-domain generalization by preserving diverse semantic features rather than strictly minimizing hallucinations, which aligns with the improved recall rates (44.9 vs 44.1) observed in our experiments.

Conclusion

This paper presents NES, a framework that addresses cross-domain degradation in ZIC. The core insight is that text-only training creates inconsistencies with visual inference, leading to hallucinations that harm cross-domain performance. Our solution employs synthetic images across training for consistent retrieval, filters negative entities from retrieved content, and applies attention-level suppression to reduce their influence in fused features. Experiments demonstrate improvements over existing methods, with notable improvements in cross-domain scenarios. Our work validates two key principles: synthetic images can effectively bridge modality gaps in cross-modal learning, and targeted hallucination suppression enables robust cross-domain generalization in zero-shot captioning tasks. Future work will explore advanced strategies for handling semantically similar negative entities and developing more precise suppression mechanisms.

References

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel Object Captioning at Scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: Semantic Propositional Image Caption Evaluation.

Andrej Karpathy; and Li Fei-Fei. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *IEEE Trans. Pattern Anal. Mach. Intell.*

Anna Rohrbach; Lisa Anne Hendricks; Kaylee Burns; Trevor Darrell; and Kate Saenko. 2018. Object Hallucination in Image Captioning. In Ellen Riloff; David Chiang; Julia Hockenmaier; and Jun'ichi Tsujii, eds., *WACV*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Goldstein, J.; Lavie, A.; Lin, C.-Y.; and Voss, C., eds., *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollar, P.; and Zitnick, C. L. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server.

David Nukrai; Ron Mokady; and Amir Globerson. 2022. Text-Only Training for Image Captioning Using Noise-Injected CLIP. In *EMNLP*.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation.

Junjie Fei; Teng Wang; Jinrui Zhang; Zhenyu He; Chengjie Wang; and Feng Zheng. 2023. Transferable Decoding with Visual Entities for Zero-Shot Image Captioning. In *KSE*.

Kim, T.; Lee, S.; Kim, S.-W.; and Kim, D.-J. 2025. ViP-Cap: Retrieval Text-Based Visual Prompts for Lightweight Image Captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Kishore Papineni; Salim Roukos; Todd Ward; and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In Pierre Isabelle; Eugene Charniak; and Dekang Lin, eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Lee, J. R.; Shin, Y.; Son, G.; and Hwang, D. 2025. Diffusion Bridge: Leveraging Diffusion Model to Reduce the Modality Gap Between Text and Vision for Zero-Shot Image Captioning. In *Cvpr*.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2023. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding.

Liu, Z.; Liu, J.; and Ma, F. 2024. Improving Cross-Modal Alignment with Synthetic Pairs for Text-Only Image Captioning. In *AAAI*.

Lu, Z.; Xu, H.; Liu, B.; and Wang, K. 2025. Negative Entity Suppression for Zero-Shot Captioning with Synthetic Images. In *AAAI*.

Ma, F.; Zhou, Y.; Rao, F.; Zhang, Y.; and Sun, X. 2024. Image Captioning with Multi-Context Synthetic Data. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Peter Young; Alice Lai; Micah Hodosh; and Julia Hockenmaier. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. In Dekang Lin; Michael Collins; and Lillian Lee, eds., *Trans. Assoc. Comput. Linguistics*.

Petryk, S.; Spencer, D.; Chen, C.; Gonzalez, J. E.; and Darrell, T. 2024. ALOHa: A New Measure for Hallucination in Captioning Models. arXiv:2104.02717.

Ramakrishna Vedantam; C. Lawrence Zitnick; and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *W-NUT*.

Ramos, R.; Elliott, D.; and Martins, B. 2023. Retrieval-Augmented Image Captioning. In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.

Rita Ramos; Bruno Martins; Desmond Elliott; and Yova Kementchedjhieva. 2023. SmallCap: Lightweight Image Captioning Prompted with Retrieval Augmentation.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2019. Object Hallucination in Image Captioning. In *WACV*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ron Mokady; Amir Hertz; and Amit H. Bermano. 2021. ClipCap: CLIP Prefix for Image Captioning. In *CoRR*.

Sarkar, R.; Zhang, W.; and Liu, X. 2025. Mitigating Hallucinations in Vision-Language Models.

Soeun Lee; Si-Woo Kim; Taewhan Kim; and Dong-Jin Kim. 2024. IFCap: Image-like Retrieval and Frequency-based Entity Filtering for Zero-shot Captioning. In *EMNLP*.

Sophia Gu; Christopher Clark; and Aniruddha Kembhavi. 2023. I Can't Believe There's No Images! Learning Visual Tasks Using Only Language Supervision. In *ICCV*.

Wei Li; Linchao Zhu; Longyin Wen; and Yi Yang. 2023. DeCap: Decoding CLIP Latents for Zero-Shot Captioning via Text-Only Training.

Yang, Z.; Ping, W.; Liu, Z.; Korthikanti, V.; Nie, W.; Huang, D.-A.; Fan, L.; Yu, Z.; Lan, S.; Li, B.; Liu, M.-Y.; Zhu, Y.; Shoeybi, M.; Catanzaro, B.; Xiao, C.; and Anandkumar, A. 2023. Re-ViLM: Retrieval-Augmented Visual Language Model for Zero and Few-Shot Image Captioning.

Zeng, D.; Shen, Y.; Lin, M.; Yi, Z.; and Ouyang, J. 2025. Zero-Shot Image Captioning with Multi-type Entity Representations. In *AAAI*.

Zequn Zeng; Yan Xie; Hao Zhang; Chiyu Chen; Zhengjue Wang; and Bo Chen. 2024. MeaCap: Memory-Augmented Zero-shot Image Captioning.