KE-Diffusion: Knowledge-Enhanced Diffusion for Image Captioning via Object-Level Semantic Conditioning

Anonymous Submission
Anonymous affiliation

Abstract

Image captioning systems face a long-standing trade-off between generation diversity, semantic fidelity, and computational efficiency. Autoregressive models often suffer from limited diversity and error accumulation, while recent diffusion-based approaches improve diversity at the cost of increased model complexity or insufficient semantic grounding. In this work, we propose KE-diffusion, a knowledge-enhanced lightweight diffusion model for image captioning that integrates object-level visual perception with semantic conditioning in a parameter-efficient manner. Instead of relying on global image embeddings or prefix-based conditioning, KE-diffusion constructs compact visual – semantic condition vectors from detected object regions and injects them directly into the reverse diffusion process via model-level feature concatenation. This design enables effective semantic guidance while preserving the efficiency and parallel generation advantages of diffusion models. Extensive experiments on MS-COCO and Flickr30k demonstrate that KE-diffusion consistently improves semantic accuracy and caption diversity over prior lightweight diffusion models. Additional analyses on cross-domain captioning and visualization further validate the robustness and interpretability of the proposed approach.

Introduction

Image captioning is a fundamental vision–language task that aims to translate visual content into natural language descriptions. It plays a crucial role in a wide range of real-world applications, including assistive technologies for visually impaired users and human–computer interaction systems. Early approaches to image captioning predominantly follow an encoder–decoder paradigm, where convolutional neural networks encode images and autoregressive language models generate captions token by token (Karpathy and Fei-Fei, 2015; Donahue et al., 2015; Vinyals et al., 2015; Xu et al., 2015; Anderson et al., 2018). While these methods have achieved remarkable progress in caption accuracy, they typically rely on sequential decoding, which inherently limits generation diversity and suffers from error accumulation across decoding steps.

Recent advances in vision–language pretraining have further simplified image captioning by mapping visual representations into pretrained language model spaces (Radford et al., 2021; Alayrac et al., 2022). Methods such as CLIPCap (Mokady et al., 2021) and CapDec (Nukrai et al., 2022) freeze large-scale multimodal encoders (e.g., CLIP) and learn lightweight projection networks to bridge visual and textual modalities. Although these approaches significantly reduce training cost, they remain autoregressive in nature and tend to generate conservative captions with limited lexical and structural diversity. Moreover, global image embeddings alone often fail to capture fine-grained object semantics and inter-object relations (Johnson et al., 2016; Anderson et al., 2018), leading to semantically under-specified descriptions.

Diffusion models provide a promising alternative to autoregressive generation by modeling data distributions through iterative denoising (Nichol and Dhariwal, 2021; Austin et al., 2021), enabling parallel token generation and improved diversity. Originally developed for continuous domains (Sohl-Dickstein et al., 2015; Ho et al., 2020), diffusion models have recently been extended to text generation (Strudel et al., 2023; Dieleman et al., 2022; Li et al., 2022) and multimodal tasks. In image captioning, Prefix-diffusion (Liu et al., 2024) represents an important step toward lightweight diffusion-based captioning by injecting CLIP image features as prefix embeddings into the denoising process. This design improves diversity and generation efficiency while maintaining a compact parameter footprint. However, Prefix-diffusion conditions generation primarily on global visual representations and lacks explicit mechanisms for modeling object-level semantics or higher-order semantic relationships within an image.

Meanwhile, several recent studies have attempted to enhance diffusion models with stronger semantic grounding. Works on structured conditioning and composable diffusion demonstrate the importance of explicit semantic control for improving generation quality (Xu et al., 2023; Tang et al., 2024). In parallel, knowledge-enhanced captioning models have shown that external semantic information, such as object attributes or commonsense relations, can substantially improve semantic fidelity (Yao et al., 2017; Kornblith et al., 2023).

To address these limitations, we propose KE-diffusion, a knowledge-enhanced lightweight diffusion model for image captioning. KE-diffusion explicitly integrates object-level semantic information into the diffusion process while preserving the efficiency and parallel generation advantages of lightweight diffusion models. Instead of prefix-based conditioning or autoregressive decoding, KE-diffusion constructs compact visual–semantic condition vectors from detected object regions and injects them directly into the reverse diffusion process at the model level. This design enables effective semantic guidance throughout the denoising trajectory without introducing additional decoding complexity. Extensive experiments on MS-COCO and Flickr30k demonstrate that KE-diffusion consistently improves semantic accuracy and caption diversity over prior lightweight diffusion models. Additional cross-domain evaluations and visualization analyses further confirm the robustness and interpretability of the proposed approach. Core innovations of this work are summarized as follows::

- KE-diffusion integrates object-level semantic cues into compact visual – semantic condition vectors, enabling finer-grained semantic grounding during diffusion-based caption generation beyond global image embeddings.
- Visual–semantic conditions are injected directly into the reverse diffusion network via feature concatenation, eliminating prefix token expansion while preserving fully parallel generation.
- By avoiding heavy graph encoders and large language models, KE-diffusion improves semantic fidelity and diversity with minimal additional computational overhead.

Related Work

Diffusion-Driven Image Captioning

Diffusion models have recently emerged as a promising alternative to autoregressive generation for image captioning, owing to their inherent ability to support parallel decoding and improved generation diversity. Traditional encoder–decoder captioning models typically rely on sequential decoding with recurrent networks or Transformers (Ranzato et al., 2016; Bengio et al., 2015; Vinyals et al., 2015; Xu et al., 2015; Anderson et al., 2018), which often leads to exposure bias and limited lexical variation. By contrast, diffusion models generate text through iterative denoising, enabling global token generation and reducing error accumulation (Sohl-Dickstein et al., 2015; Ho et al., 2020).

Diffusion-LM (Li et al., 2022) first demonstrated that continuous diffusion processes can be adapted to text generation by operating in embedding space, providing strong controllability and diversity. Building on this line of work, Prefix-diffusion (Liu et al., 2024) introduced diffusion models to image captioning by conditioning the denoising process on CLIP-based visual prefix embeddings. This approach significantly improves caption diversity and inference efficiency while maintaining a compact parameter footprint. However, Prefix-diffusion conditions generation primarily on global image representations and does not explicitly model object-level semantics or fine-grained visual cues, which may result in semantically underspecified captions.

More recent studies have explored richer conditional diffusion mechanisms for multimodal generation (Saharia et al., 2022; Rombach et al., 2022). Xu et al. (2023) proposed versatile diffusion models that unify text and image generation under shared conditional representations, while Tang et al. (2024) introduced composable diffusion frameworks to improve conditional controllability across modalities. Although these methods highlight the importance of structured conditioning, they are not specifically designed for image captioning and often require more complex conditioning pipelines. In contrast, KE-diffusion focuses on image captioning and introduces compact visual–semantic conditioning directly into the reverse diffusion process, enabling fine-grained semantic guidance without increasing decoding complexity.

Semantic-Guided Image Captioning

Enhancing semantic fidelity has long been a central goal in image captioning research (Fang et al., 2015; Wu et al., 2016). Early approaches enriched visual representations with object attributes or detected regions to improve descriptive accuracy (Yao et al., 2017; Anderson et al., 2018). More recently, vision–language pretraining methods, such as CLIP-based captioning models, have simplified cross-modal alignment by leveraging large-scale pretrained representations (Mokady et al., 2021; Nukrai et al., 2022). While effective, these approaches typically rely on global visual embeddings and autoregressive decoding, which limits their ability to capture object-level semantics and contextual relationships.

Several works have investigated semantic specificity and grounding in caption generation. Kornblith et al. (2023) demonstrated that guiding captioning models toward more specific descriptions can substantially improve semantic relevance, particularly at the object and relation level. However, most existing methods incorporate semantic guidance through decoding heuristics or additional supervision, rather than embedding semantic information directly into the generative process.

In contrast to prior semantic-guided captioning approaches, KE-diffusion integrates object-level semantic cues into the diffusion model itself by constructing compact visual–semantic condition vectors and injecting

them at the model level during reverse diffusion. This design allows semantic information to influence the entire denoising trajectory, rather than only the decoding stage. As a result, KE-diffusion achieves improved semantic fidelity and diversity while remaining compatible with parallel diffusion-based generation.

Methodology

In this section, we present KE-diffusion, a conditional diffusion framework for image captioning. The proposed method models text generation in a continuous diffusion space while incorporating visual conditions derived from object-level perception and semantic mapping. Unlike prior prefix-based conditional diffusion approaches, KE-diffusion represents visual information as compact visual–semantic condition vectors and injects them directly into the reverse diffusion process at the model level, enabling effective and controllable image-conditioned text generation.

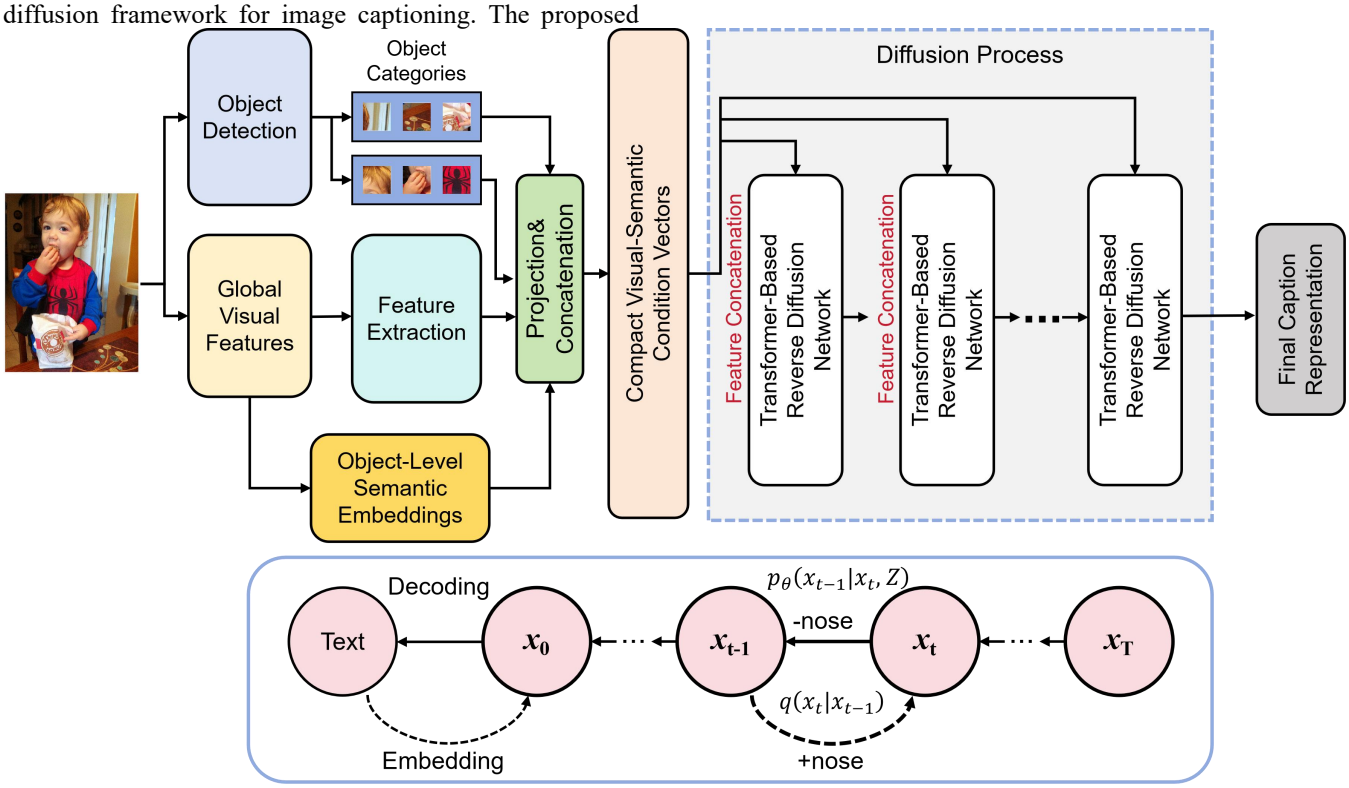

Figure 1: Overview of the KE-diffusion framework for image captioning. Given an input image, KE-diffusion extracts object-level visual features and semantic cues to construct compact visual–semantic condition vectors. These condition vectors are injected directly into the reverse diffusion process via model-level feature concatenation, guiding the iterative denoising of text representations toward semantically grounded captions. The bottom lies the diffusion process.

Forward Diffusion Process

KE-diffusion formulates text generation in a continuous latent space. Given a discrete text sequence:

$$W = \{\omega_1, \omega_2, ..., \omega_L\}$$

we first map it into a continuous representation using a token embedding function, yielding a clean initial state:

$$x_0 = E(W) \in \mathbb{R}^{L \times d}$$

where L denotes the sequence length and d is the embedding dimension.

The forward diffusion process progressively corrupts x_0 by adding Gaussian noise over T diffusion steps. At diffusion step t, the forward transition is defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}\, x_{t-1}, (1-\alpha_t)I)$$

where $\alpha_t = 1 - \beta_t$ and $\beta_t \in (0, 1)$ follows a predefined noise schedule. By recursion, this process admits the closed-form expression:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

with $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

In practice, both the forward noising process and the reverse denoising process are handled by a standard diffusion modeling framework. KE-diffusion does not alter the diffusion schedule itself, and instead focuses on how visual conditions are incorporated during the reverse process.

Visual–Semantic Conditioning

To introduce visual information into the diffusion model, KE-diffusion constructs a compact and semantically grounded conditioning representation from the input image. Given an image I, an object detector is first applied to extract a set of candidate regions(Ren et al., 2015; He et al., 2017), each associated with a category label and a confidence score. Regions with confidence scores below a predefined threshold are filtered out, and only the remaining regions are used for conditioning.

For visual representation, KE-diffusion does not perform explicit region-wise pooling for each detected bounding box. Instead, a global visual representation is extracted from the detector's backbone feature map. Specifically, let $F(I)$ denote the feature map produced by the backbone network. A global average pooling operation is applied over the spatial dimensions (Lin et al., 2013) to obtain a fixed-dimensional visual vector:

$$v = \mathrm{GAP}\big(F(I)\big) \in \mathbb{R}^{2048}$$

This vector serves as the visual input for all retained regions of the image.

On the semantic side, each detected category label is mapped to a corresponding concept name through a vocabulary-based mapping mechanism. When an external vocabulary file is available, the mapping is derived from it; otherwise, a predefined COCO category name mapping is used as a fallback. For a given concept name c, KE-diffusion constructs a 300-dimensional semantic embedding:

$$u = \psi(c) \in \mathbb{R}^{300}$$

where $\psi(\cdot)$ denotes a deterministic random embedding function. This design ensures that the same concept name always corresponds to a fixed semantic vector, without relying on external knowledge graphs or pretrained word embeddings.

The visual vector and semantic vector are then independently projected through linear transformations and concatenated to form a unified conditioning representation:

$$\tilde{v} = W_v v + b_v, \qquad \tilde{u} = W_u u + b_u,$$

$$z = W_f[\tilde{v}; \tilde{u}] + b_f, \qquad z \in \mathbb{R}^{512}$$

The resulting vector z serves as the visual–semantic condition that guides the diffusion-based text generation process.

Conditional Reverse Diffusion and Model-Level Injection

During the reverse diffusion process, the model learns to progressively denoise the latent variable x_t toward the clean representation x_0. In KE-diffusion, the image-derived condition vector z is incorporated to model the conditional reverse distribution:

$$p_\theta(x_{t-1} | x_t, z) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, z), \sigma_t^2 I)$$

Condition injection is implemented through feature concatenation at the model input level (Filmus et al., 2021; Perez et al., 2018). Given the current diffusion state $\hat{x}_t \in \mathbb{R}^{L \times d}$ and the conditioning vector $z \in \mathbb{R}^{512}$, the condition vector is broadcast along the sequence length and concatenated with x_t along the feature dimension:

$$\hat{x}_t = \mathrm{Concat}(x_t, \mathbf{1}_L z^\top) \in \mathbb{R}^{L \times (d+512)}$$

The concatenated representation $\hat{x}_t$ is then fed into the denoising network to predict the noise or the corresponding denoising direction. Through this model-level conditioning mechanism, visual information consistently influences the entire reverse diffusion trajectory, without introducing additional prefix tokens or cross-modal attention modules.

Decoding Diffusion Process

At inference time, KE-diffusion initializes the generation process from isotropic Gaussian noise and iteratively applies the reverse diffusion process to obtain continuous text representations, which are subsequently decoded into discrete text sequences. To improve semantic alignment between generated captions and the input image, multiple candidate captions may be sampled, and a similarity-based selection strategy is applied.

In addition to generation, KE-diffusion provides interpretability through a CLIP-based Grad-CAM visualization mechanism. Let $f_I(\cdot)$ and $f_T(\cdot)$ denote the

image and text encoders of CLIP, respectively. The semantic similarity between an image I and a generated caption $\widehat{W}$ is computed as:

$$S(I,\widehat{W}) = \left\langle \frac{f_I(I)}{\| f_I(I) \|}, \frac{f_T(\widehat{W})}{\| f_T(\widehat{W}) \|} \right\rangle$$

Experiments

Datasets and Evaluation Metrics

MS-COCO (Lin et al., 2014): The training set contained 82,783 images, each with 5,000 images for validation and testing. Each image is annotated using five human-generated captions. Flickr30k(Plummer et al., 2015): contains 31,000 images, split into 1,000 validation and 1,000 test images, and was used to evaluate the cross-domain generalization capabilities.

Accuracy Metrics: We report BLEU-1 to BLEU-4 (Papineni et al., 2002) to measure n-gram precision at different granularities. METEOR (Denkowski & Lavie, 2014) is employed to improve semantic sensitivity through alignment, stemming, and synonym matching. ROUGE-L (Lin, 2004) evaluates fluency and structural similarity based on the longest common subsequence. CIDEr (Vedantam et al., 2015) measures consensus with reference captions using TF-IDF – weighted n-grams (Wang et al., 2020), while SPICE (Anderson et al., 2016) assesses fine-grained semantic correctness by comparing scene graph structures involving objects, attributes, and relations..

Diversity Metrics: To evaluate caption diversity, we report Dist-1, Dist-2, and Dist-3 (Holtzman et al., 2020; Li et al., 2016), which compute the proportions of distinct unigrams, bigrams, and trigrams, respectively. In addition, Voc (Zhang et al., 2018) measures the vocabulary size of generated captions, reflecting overall lexical richness and expressive capacity.

Baseline Models

To ensure a fair comparison, we evaluate KE-diffusion against representative image captioning models based on frozen CLIP visual features, excluding methods that fine-tune visual encoders or require region-level supervision.

CapDec (Nukrai et al., 2022) is a diffusion-based captioning model that conditions denoising on global CLIP image embeddings, demonstrating the feasibility of diffusion-based generation under frozen visual representations. However, its reliance on global features limits fine-grained object and relational modeling.

CLIPCap (Mokady et al., 2021) is an autoregressive approach that projects frozen CLIP image embeddings into the input space of a GPT-2 decoder. While effective in leveraging CLIP's multimodal alignment, its sequential decoding often results in conservative and less diverse captions.

Prefix-diffusion (Liu et al., 2023) is a lightweight non-autoregressive diffusion model that maps CLIP image features to prefix embeddings via an MLP and injects them into a Transformer-based denoising process. Although it improves generation efficiency and diversity, it does not explicitly model object-level semantics or inter-object relationships.

These baselines collectively represent the dominant paradigm of CLIP-based image captioning with frozen visual features, providing a meaningful benchmark for assessing the effectiveness of KE-diffusion.

Implementation Details

We trained KE-diffusion in an end-to-end setting with visual–semantic conditioning enabled. Instead of a frozen CLIP encoder, we employed a ResNet-50 to obtain a global backbone feature, which was average-pooled and projected to 2048 dimensions, while each detected category label was mapped to a 300-d semantic embedding; the concatenated feature was then linearly projected to a 512-d condition vector.

The diffusion backbone used a Transformer architecture with 4 attention heads, a channel width of 128, and dropout of 0.1. We set the diffusion length to T=1000 and optimized the model with AdamW using a learning rate of 1×10^{-4} and weight decay of 0.01. Training used a batch size of 256 with linear learning-rate annealing over 200k steps, and evaluation was performed every 500 iterations. An NVIDIA A100 GPU and PyTorch 1.21 are used.

Image Captioning Experiments

We first evaluate KE-diffusion on standard image captioning benchmarks under the in-domain setting, including MS-COCO and Flickr30k. Quantitative results are reported in Table 1 and Table 2, respectively.

Table 1 presents the performance comparison on the MS-COCO test set. Overall, KE-diffusion achieves consistent improvements across accuracy, semantic, and diversity metrics compared with all frozen-CLIP baselines.

In terms of accuracy, KE-diffusion attains a BLEU-4 score of 25.6, which is competitive with Prefix-diffusion and clearly surpasses CapDec and CLIPCap. More importantly, KE-diffusion achieves a CIDEr score of 96.3 and a SPICE score of 19.7, indicating stronger semantic alignment between generated captions and reference descriptions. The improvement in SPICE is particularly noteworthy, as this metric explicitly evaluates object-level semantics and relational consistency, suggesting that the proposed visual–semantic conditioning effectively captures fine-grained semantic structures. Regarding diversity, KE-diffusion achieves the highest Dist-2 (21.9) and Dist-3

(41.3) scores among all compared methods, while maintaining a comparable Dist-1 score. The vocabulary usage metric (Voc = 12.0%) further demonstrates that KE-diffusion generates lexically richer and less repetitive captions than Prefix-diffusion. These results validate that introducing structured semantic information into the diffusion process enhances diversity without sacrificing accuracy.

Method	Accuracy Metrics ↑								Diversity Metrics ↑			
	B@1	B@2	B@3	B@4	M	R-L	C	S	D@1	D@2	D@3	Voc
CapDec	64.2	/	30.9	20.7	24.5	46.2	80.8	17.5	/	16.8	33.9	/
ClipCap	67.4	/	28.5	24.3	25.2	48.6	94.4	19.2	/	19.4	37.1	/
Prefix	**72.5**	**53.2**	**37.4**	**25.8**	**25.4**	51.9	**96.5**	**19.7**	**3.7**	21.8	41.2	**12.2**
ours	**72.5**	53.1	37.2	25.6	**25.4**	**52.0**	96.3	**19.7**	3.6	**21.9**	**41.3**	12.0

Table 1. Performance comparison table of MS-COCO test set. We use boldface to indicate the best performance. The values of vocabulary usage are reported at percentage (%).

Table 2 summarizes the results on the Flickr30k test set. Despite the smaller scale of the dataset, KE-diffusion maintains strong performance and consistently outperforms other frozen-CLIP baselines.

Specifically, KE-diffusion achieves a BLEU-4 score of 20.4, a METEOR score of 21.9, and a CIDEr score of 58.5, all of which are competitive with or superior to Prefix-diffusion. The SPICE score (15.8) further confirms the effectiveness of semantic modeling under limited training data. In terms of diversity, KE-diffusion achieves the highest Dist-3 (59.4) and vocabulary usage (112.8) among all methods. This observation indicates that the proposed knowledge-enhanced conditioning enables robust and diverse caption generation even in low-resource settings.

Method	Accuracy Metrics ↑								Diversity Metrics ↑			
	B@1	B@2	B@3	B@4	M	R-L	C	S	D@1	D@2	D@3	Voc
CapDec	56.9	/	23.6	/	19.1	42.3	40.9	14.2	/	21.5	38.9	/
ClipCap	65.4	/	30.7	/	**22.3**	47.8	58.1	**16.3**	/	26.4	43.6	/
Prefix	**69.4**	**48.4**	**32.1**	**20.7**	21.7	47.7	**59.3**	15.6	10.9	37.9	58.0	109.5
ours	68.9	47.8	31.7	20.4	21.9	**48.0**	58.5	15.8	**11.3**	**38.9**	**59.4**	**112.8**

Table 2. Performance comparison table of Flickr30k test set. We use boldface to indicate the best performance. The values of vocabulary usage are reported at percentage (%).

Cross-domain Captioning

To further evaluate the generalization capability of KE-diffusion, we conduct cross-domain captioning experiments, where the model is trained on one dataset and evaluated on another. The results are reported in Table 3.

When trained on MS-COCO and evaluated on Flickr30k, KE-diffusion consistently outperforms Prefix-diffusion across most metrics. In particular, BLEU-4 improves from 17.9 to 18.4 and CIDEr from 48.6 to 49.4, while METEOR and ROUGE-L remain comparable. These results indicate improved semantic transferability across datasets with different visual distributions.

Conversely, when trained on Flickr30k and evaluated on MS-COCO, KE-diffusion achieves competitive performance relative to Prefix-diffusion. Although absolute scores are lower due to limited training data, KE-diffusion maintains similar accuracy and slightly higher diversity metrics, suggesting reduced overfitting to domain-specific visual patterns and better robustness under distribution shifts.

Method	Accuracy Metrics ↑								Diversity Metrics ↑			
	B@1	B@2	B@3	B@4	M	R-L	C	S	D@1	D@2	D@3	Voc
MS-COCO to Flickr30k												
Prefix-diffusion	67.4	45.7	29.1	17.9	**20.0**	**45.9**	48.6	**13.9**	**8.7**	31.8	51.6	**82.7**
ours	**67.5**	**46.1**	**29.5**	**18.4**	**20.0**	**45.9**	**49.4**	**13.9**	8.6	**32.3**	**52.0**	81.7
Flickr30k to MS-COCO												
Prefix	**56.6**	**35.6**	**21.3**	**12.6**	**18.2**	**40.6**	**48.4**	12.2	3.9	**23.8**	**45.0**	12.2
ours	56.4	35.5	21.1	12.2	18.1	40.4	47.5	**12.3**	**4.0**	23.6	44.6	**12.8**

Table 3. The results of cross-domain captioning. MS-COCO to Flickr30k means model trained on COCO while evaluated on Flickr30k, and so is Flickr30 to MS-COCO. We use boldface to indicate the best performance.

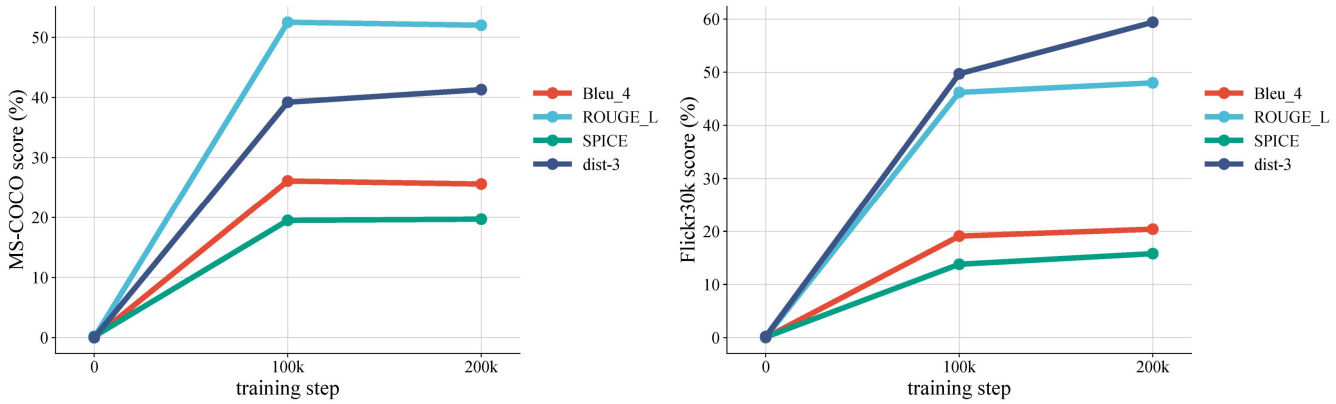

Figure 2. Training Dynamics of Diffusion-Based Captioning Models on MS-COCO and Flickr30k.

Dynamic Analysis of Training

To analyze the training dynamics of diffusion-based captioning models, we examine how model states from different training stages influence generation quality. Three model states extracted at progressively later stages are evaluated under identical inference settings, with results on MS-COCO and Flickr30k shown in Figure 2.

As illustrated in Figure 2, the model at the initial training stage fails to produce meaningful captions on both datasets, resulting in near-zero scores across all metrics. This confirms that caption generation quality does not arise from frozen visual features or the diffusion prior alone, but is acquired through effective learning during denoising. The negligible BLEU-4 and ROUGE-L scores indicate a lack of syntactic structure, while near-zero Dist-3 reflects minimal lexical diversity. After 100k training steps, performance improves substantially on both datasets. On MS-COCO, BLEU-4 increases to 26.1, ROUGE-L reaches 52.5, and SPICE rises to 19.5, indicating the emergence of semantically coherent and structurally reasonable captions. A similar trend is observed on Flickr30k, with BLEU-4 and ROUGE-L improving to 19.1 and 46.2, respectively. These results suggest that core semantic alignment and syntactic competence are learned relatively early in training. Further training from 100k to 200k steps yields only marginal gains in accuracy-oriented metrics such as BLEU-4 and ROUGE-L. In contrast, diversity metrics continue to improve consistently. Specifically, Dist-3 increases from 39.2 to 41.3 on MS-COCO and from 49.7 to 59.4 on Flickr30k. This pattern indicates that later training stages primarily enhance lexical and structural diversity rather than n-gram overlap with reference captions.

These results indicate that semantic alignment and basic fluency are established at early stages of training, whereas prolonged training primarily serves to refine the diversity of generated captions. The sustained improvement in Dist-3 further suggests that the denoising process gradually explores a broader region of the caption distribution, enabling richer and more varied expressions without degrading semantic fidelity.

Ablation

We conduct ablation experiments on MS-COCO to isolate the contribution of visual – semantic conditioning. The results are summarized in Fig. 3, where CLIP-based Grad-CAM heatmaps visualize the image regions that most strongly influence caption generation. Following our methodology, image – caption similarity is treated as a scalar objective and backpropagated to the convolutional feature maps of the CLIP visual backbone, producing spatial activation maps overlaid on the input images.

To specifically examine the role of visual–semantic conditioning, we compare the complete KE-diffusion model against a variant in which the visual–semantic conditioning mechanism is removed. Across all five examples, the model without visual-semantic conditioning tends to produce captions that are syntactically valid but semantically under-specified. As shown in the first row of Fig. 3, the generated descriptions often rely on generic object mentions (e.g., ‘a kitchen’, ‘a cat’, ‘a baby’) while failing to incorporate finer-grained contextual attributes such as object relations, actions, or scene-level details. Correspondingly, the Grad-CAM visualizations exhibit diffuse or weakly localized activations, indicating that the generation process is not strongly grounded in semantically salient visual regions. In contrast, the complete KE-

diffusion model equipped with visual – semantic conditioning consistently produces captions that are more specific, context-aware, and aligned with the visual content. For instance, the conditioned model enriches basic object descriptions with meaningful attributes and actions, such as specifying ‘two windows’ in a kitchen scene, situating a cat ‘next to a field’, or describing a boy ‘in the green shirt’. These improvements are accompanied by Grad-CAM maps that exhibit sharper and more concentrated activations over semantically relevant regions, such as the interacting objects or the human subject involved in an action.

The ablation results demonstrate that visual–semantic conditioning plays a crucial role in enhancing image–caption alignment. By explicitly coupling visual features with semantic guidance throughout the diffusion process, the model attends more effectively to informative image regions and generates captions with improved visual grounding and semantic fidelity.

Figure 3. Ablation results of visual–semantic conditioning visualized via CLIP-based Grad-CAM.

Conclusion and Future Work

In this paper, we proposed KE-diffusion, a diffusion-based image captioning framework that enhances semantic grounding by incorporating object-level visual–semantic conditioning into the reverse diffusion process. Unlike prior captioning methods that rely on global image embeddings or prefix-based conditioning, KE-diffusion injects compact visual–semantic condition vectors directly at the model level, allowing semantic cues to guide the entire denoising trajectory while preserving the parallel generation advantages of diffusion models.

Comprehensive experiments on MS-COCO and Flickr30k demonstrate that KE-diffusion consistently improves semantic fidelity and caption diversity compared with existing CLIP-based diffusion and autoregressive baselines. Cross-domain evaluations further show that the proposed conditioning strategy generalizes well across datasets with different visual distributions. In addition, visualization analyses provide qualitative evidence that KE-diffusion achieves stronger alignment between salient image regions and generated descriptions, supporting the effectiveness of model-level visual–semantic conditioning.

Several directions remain for future work. First, incorporating richer object interactions or relational cues may further enhance fine-grained semantic expressiveness. Second, extending KE-diffusion to controllable or user-guided caption generation could broaden its applicability in interactive vision–language systems. Finally, applying the proposed conditioning framework to other multimodal generation tasks, represents a promising avenue for future research.

Limitations

Despite its effectiveness, KE-diffusion has several limitations. First, the proposed method relies on object detection to construct visual–semantic condition vectors. Inaccurate or incomplete detections may introduce noise into the conditioning signals, potentially leading to missing or biased semantic guidance during the diffusion process. This limitation may be more pronounced in images with heavy occlusion, small objects, or abstract visual concepts that are challenging for standard detectors. In addition, the semantic embeddings are derived solely from object

category names, without explicitly modeling relational or commonsense knowledge, which may limit the model's ability to capture complex interactions among objects.

Second, KE-diffusion is evaluated primarily on English image captioning benchmarks. Its generalization to multilingual settings or culturally diverse visual domains remains unexplored. Moreover, although diffusion-based generation improves diversity, it typically incurs higher inference latency than strictly autoregressive lightweight models, which may constrain applicability in real-time or resource-limited scenarios. Finally, our evaluation mainly relies on automatic metrics and qualitative analyses, which may not fully reflect human preferences or downstream task performance.

References

Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, Antoine Miech, et al. 2022.
Flamingo: A visual language model for few-shot learning.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Anderson, Peter, Basura Fernando, Mark Johnson, and Stephen Gould. 2018.
Bottom-up and top-down attention for image captioning and visual question answering.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086.

Anderson, Peter, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2016.
SPICE: Semantic propositional image caption evaluation.
In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 382–398.

Austin, Jacob, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021.
Structured denoising diffusion models in discrete state spaces.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bengio, Samy, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015.
Scheduled sampling for sequence prediction with recurrent neural networks.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Denkowski, Michael, and Alon Lavie. 2014.
METEOR universal: Language specific translation evaluation for any target language.
In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380.

Dieleman, Sander, et al. 2022.
Diffusion models for text generation.
In *Proceedings of the ICML Workshop on Deep Generative Models*.

Donahue, Jeff, Lisa Anne Hendricks, Sergio Guadarrama, et al. 2015.
Long-term recurrent convolutional networks for visual recognition and description.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634.

Fang, Hao, Saurabh Gupta, Forrest Iandola, et al. 2015.
From captions to visual concepts and back.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1473–1482.

He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017.
Mask R-CNN.
In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel. 2020.
Denoising diffusion probabilistic models.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020.
The curious case of neural text degeneration.
In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Johnson, Justin, Andrej Karpathy, and Li Fei-Fei. 2016.
DenseCap: Fully convolutional localization networks for dense captioning.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4565–4574.

Karpathy, Andrej, and Li Fei-Fei. 2015.
Deep visual-semantic alignments for generating image descriptions.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137.

Kornblith, Simon, Ting Chen, and Mohammad Norouzi. 2023.
Specificity improves image captioning.
In *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 16063–16072.

Li, Jiwei, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016.
A diversity-promoting objective function for neural conversation models.
In *Proceedings of NAACL-HLT*, pages 110–119.

Li, Xiang Lisa, John D. McClelland, Yiming Wang, and Jason D. Lee. 2022.
Diffusion-LM improves controllable text generation.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Lin, Chin-Yew. 2004.
ROUGE: A package for automatic evaluation of summaries.
In *Text Summarization Branches Out*, pages 74–81.

Lin, Min, Qiang Chen, and Shuicheng Yan. 2013.
Network in network.
In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, et al. 2014.
Microsoft COCO: Common objects in context.
In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755.

Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024.
Prefix-diffusion: Lightweight diffusion-based image captioning.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mokady, Ron, Amir Hertz, and Amit Bermano. 2021.
CLIPCap: CLIP prefix for image captioning.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–868.

Nichol, Alex, and Prafulla Dhariwal. 2021.
Improved denoising diffusion probabilistic models.
In *Proceedings of the International Conference on Machine Learning (ICML)*.

Nukrai, Denis, Ron Mokady, and Amit Bermano. 2022.
CapDec: A diffusion-based image captioning model.
In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002.
BLEU: A method for automatic evaluation of machine translation.
In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Perez, Ethan, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018.
FiLM: Visual reasoning with a general conditioning layer.
In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Plummer, Bryan A., Liwei Wang, Chris M. Cervantes, et al. 2015.
Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models.
In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649.

Radford, Alec, Jong Wook Kim, Chris Hallacy, et al. 2021.
Learning transferable visual models from natural language supervision.
In *Proceedings of the International Conference on Machine Learning (ICML)*.

Ranzato, Marc'Aurelio, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016.
Sequence level training with recurrent neural networks.
In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. 2015.
Faster R-CNN: Towards real-time object detection with region proposal networks.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022.
High-resolution image synthesis with latent diffusion models.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Saharia, Chitwan, William Chan, Saurabh Saxena, et al. 2022.
Imagen: Text-to-image diffusion models.
In *Proceedings of the International Conference on Machine Learning (ICML)*.

Sohl-Dickstein, Jascha, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015.
Deep unsupervised learning using nonequilibrium thermodynamics.

In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2256–2265.

Strudel, Robin, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. 2023.
Self-conditioned diffusion models for text generation.
In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Tang, Zecheng, Xin Wang, and Yixin Zhu. 2024.
Composable diffusion for multimodal conditional generation.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. 2015.
CIDEr: Consensus-based image description evaluation.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015.
Show and tell: A neural image caption generator.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164.

Xu, Hu, Bing Liu, Lei Lei, and Wei Liu. 2023.
Versatile diffusion: Text, images and variations all in one diffusion model.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, Kelvin, Jimmy Ba, Ryan Kiros, et al. 2015.
Show, attend and tell: Neural image caption generation with visual attention.
In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2048–2057.

Yao, Ting, Yingwei Pan, Yehao Li, and Tao Mei. 2017.
Incorporating copying mechanism in image captioning for learning novel objects.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5263–5271.

Zhang, Yizhe, Michel Galley, Jianfeng Gao, et al. 2018.
Generating informative and diverse conversational responses via adversarial information maximization.
In *Advances in Neural Information Processing Systems (NeurIPS)*.