How Structured Knowledge Constrains Generative Diffusion Models for Domain Visual Description

Ronghui Liu[1,2] **Wenfeng Deng**[1,*] **Nan Zhou**[1] **Xiaojun Liang**[1] **Wei Cui**[2]
[1]Pengcheng Laboratory, Shenzhen, Guangdong, China
[2]South China University of Technology, Guangzhou, Guangdong, China
{liurh@pcl.ac.cn, dengwf@pcl.ac.cn, zhou@pcl.ac.cn, liangxj@pcl.ac.cn, aucuiwei@scut.edu.cn}
*Corresponding author

Abstract

Visual description is a fundamental yet challenging task that reflects how perceptual information is transformed into structured linguistic descriptions. Recent diffusion-based models have shown strong potential for parallel decoding and global semantic modeling. However, existing approaches largely treat generation as an unconstrained process, lacking mechanisms to incorporate structured knowledge, which often results in conceptually implausible or hallucinated descriptions, especially in domain-specific contexts. To investigate how structured knowledge constrains generative diffusion processes, we propose KenDiC, a **K**nowledge-**en**hanced **Di**ffusion-based **C**aptioner for visual description. KenDiC integrates a domain-adaptive visual encoder (DAVE) trained via contrastive learning to align perceptual and linguistic representations, and a domain term vocabulary (DTV) that constrains decoding to guide concept selection during generation. To support systematic analysis, we construct II2T-Bench, a domain-centric benchmark with expert annotations. Experimental results show that structured knowledge constraints significantly reduce hallucinations and improve semantic fidelity, suggesting a computational account of how knowledge guides generative visual description.

Keywords: Generative AI, Visual Description, Diffusion Models, Structured Knowledge

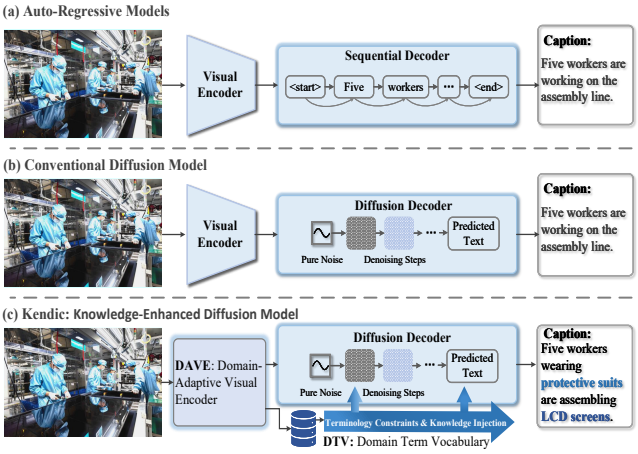

Figure 1: Comparison between (a) autoregressive image captioning models, (b) conventional diffusion-based captioning models, and (c) the proposed KenDiC framework. Unlike prior approaches, KenDiC incorporates domain-specific terminology constraints and contrastive fine-tuning of the visual encoder, enabling knowledge-enhanced latent diffusion and accurate generation of domain captions.

Introduction

Visual description refers to the process of transforming perceptual input into structured, conceptually grounded linguistic representations. In computational settings, this process is often instantiated as image-to-text generation at the intersection of computer vision and natural language processing (Y. Jiang et al., 2025). Beyond its practical applications in content summarization, intelligent monitoring, human–AI interaction, and assistive technologies (Ghandi et al., 2023; Żelaszczyk & Mańdziuk, 2023), visual description captures a core cognitive behavior in which prior knowledge and domain conventions constrain what aspects of a scene are described and how they are linguistically expressed.

Early computational models of visual description predominantly adopted encoder–decoder architectures, where convolutional neural networks extract visual features and recurrent neural networks decode them into textual sequences (Daneshfar et al., 2024; Ghandi et al., 2023; Zohourianshahzadi & Kalita, 2022). More recent approaches employ transformer-based autoregressive (AR) architectures that generate descriptions token by token while conditioning on previously produced outputs (Cornia et al., 2020; Fang et al., 2022; Y. Li et al., 2022). Although these models have achieved strong performance on general benchmarks, their sequential decoding paradigm limits parallelization, suffers from exposure bias, and often fails to capture global semantic structure, leading to reduced robustness and conceptual fidelity in complex visual scenes.

Diffusion models, particularly denoising diffusion probabilistic models (Ho et al., 2020), provide an alternative framework for modeling visual description as a global generative process. Their success in visual synthesis (Gao et al., 2023) and text generation (Wu et al., 2023) suggests strong capacity for parallel generation and holistic semantic planning. Recent studies (He et al., 2023; X. Li et al., 2022; Liu et al., 2024) have applied diffusion models to image-to-text generation by treating the diffusion process as a text decoder within an encoder–decoder architecture. While these methods mitigate exposure bias and improve sentence-level coherence, they largely model description as an unconstrained stochastic process, overlooking how structured knowledge guides concept selection. As a result, diffusion-based models remain prone to semantically plausible yet conceptually incorrect or hallu-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

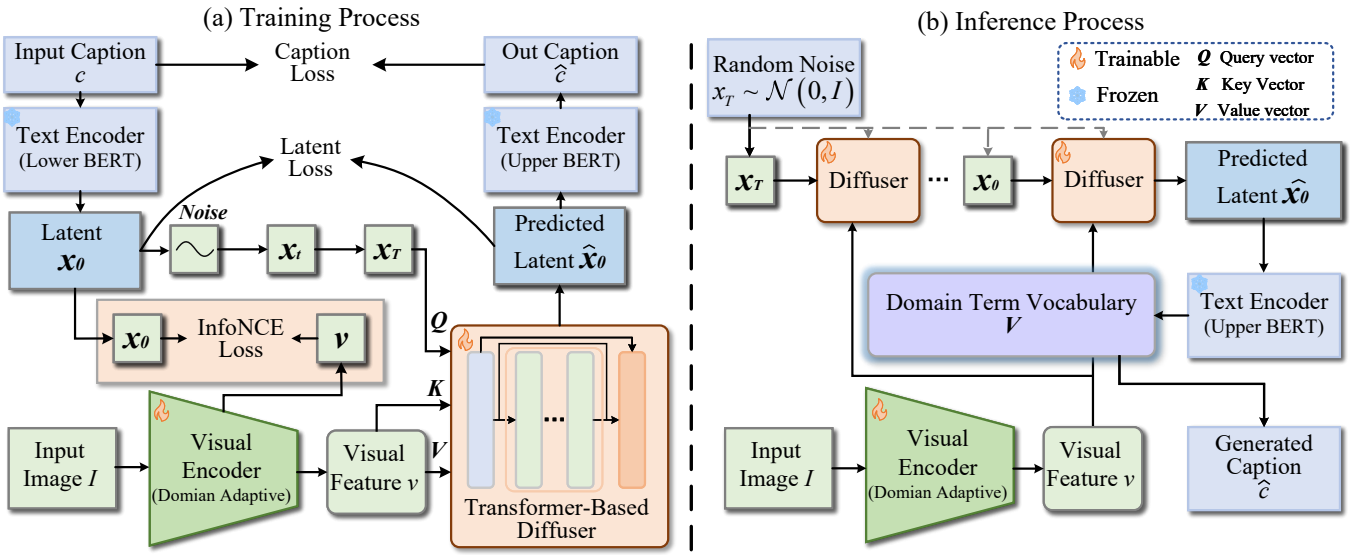

Figure 2: An overview of KenDiC. It consists of an image encoder, a text encoder, a diffuser, and a text decoder. Panel (a) shows the training process, and panel (b) shows the inference process.

cinated descriptions, particularly in domain-specific contexts (Y. Huang et al., 2025).

In this work, we examine how structured knowledge constrains generative diffusion models for visual description. We propose **KenDiC**, a **K**nowledge-**en**hanced **Di**ffusion-based **C**aptioner that explicitly integrates domain knowledge into the diffusion-based generation process. As illustrated in Figure 1, KenDiC incorporates two complementary mechanisms. First, a Domain-Adaptive Visual Encoder (DAVE) aligns perceptual representations with domain-relevant linguistic concepts via contrastive learning, stabilizing latent semantic recovery during diffusion denoising. Second, a Domain Term Vocabulary (DTV) constrains decoding at the terminology level, guiding concept realization and reducing domain-specific hallucination without sacrificing generative flexibility. To support systematic evaluation, we further construct II2T-Bench, a domain-centric image-to-text benchmark. Extensive experiments demonstrate that structured knowledge constraints lead to more faithful and reliable visual descriptions, offering a computational account of how knowledge guides generative visual language.

Related Work

Visual Description and Image-to-Text Generation

Visual description, commonly instantiated as image-to-text generation, aims to produce natural language descriptions grounded in visual content (Żelaszczyk & Mańdziuk, 2023). Early approaches relied on template-based (Kojima et al., 2002) and retrieval-based methods (Hodosh et al., 2013), which offered limited flexibility and conceptual generalization. With the advent of deep learning, autoregressive encoder–decoder architectures became dominant, sequentially mapping visual features to textual outputs (Karpathy & Fei-Fei, 2015; Vinyals et al., 2015).

Subsequent advances introduced attention mechanisms (Anderson et al., 2018; You et al., 2016; Zohourianshahzadi & Kalita, 2022) and transformer-based architectures with large-scale vision–language pre-training (L. Huang et al., 2019; Zhou et al., 2020), substantially improving fluency and visual grounding. Nevertheless, these approaches largely treat visual description as an unconstrained generation process. In domain-specific settings, where accurate concept selection and terminology usage are critical, such models often struggle to maintain semantic precision, highlighting the need for mechanisms that explicitly constrain generation with structured knowledge.

Diffusion-Based Visual Description

Diffusion models frame generation as an iterative denoising process and have emerged as a powerful paradigm for multimodal generation (Croitoru et al., 2023; Ho et al., 2020). When applied to visual description, diffusion-based models enable parallel decoding and global semantic planning, alleviating limitations of autoregressive generation. Early work explored conditioning diffusion on pre-trained vision–language representations, such as CLIP-Diffusion-LM (S. Xu, 2022), while later studies extended diffusion frameworks to unified multimodal generation, including Versatile Diffusion (X. Xu et al., 2023) and CoDi (Tang et al., 2023). To improve semantic alignment, recent methods incorporate explicit conditioning or data augmentation strategies (D. Jiang et al., 2024; Luo et al., 2023; Ma et al., 2024). Notably, LaDiC (Wang et al., 2024) demonstrated that latent diffusion models can achieve competitive performance for image captioning. However, most diffusion-based approaches model generation as a largely unconstrained stochastic process and are trained on

general-purpose data. In domain-centric visual description, the absence of structured knowledge constraints often leads to semantic drift and hallucinated concepts, motivating the development of knowledge-constrained diffusion models that explicitly guide concept selection during generation.

Methods

As shown in Figure 2, KenDiC models visual description as a knowledge-constrained generative diffusion process. It consists of four components: a text encoder, a visual encoder, a transformer-based diffuser, and a text decoder. During training, a caption is encoded into a latent semantic representation x_0 using the lower layers of a split BERT model and progressively noised into x_T via a forward Gaussian process, where x_t denotes the latent variable at diffusion step t, and x_T is the terminal noisy state. In parallel, the input image is mapped to a visual representation $v = F_v(I)$ by a domain-adaptive visual encoder. The diffuser f_θ learns to recover $\hat{x}_0$ from x_t conditioned on v, which is then decoded into text by the upper BERT layers. At inference time, generation starts from Gaussian noise $x_T \sim \mathcal{N}(0, I)$ and iteratively denoises to $\hat{x}_0$, enabling non-autoregressive visual description.

The total training objective jointly minimizes the latent loss and caption loss as follows:

$$\mathcal{L} = \|f_\theta(x_t, v, t) - x_0\|^2 + \lambda \cdot \mathrm{CE}(\hat{c}, c) \tag{1}$$

Latent Semantic Representation

Following LaDiC (Wang et al., 2024), we adopt a 12-layer split-BERT architecture, with layers 1–6 serving as the text encoder and layers 7–12 as the decoder. This design provides a stable latent semantic space for diffusion-based generation while preserving decoding fluency. Given a caption c, the encoder produces a latent representation $x_0 = \mathcal{E}_{\text{text}}(c)$.

During denoising, latent evolution is conditioned on both visual input and terminology-level semantic cues:

$$x_{t-1} = \mu_\theta(x_t, t, F_v(I), F_t(c)) + \Sigma_\theta(x_t, t) \cdot \epsilon, \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $F_t(c)$ denotes terminology embeddings derived from the ground-truth caption during training. The recovered latent $\hat{x}_0$ is decoded into natural language by $\mathcal{D}_{\text{text}}$. This formulation allows the diffusion process to operate over a conceptually structured latent space rather than unconstrained surface text.

Domain-Adaptive Visual Encoder

To align perceptual representations with domain-relevant concepts, we introduce a DAVE trained via self-supervised contrastive learning on image–caption pairs. Given a mini-batch $\{(I_i, c_i)\}_{i=1}^N$, matched image–caption pairs are treated as positives, while mismatched pairs serve as negatives, enabling domain-aware alignment without additional supervision.

The visual encoder $F_v(\cdot)$ maps each image I_i to an embedding v_i, and the corresponding caption c_i is encoded into a textual embedding t_i using a frozen text encoder. A similarity matrix $S \in \mathbb{R}^{N \times N}$ is computed using scaled cosine similarity:

$$S_{ij} = \frac{v_i^\top t_j}{\tau}, \tag{3}$$

where τ is a temperature parameter.

We optimize a symmetric InfoNCE objective:

$$\mathcal{L}_{\text{CL}} = -\frac{1}{2}\left(\log \frac{\exp(S_{ii})}{\sum_j \exp(S_{ij})} + \log \frac{\exp(S_{ii})}{\sum_j \exp(S_{ji})}\right), \tag{4}$$

Domain Term Vocabulary Augmentation

To examine how structured knowledge constrains generative diffusion for domain visual description, we introduce a domain term vocabulary that guides concept selection during decoding. In domain-specific visual description, accurate interpretation depends not only on perceptual input but also on domain conventions that determine which concepts are valid and how they should be expressed. Without such constraints, generative models often produce visually plausible yet conceptually invalid descriptions.

We define the domain vocabulary as $V = \{t_1, t_2, \ldots, t_K\}$, constructed from domain standards, reference documents, and expert-annotated corpora. The vocabulary covers core domain concepts, including equipment, actions, and procedures, and serves as an explicit representation of structured domain knowledge.

During decoding, token probabilities are computed from the decoder hidden state h_i:

$$p(y_i = w \mid h_i) = \frac{\exp(v_w^\top h_i)}{\sum_u \exp(v_u^\top h_i)}, \tag{5}$$

where v_w denotes the embedding of token w. The vocabulary constraint is applied at the logit level prior to the softmax operation, rather than being fed into the decoder as an input. To constrain linguistic realization toward domain-consistent concepts, tokens belonging to V are assigned a higher preference prior to normalization:

$$\widetilde{v_w^\top h_i} = \begin{cases} \beta \cdot v_w^\top h_i, & w \in V, \\ v_w^\top h_i, & \text{otherwise}, \end{cases} \tag{6}$$

with $\beta > 1$ controlling the strength of the constraint.

This constraint operates exclusively at decoding time and does not alter model training. By restricting the space of plausible concepts during generation, the domain vocabulary constrains the diffusion-based generative process for domain visual description, reducing hallucination while preserving fluency and contextual coherence. While the proposed vocabulary constraint is implemented as a simple logit reweighting, its role is not to enforce hard selection but to bias the generative process toward domain-consistent concepts.

Results and Analysis

Experimental Settings

Dataset: To evaluate domain visual description under structured knowledge constraints, we construct II2T-Bench, a

Table 1: Comparative performance of different models.

Models	Evaluation Metrics					
	BLEU@1	BLEU@4	CIDEr	METEOR	ROUGE-L	BERT-S
Autoregressive-based Models						
CLIPCap (Mokady et al., 2021)	26.2	1.5	2.1	10.3	16.7	54.4
ViTCap (Fang et al., 2022)	34.3	4.8	6.8	5.6	18.7	32.6
BLIP (J. Li et al., 2022)	56.3	15.8	31.7	18.7	35.0	68.1
Diffusion-based Models						
DDCap (Zhu et al., 2022)	49.3	11.6	26.0	16.9	32.4	65.2
Prefix-diffusion (Liu et al., 2024)	55.8	15.6	**36.0**	18.5	34.6	67.8
LaDiC (Wang et al., 2024)	57.8	15.2	28.1	17.2	33.6	65.7
KenDiC (Ours)	**62.5**	**19.3**	32.3	**20.2**	**39.5**	**71.2**

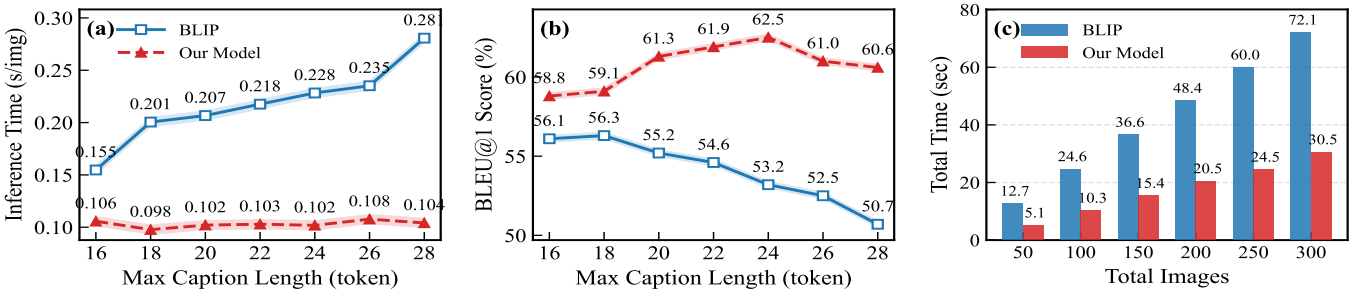

Figure 3: Performance comparison between BLIP and KenDiC.

domain-centric image-to-text benchmark collected from real industrial environments. From over 5,000 images, 3,300 representative samples are retained after filtering ambiguous scenes, severe occlusion, and insufficient visibility of key entities. Each image is annotated with five expert-written descriptions following a four-dimensional schema covering scene, personnel, actions, and equipment, yielding 16,500 captions. The dataset is split into training, validation, and test sets with an 8:1:1 ratio at the image level, ensuring no scene overlap. II2T-Bench spans diverse industrial contexts, including metallurgical processing, equipment operation, safety inspection, and material handling. The dataset will be publicly released upon acceptance.

Evaluation Metrics: We assess description quality using standard automatic metrics, including BLEU-1 and BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee & Lavie, 2005), and CIDEr (Vedantam et al., 2015), which quantify content accuracy and indirectly reflect hallucinated or incorrect entities. To evaluate semantic alignment beyond surface-level overlap, we additionally report BERTScore (BERT-S) (Zhang et al., 2019), measuring contextual similarity in embedding space. Although standard metrics such as BLEU, ROUGE, METEOR, and CIDEr do not explicitly measure hallucination, they indirectly reflect semantic correctness through agreement with expert references. In our domain setting, the use of expert-annotated captions reduces ambiguity, making these metrics more indicative of factual consistency.

Implementation Details: KenDiC adopts a split-BERT encoder–decoder architecture (6+6 layers) following LaDiC (Wang et al., 2024), with $\text{BLIP}_{\text{base}}$ ViT-B/16 as the visual backbone. The diffuser consists of 12 Transformer blocks with cross-attention. Diffusion runs for $T = 1500$ steps using cosine noise scheduling, and deterministic DDIM sampling is applied at inference. Models are trained for 300 epochs with Adam (learning rate 1×10^{-4}, batch size 48) on two RTX 3090 GPUs, with gradient clipping (norm 1.0) and $\lambda = 0.5$ to balance denoising and caption losses.

Comparative Study

We compare KenDiC with representative autoregressive (AR) image captioning models, including CLIPCap (Mokady et al., 2021), ViTCap (Fang et al., 2022), and BLIP (J. Li et al., 2022), as well as diffusion-based approaches such as DDCap (Zhu et al., 2022), Prefix-diffusion (Liu et al., 2024), and LaDiC (Wang et al., 2024). All baselines were reimplemented and trained on II2T-Bench using identical data splits and training protocols to ensure fair comparison.

As reported in Table 1, KenDiC achieves the strongest overall performance across most metrics. Compared with the strongest AR baseline (BLIP), KenDiC improves BLEU@4, METEOR, and ROUGE-L, indicating more precise and se-

BLIP: A woker is waking on the ground.
LaDiC: A worker operates in flames and smoke wearing protective gear.
Ours: A worker in protective clothing is crossing over an iron trench.

BLIP: A stream of light flows across a dirty ground.
LaDiC: An orange line appears along the ground surface.
Ours: Molten slag flows along an iron trench on the factory floor.

BLIP: A worker is cutting metal with sparks flying.
LaDiC: A worker is grinding steel with an angle grinder.
Ours: A worker cutting steel with an angle grinder without protective gear.

BLIP: Two gas cylinders lying on a factory floor.
LaDiC: Two industrial cylinders, blue and yellow, placed on a grimy ground.
Ours: Two industrial gas cylinders on a dirty floor, with inverted oxygen bottle.

Figure 4: Qualitative comparison of captions generated by different models.

Table 2: Ablation results for the proposed KenDiC.

Configs	Evaluation Metrics					
	BLEU@1	BLEU@4	CIDEr	Meteor	ROUGE-L	BERT-S
Baseline	57.8	15.2	28.1	17.2	33.6	65.7
Baseline+DTV	59.4	16.8	29.6	17.5	35.2	68.2
Baseline+DTV+DAVE	62.5	19.3	32.3	20.2	39.5	71.2

mantically faithful domain visual descriptions. While Prefix-Diffusion attains the highest CIDEr score, KenDiC yields more balanced improvements across lexical and semantic metrics, including the best BERTScore, reflecting stronger alignment between generated descriptions and domain-relevant concepts. These results suggest that explicitly constraining diffusion with structured knowledge improves domain visual description beyond unconstrained autoregressive and diffusion-based generation.

Performance Analysis

We further analyze KenDiC in terms of efficiency, robustness, and descriptive precision. As shown in Figure 3, KenDiC maintains stable inference time as caption length increases, whereas the decoding time of BLIP grows substantially, reflecting the advantage of non-autoregressive diffusion for scalable visual description. Moreover, KenDiC preserves higher BLEU@1 scores under longer descriptions, indicating reduced semantic drift during generation. Across increasing test set sizes, KenDiC consistently achieves lower total inference time, demonstrating scalability for domain settings.

Qualitative examples in Figure 4 illustrate how structured knowledge constrains generative diffusion at the conceptual level. Compared with BLIP (J. Li et al., 2022) and LaDiC (Wang et al., 2024), KenDiC produces more specific and accurate domain visual descriptions, correctly identifying critical elements such as equipment, actions, and safety-related details. In contrast, baseline models often generate generic or underspecified phrases. These examples qualitatively confirm that knowledge constraints guide concept selection rather than merely improving surface fluency.

Ablation Study

We conduct an ablation study to isolate the effects of structured knowledge constraints. As shown in Table 2, the baseline model without knowledge components achieves reasonable performance, reflecting general image–text alignment. Introducing the domain term vocabulary (DTV) consistently improves all metrics, indicating that constraining concept selection during decoding reduces domain-specific hallucination. Further incorporating the domain-adaptive visual encoder (DAVE) yields the best performance, suggesting that aligning perceptual representations with domain-relevant concepts complements linguistic constraints. Together, these results demonstrate that domain visual description benefits from both perceptual alignment and explicit conceptual constraints within the generative diffusion process.

Conclusion

This work investigated how structured knowledge constrains generative diffusion models for domain visual description. We proposed KenDiC, a knowledge-constrained diffusion framework that models visual description as a generative process guided by domain-specific conceptual structure rather than unconstrained text production. By integrating a domain-adaptive visual encoder that aligns perceptual representations with domain-relevant concepts, and a terminology-guided decoding mechanism that constrains concept selection during linguistic realization, KenDiC explicitly incorporates structured knowledge into the diffusion-based generation process.

Experiments on II2T-Bench demonstrate that such constraints significantly reduce semantic drift and domain-specific hallucination while preserving fluency and descriptive accuracy. Beyond empirical performance, these results support a computational account of domain visual description in which structured knowledge functions as a constraint on generative diffusion, shaping both the space of plausible concepts and their linguistic expression. This perspective aligns with cognitive theories of constrained language production and suggests that effective domain visual description emerges from the interaction between perceptual representations, generative processes, and structured conceptual knowledge.

Acknowledgments

This work was supported in part by the Guangdong S&T Programme (Grant No. 2024B0101010003), and in part by the Major Key Project of Pengcheng Laboratory (PCL) (Grant No. PCL2025A13 and No. PCL2025A12).

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086. https://doi.org/10.48550/arXiv.1707.07998

Banerjee, S., & Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72. https://doi.org/10.5555/1626355.1626389

Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-memory transformer for image captioning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10578–10587. https://doi.org/10.1109/cvpr42600.2020.01059

Croitoru, F.-A., Hondru, V., Ionescu, R. T., & Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10850–10869. https://doi.org/10.1109/TPAMI.2023.3261988

Daneshfar, F., Bartani, A., & Lotfi, P. (2024). Image captioning by diffusion models: A survey. *Engineering Applications of Artificial Intelligence*, *138*, 109288. https://doi.org/10.1016/j.engappai.2024.109288

Fang, Z., Wang, J., Hu, X., Liang, L., Gan, Z., Wang, L., Yang, Y., & Liu, Z. (2022). Injecting semantic concepts into end-to-end image captioning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18009–18019. https://doi.org/Injectingsemanticconceptsintoend-to-endimagecaptioning

Gao, S., Liu, X., Zeng, B., Xu, S., Li, Y., Luo, X., Liu, J., Zhen, X., & Zhang, B. (2023). Implicit diffusion models for continuous super-resolution. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10021–10030. https://doi.org/10.48550/arXiv.2303.16491

Ghandi, T., Pourreza, H., & Mahyar, H. (2023). Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, *56*(3), 1–39. https://doi.org/10.1145/3617592

He, Y., Cai, Z., Gan, X., & Chang, B. (2023). Diffcap: Exploring continuous diffusion on image captioning. *arXiv preprint arXiv:2305.12144*. https://doi.org/10.48550/arXiv.2305.12144

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840–6851. https://doi.org/10.48550/arXiv.2006.11239

Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, *47*, 853–899. https://doi.org/10.1613/jair.3994

Huang, L., Wang, W., Chen, J., & Wei, X.-Y. (2019). Attention on attention for image captioning. *Proceedings of the IEEE/CVF international conference on computer vision*, 4634–4643. https://doi.org/10.48550/arXiv.1707.07998v3

Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Cao, L., & Chen, S. (2025). Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.48550/arXiv.2402.17525

Jiang, D., Song, G., Wu, X., Zhang, R., Shen, D., Zong, Z., Liu, Y., & Li, H. (2024). Comat: Aligning text-to-image diffusion model with image-to-text concept matching. *Advances in Neural Information Processing Systems*, *37*, 76177–76209. https://doi.org/10.48550/arXiv.2404.03653

Jiang, Y., Dale, R., & Lu, H. (2025). Understanding human heuristics in context-sensitive image captioning. https://escholarship.org/uc/item/6x4796vt

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3128–3137. https://doi.org/10.1109/cvpr.2015.7298932

Kojima, A., Tamura, T., & Fukunaga, K. (2002). Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, *50*, 171–184. https://doi.org/10.1023/A:1020346032608

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *International conference on machine learning*, 12888–12900. https://doi.org/10.48550/arXiv.2201.12086

Li, X., Thickstun, J., Gulrajani, I., Liang, P. S., & Hashimoto, T. B. (2022). Diffusion-lm improves controllable text gener-

ation. *Advances in neural information processing systems*, *35*, 4328–4343. 10.48550/arXiv.2205.14217

Li, Y., Pan, Y., Yao, T., & Mei, T. (2022). Comprehending and ordering semantics for image captioning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17990–17999. https://doi.org/10.48550/arXiv.2206.06930

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text summarization branches out*, 74–81. https://aclanthology.org/W04-1013.pdf

Liu, G., Li, Y., Fei, Z., Fu, H., Luo, X., & Guo, Y. (2024). Prefix-diffusion: A lightweight diffusion model for diverse image captioning. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 12954–12965. https://doi.org/10.48550/arXiv.2309.04965

Luo, J., Li, Y., Pan, Y., Yao, T., Feng, J., Chao, H., & Mei, T. (2023). Semantic-conditional diffusion networks for image captioning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 23359–23368. https://doi.org/10.48550/arXiv.2212.03099

Ma, F., Zhou, Y., Rao, F., Zhang, Y., & Sun, X. (2024). Image captioning with multi-context synthetic data. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*, 4089–4097. https://doi.org/10.1609/aaai.v38i5.28203

Mokady, R., Hertz, A., & Bermano, A. H. (2021). Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*. https://doi.org/10.48550/arXiv.2111.09734

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Tang, Z., Yang, Z., Zhu, C., Zeng, M., & Bansal, M. (2023). Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, *36*, 16083–16099. https://doi.org/10.48550/arXiv.2305.11846

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575. https://doi.org/10.1109/cvpr.2015.7299087

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164. https://doi.org/10.1109/cvpr.2015.7298935

Wang, Y., Ren, S., Gao, R., Yao, L., Guo, Q., An, K., Bai, J., & Sun, X. (2024). Ladic: Are diffusion models really inferior to autoregressive counterparts for image-to-text generation? *arXiv preprint arXiv:2404.10763*. https://doi.org/10.48550/arXiv.2404.10763

Wu, T., Fan, Z., Liu, X., Zheng, H.-T., Gong, Y., Jiao, J., Li, J., Guo, J., Duan, N., Chen, W., et al. (2023). Ar-diffusion: Auto-regressive diffusion model for text generation. *Advances in Neural Information Processing Systems*, *36*, 39957–39974. https://doi.org/10.48550/arXiv.2305.09515

Xu, S. (2022). Clip-diffusion-lm: Apply diffusion model on image captioning. *arXiv preprint arXiv:2210.04559*. https://doi.org/10.48550/arXiv.2210.04559

Xu, X., Wang, Z., Zhang, G., Wang, K., & Shi, H. (2023). Versatile diffusion: Text, images and variations all in one diffusion model. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7754–7765. https://doi.org/10.48550/arXiv.2211.08332

You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4651–4659. https://doi.org/10.48550/arXiv.1603.03925

Żelaszczyk, M., & Mańdziuk, J. (2023). Cross-modal text and visual generation: A systematic review. part 1: Image to text. *Information Fusion*, *93*, 302–329. https://doi.org/10.1016/j.inffus.2023.01.008

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*. https://doi.org/10.48550/arXiv.1904.09675

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., & Gao, J. (2020). Unified vision-language pre-training for image captioning and vqa. *Proceedings of the AAAI conference on artificial intelligence*, *34*, 13041–13049. https://doi.org/10.48550/arXiv.1909.11059

Zhu, Z., Wei, Y., Wang, J., Gan, Z., Zhang, Z., Wang, L., Hua, G., Wang, L., Liu, Z., & Hu, H. (2022). Exploring discrete diffusion models for image captioning. *arXiv preprint arXiv:2211.11694*. https://doi.org/10.48550/arXiv.2211.11694

Zohourianshahzadi, Z., & Kalita, J. K. (2022). Neural attention for image captioning: Review of outstanding methods. *Artificial Intelligence Review*, *55*(5), 3833–3862. https://doi.org/10.1007/s10462-021-10092-2