Deep-learning-driven Cross-modal Image Fusion For Cartoon Captioning

Lijuan Feng, Jiangjiang Li, Yachao Zhang, and *Yandong Han

School of Electronics and Electrical Engineering, Zhengzhou University of Science and Technology, Zhengzhou 450064 China

Corresponding author. E-mail: yandonghanzust@163.com

Received: Jun. 25, 2025; Accepted: Aug. 15, 2025

Deep learning has revolutionized cross-modal understanding through its ability to process heterogeneous visual-textual data, which achieves encouraging performance in the image captioning domain. However, existing methods struggle with redundant cross-modal representations and autoregressive generation bottlenecks, leading to semantically misaligned captions. To address these challenges, an innovative data-driven cartoon image captioning framework (IBA-CD) is proposed through two synergistic advancements. First, IBA-CD develops an information bottleneck-driven cross-modal alignment mechanism that fusing principles from information theory and deep learning, to optimize semantic distillation. Such a mechanism suppresses redundant visual-textual mutual information while maximizing task-relevant correlations through variational inference. Second, IBA-CD pioneers a cascaded diffusion generation paradigm that reimagines text synthesis through bidirectional Transformer-based denoising processes, establishing non-autoregressive generation with multi-stage visual-semantic refinement. IBA-CD achieves component synergy through a bidirectional closed-loop mechanism: The information bottleneck alignment dynamically injects distilled compact semantic features as conditional guidance into each denoising step of the diffusion network, enabling fine-grained visual-textual alignment through cross-modal attention mechanisms. Simultaneously, quality feedback from the generation process proactively optimizes the feature alignment intensity, forming an iterative refinement cycle that evolves from semantic compression to generation correction. This collaborative framework ultimately accomplishes efficient and precise cross-modal reasoning through tightly coupled visual-semantic distillation and progressive generative enhancement. Extensive experiments on benchmark datasets verify significant improvements in the cartoon image captioning task.

Keywords: Image captioning; information bottleneck alignment; cascaded diffusion network

http://dx.doi.org/10.6180/jase.202605_29(5).0016

1. Introduction

As an emerging branch of multi-modal tasks, image captioning is gaining increasing attention, whose goal is to generate natural language sentences to describe image content [1, 2]. Inspired by deep learning's success in machine translation, most image captioning models adopt an encoder-decoder architecture. The encoder, often a convolutional neural network (CNN), extracts visual features from images, while the decoder, typically a recurrent neural network (RNN) or long short-term memory (LSTM) network, generates corresponding textual descriptions [3, 4]. Early work used CNNs as encoders and RNNs/LSTMs as decoders. To better focus on relevant image regions, Xu et al. used target detectors like Fast-RCNN and introduced a top-down attention mechanism [5]. Nguyen et al. enhanced this by adding a previous-moment attention vector to ensure visual coherence and proposed a dual-attention mechanism to capture more comprehensive semantic information [6].

Inspired by Transformer and BERT language models,

many scholars are studying Transformer-based image captioning models to better capture the relationship between image features and generated sequences [7, 8]. For instance, Luo et al. proposed masked non-auto-regressive decoding for generating more semantic and diverse annotations in parallel [9]. Zhao et al. developed a semi-auto-regressive Transformer that predicts groups of words in parallel and generates them from left to right, better balancing captioning speed and quality [10]. With diffusion models achieving success in image generation, they are becoming a new research direction for text generation. Unlike typical one-off discrete sentence generation, diffusion processes are parameterized Markov chains that gradually add Gaussian noise to sentences. By learning each reverse state transition, original sentences can be restored from noisy data for text generation [11]. Luo et al. proposed encoding discrete text into bits and using self-conditioning diffusion models for annotation generation [12]. Liu et al. naturally transformed discrete markers and applied continuous diffusion on them, successfully integrating extracted image features for diffusion-based text generation [13]. However, most diffusion-based image captioning methods have issues like generating text mismatched with input images, lacking effective semantic control, and having word repetition/omission. There's also the problem of insufficient training due to inadequate noise.

This study introduces an information bottleneck-driven cross-modal alignment and cascaded diffusion network (IBA-CD) for cartoon image captioning, addressing the challenge of generating accurate textual descriptions from visual inputs. The framework comprises two key components: (1) A cross-modal alignment module that employs Transformers to extract image-text representations, enhanced by an information bottleneck mechanism to eliminate redundancy while maximizing semantic correlation. This is achieved through a novel loss function minimizing mutual information between representations and raw data, while maximizing cross-modal mutual information using variational optimization and Monte Carlo approximation. (2) A cascaded diffusion network featuring multiple Transformer-based diffusion models that progressively denoise text sequences through bidirectional semantic guidance, enabling non-autoregressive generation with enhanced coherence. The architecture utilizes cross-attention layers to integrate aligned visual features during the reverse denoising process, supervised by a hybrid loss combining cross-entropy captioning loss, diffusion reconstruction loss, and the information bottleneck alignment loss. Experimental validation demonstrates significant improvements in caption quality through this dual optimization of compact semantic alignment and iterative diffusion-based refinement.

The key contributions of IBA-CD include:

- We introduce an information bottleneck-driven cross-modal alignment mechanism to enhance cartoon-text semantic correlation through variational mutual information optimization and redundancy-aware compression.
- We propose a cascaded diffusion generation architecture enabling non-autoregressive, visually grounded text synthesis via bidirectional diffusion processes and multi-stage Transformer-based iterative refinement
- Extensive experiments conducted two datasets verify that IBA-CD shows a new baseline in the cartoon image captioning task.

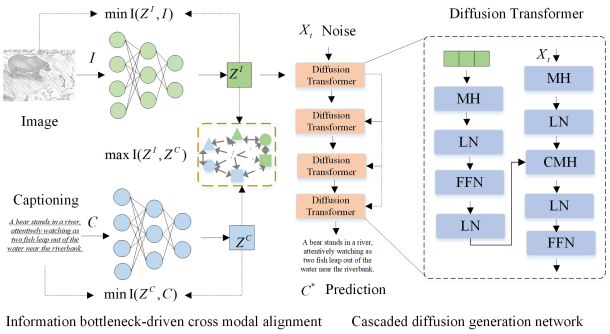

Fig. 1. The overall architecture of IBA-CD. It contains an information bottleneck-driven cross modal alignment and a cascaded diffusion generation network.

2. Method

Cartoon image captioning is a cross-modal generation task that aims to map visual information to natural language descriptions. It seeks to transform a cartoon image I into an accurate and meaningful textual description C through a mathematical model $f : I \rightarrow C$, where f is typically implemented by deep learning algorithms to maximize the accuracy and relevance of the description. To achieve the above task, we propose an information bottleneck-driven cross-modal alignment and cascaded diffusion network (IBA-CD), as shown in Fig. 1.

2.1. Information Bottleneck-driven Cross-modal Alignment

Information bottleneck-driven cross modal alignment (IBCMA) aims to achieve accurate semantic alignments between cartoon and text representations. To this end,

IBCMA utilizes image Transformer and text Transformer to conduct a multi-modal feature extractor for extract the corresponding representations. meanwhile, to eliminate redundant information between modalities, IBCMA designs cross modal information bottleneck regularization terms to minimize the mutual information between data and representations in each modality and maximize the mutual information between cartoon and text representations

Given a cartoon and text pair (I, C), cartoon and text representations are obtained via the multi-modal feature extractor:

$$\begin{aligned} Z^I &= Transformer^I(I) \\ Z^C &= Transformer^C(C) \end{aligned} \tag{1}$$

where Z^I and Z^C denote cartoon and text representations, respectively. Then, IBCMA introduces information bottleneck to refine representations:

$$L_{IBCAM} = I(Z^I, I) + I(Z^C, C) - \alpha I(Z^I, Z^C) \tag{2}$$

where α denotes a trade-off parament. In the context of Cartoon image captioning, the information bottleneck loss L_{IBCAM} plays a crucial role in optimizing the representations for better performance. By minimizing the first term $I(Z^I, I)$, we aim to ensure that the cartoon representation Z^I captures only the essential and relevant information from the cartoon image I, discarding any redundant or noisy details that might be present in the original image data. Similarly, minimizing the second term $I(Z^C, C)$ helps to refine the text representation Z^C so that it focuses solely on the key aspects of the caption C, eliminating unnecessary or irrelevant textual elements. This process of minimizing the mutual information between the representations and their original sources effectively compresses the information, making the representations more concise and focused on the most important features for the task. On the other hand, maximizing the third term $I(Z^I, Z^C)$ encourages a strong and meaningful correlation between the cartoon and text representations. This is crucial for Cartoon image captioning because it ensures that the refined representations of the cartoon image and its corresponding caption are closely aligned and can effectively communicate with each other. By maximizing this mutual information, the model is able to learn a joint representation space where the cartoon and text are well-matched, leading to more accurate and coherent captions that truly reflect the content of the cartoon image.

To obtain numerical solution of L_{IBCAM}, a variational optimization method **[30, 31, 33]** is designed to compute the mutual information. Specifically, The derived process of the first term is as follows:

$$I(I; Z^I) = \int_{z^I}\int_i p(i, z^I) \log \frac{p(i, z^I)}{p(i)p(z^I)} = \int_{z^I}\int_i p(i, z^I) \log \frac{p(i \mid z^I)}{p(i)} \tag{3}$$

where $p(i, z^I)$, $p(i)$, and $p(z^I)$ denote the joint probability density function and the marginal probability density functions, respectively. Since the difficulty in solving the posterior probability distribution $p(i \mid z^I)$, a variational estimate $q(i)$ of $p(i)$ is used based on the variational inference for approximating $p(i \mid z^I)$. Since the KL divergence is non-negative, having:

$$KL(p(i) \| q(i)) > 0 \Rightarrow \int p(i) \log p(i) > \int p(i) \log q(i) \Rightarrow p(i) > q(i) \tag{4}$$

Now, $I(I; Z^I)$ can be rewritten as follows:

$$I(I; Z^I) = \iint p(i, z^I) \log \frac{p(i \mid z^I)}{p(i)} < \iint p(i, z^I) \log \frac{p(i \mid z^I)}{q(i)}. \tag{5}$$

According to the Bayesian theory $p(i, z^I) = p(z^I)p(i \mid z^I)$, $I(I; Z^I)$ can be optimized as:

$$I(I; Z^I) < \iint p(z^I)p(I \mid z^I) \log \frac{p(i \mid z^I)}{q(i)} \tag{6}$$

Then, the Monte Carlo sampling strategy is used to approximate $p(z^I)$ to eliminate extraneous elements, having:

$$I(I; Z^I) < \int p(i \mid z^I) \log \frac{p(i \mid z^I)}{q(i)}. \tag{7}$$

For simplicity, considering $p(i \mid z^I)$ as a Gaussian distribution with the mean μ and variances σ, we reparameterize z^I as:

$$z^I = \mu(i) + \sigma(i). \tag{8}$$

Next, $I(I; Z^I)$ can be expressed as:

$$I(I; Z^I) \approx \mathbb{E}_\theta\{KL[p(i \mid z^I) \| q(i)]\} \tag{9}$$

where θ denotes the standard normal distribution. Similarity, $I(C; Z^C)$ can be expressed as:

$$I(C; Z^C) \approx \mathbb{E}_\theta\{KL[p(c \mid z^C) \| q(c)]\} \tag{10}$$

For the optimization of the third term $I(Z^C, Z^I)$, the joint probability is first computed as follows:

$$p(z^C, z^I) = \frac{1}{2}\left(\sum_{n=1}^{bn} z_n^C \times (z_n^I)^T + \sum_{n=1}^{bn} z_n^I \times (z_n^C)^T\right) \tag{11}$$

where bn is the batchsize number. Then, the marginal probabilities of $p(z^C)$ and $p(z^T)$ are calculated to obtain $I(Z^C, Z^I)$

$$I(Z^C, Z^I) = p(z^C, z^I) \log\left(\frac{p(z^C, z^I)}{p(z^C)p(z^I)}\right) \tag{12}$$

2.2. Cascaded Diffusion Generation Network

Cascaded diffusion generation network aims at implementing bidirectional text information transfer and generating all words simultaneously using the diffusion model paradigm in a non-autoregressive approach, inspired by the successful application of cascaded diffusion models in image generation, involves stacking multiple diffusion models in a cascade. Each diffusion model diffuses based on the output of the previous diffusion model. Each diffusion model layer takes the output of the previous layer and semantic alignment features as input, using the semantic alignment features as semantic conditions to control the diffusion process, and passes the results of each layer and control conditions through layer by layer. The cascade structure can gradually enhance the output sentences, achieving better visual feature guidance and language coherence. The expression for the cascaded diffusion process is:

$$F(X_t, \gamma(t'), z^I) = \prod_{n=1}^{N} f_n(X_t, \gamma(t'), z^I). \tag{13}$$

where N is the number of cascaded diffusion models, $f_i(\cdot)$ is a single diffusion model, X_t is the noise sequence used for generating text, and $\gamma(t')$ is a monotonically increasing function from 0 to 1.

Each diffusion model is divided into two parts: the forward process and the reverse process. $q(X_t|X_{t-1})$ is the conditional probability distribution of the forward diffusion process; $p(X_{t-1}|X_t, Z^I)$ is the conditional probability distribution of the reverse denoising process. Z^I is the output of the information bottleneck-driven cross-modal alignment module, used to condition the image features for the reverse process as semantic information; C is the original text sequence. The forward process is the process of gradually adding noise, through which high-frequency noise is continuously added to transform the original data distribution into a simple standard Gaussian distribution. The reverse process is the process of denoising, gradually removing high-frequency noise from the Gaussian distribution to approach the real data distribution, achieving the goal of text generation. The diffusion model based on the Transformer encoder-decoder structure includes a visual encoder and a text decoder. Each encoder contains one self-attention layer for multimodal features and one feedforward network. The first layer of the encoder receives the multimodal features Z^I aligned with the image semantics, and encodes them as conditional semantics for transmission. The decoder includes one self-attention layer for the noise sequence, one cross-attention layer, and one feedforward network. The self-attention layer of the first layer of the decoder receives the noise sequence $X_t = [x_0^t, x_1^t, \cdots, x_n^t]$ at time step t, and then passes through the cross-attention layer, combining the multimodal features after encoding to control the noise sequence, ultimately generating an image description. The calculation for each Transformer structure is as follows:

$$\mathrm{MH}(Q, K, V) = \mathrm{Concat}(h_1, \cdots, h_p)W^O, \tag{14}$$

$$h_i = \mathrm{Att}\left(QW_i^Q, KW_i^K, QW_i^V\right), \tag{15}$$

$$\mathrm{Att}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{16}$$

$$H^I = \mathrm{LN}(Z^I + \mathrm{FFN}(T(Z^I))), \tag{17}$$

$$T(X) = \mathrm{LN}(X + \mathrm{MH}(X, X, X)), \tag{18}$$

$$O = \mathrm{FFN}(\mathrm{LN}(X_t + \mathrm{MH}(T(X_t), H^I, H^I))). \tag{19}$$

where MH represents the multi-head attention layer, Att represents the attention layer, softmax is the activation function, LN represents layer normalization, FFN represents the feedforward neural network, Q, K, and V are all input feature matrices, h_i represents the output of the single-head attention, W^O is the linear projection parameter, W^Q, W^K, and W^V are all attention head parameters, d_k is the scaling factor, X is the feature used as the residual connection.

The forward process is defined as a Markov chain, gradually adding noise to the data X_0 to obtain a series of increasingly noisy X_t, where t ranges from 0 to 1.0, representing the transition from the least noisy to the most noisy. To alleviate the insufficient training caused by standard Gaussian noise, the noise scale at each step is enhanced, and the covariance matrix of the forward process $q(X_t|X_{t-1})$ is extended from $\beta_t E$ to $P_t\beta_t E$. For any $t \in [0, 1]$, the forward state transition from X_0 to X_t is calculated as:

$$q(X_t|X_{t-1}) = \mathcal{N}\left(X_t; \sqrt{1-\beta_t}X_{t-1}, P_t\beta_t E\right). \tag{20}$$

where β_t is the noise level added at time step t, P is the noise enhancement level, and E is the identity matrix.

The reverse process of the diffusion model is achieved through the reverse state transition: $X_T \to X_{T-1} \to \cdots \to X_0$. This process realizes the generation of sentences related to a given image from the diffusion model based on Transformer. The reverse process samples a series of latent states X_T, and estimates X_0 by iteratively removing noise through the denoising function f at each X_T.

$$p_\theta(X_{t-1}|X_t, Z^I) = \mathcal{N}(X_{t-1}; \mu_\theta(X_t, t), \sum_\theta(X_t, t)). \tag{21}$$

where $\mu_\theta()$ and $\sum_\theta()$ are the predicted mean and covariance of $q(X_{t-1}|X_t)$, respectively, and θ represents the model parameters.

The training of the image description model typically uses cross-entropy loss:

$$L_{CE} = \sum_{t=1}^{T} \lg(p_\theta(C_t^* | C_{1:T})). \tag{22}$$

where C_t^* is the predicted sequence value at the current position, and $C_{1:T}$ is the target true sequence value. In the diffusion model, training is conducted through L2 regression loss to reconstruct X_0, and denoising processing is performed:

$$L_{REG} = \mathbb{E}_{t\sim U(0,T),\epsilon\sim\mathcal{N}(0,1)} \left\| f\left(\sqrt{\gamma(t)}X_0 + \sqrt{1-\gamma(t)}\epsilon, t\right) - X_0 \right\|^2 \tag{23}$$

2.3. The overall loss function

We define the following loss L to train the proposed method for obtaining cartoon image captioning:

$$L = L_{IBCAM} + \lambda L_{CE} + \gamma L_{REG} \tag{24}$$

L_{IBCAM} achieves accurate semantic alignment by minimizing the redundant information between cartoon and text representations and maximizing their mutual information, laying the foundation for generating descriptions that are highly relevant to the image content. L_{CE} supervises the model's output by measuring the difference between the generated captions and the ground-truth captions, thereby improving the accuracy and relevance of the generated descriptions. L_{REG} optimizes the denoising capability in the training of the diffusion model, ensuring that the generated captions are coherent and natural. The hyperparameters λ and γ control the weights of L_{CE} and L_{REG} in the overall loss, respectively. By adjusting them, a balance can be achieved among semantic alignment, generation accuracy, and denoising capability.

3. Results and discussion

3.1. Setup

Datasets and Metrics: Following previous works [14–17], Two cartoon image captioning datasets are used as benchmark datasets to conduct performance comparisons of different image captioning models. CartoonCap-9k, the largest cartoon-focused benchmark, contains 90,00 images spanning diverse styles, each paired with 5 creative captions describing character interactions and scene narratives. It follows an 8k/0.5k/0.5k split for training, validation, and testing, and includes metadata such as art styles and scene categories to support multimodal research. ToonFables-32k, extended from the ToonTales-8k dataset, focuses on cartoon-specific narratives, comprising 32,000 images from animated films and webcomics. Its annotations emphasize unique elements like humor, magical transformations, and anthropomorphic dialogues, with 5 captions per image incorporating emotional arcs and cultural references. Public splits (29k/1k/2k) ensure fair benchmarking, while character bounding boxes mitigate recognition bias for stylized objects. Evaluation metrics use include BLEU, METEOR, ROUGE, and CIDEr to fairly assess the quality of the generated captions.

Implementation Details. The multimodal encoder is initialized with BERT pre-trained weights, with the weights of the cross-attention layer parameters randomly initialized. Other training hyperparameters follow the configuration in CLIP. During text processing, punctuation is removed, and letters are converted to lowercase. Descriptions are truncated to 20 words, and tokenization is performed using the SpaCy toolkit. In the word embedding process, words in the vocabulary are mapped to a real number vector space. All input images are resized so that the maximum dimensions of the shorter and longer sides are 384 and 640, respectively. Each diffusion model's encoder and decoder consist of 3 Transformer layers, each containing 8 attention heads and 512 hidden states. The model is trained for 60 epochs using cross-entropy loss and L2 loss, with the network learning rate fixed at 0.00001.

3.2. Comparison with baselines

Baselines. The seven baselines in the toxic text detection are used as comparison methods, including M-FFN [1], Haav [2], PromptCap [3], CLIPScore [7], textcap [8], SCD [9], Smallcap [10]and Evcap [11].

IBA-CD demonstrates significant performance improvements over the baseline methods on both the CartoonCap-9k and ToonFables-32k datasets, as shown in Tables 1 and 2. The advantages of our IBA-CD method can be attributed to the following two key aspects: (1) The Information Bottleneck-driven Cross-modal Alignment module plays a crucial role in optimizing the semantic alignment between cartoon images and their corresponding captions. By minimizing the mutual information between the original data and their representations while maximizing the mutual information between the cartoon and text representations, our method effectively eliminates redundant information and focuses on the most relevant features. This ensures that the generated captions are highly coherent with the visual content of the cartoon images, leading to improved performance in terms of BLEU, METEOR, and CIDEr scores. (2) The Cascaded Diffusion Generation Network leverages the diffusion model paradigm to generate captions in a non-autoregressive manner. By stacking multiple diffu-

Table 1. Comparison results on CartoonCap-9k dataset across four metrics.

Metric	M-FFN	Haav	PromptCap	CLIPScore	textcap	SCD	Ours
BLEU-1	65.25	65.92	71.83	75.37	74.11	76.79	**80.20**
BLEU-4	26.11	25.92	27.83	33.47	33.11	35.79	**39.20**
METEOR	15.01	21.77	23.02	23.10	24.76	25.87	**28.84**
ROUGE	39.77	43.11	47.37	50.40	53.00	51.08	**56.32**
CIDEr	105.8	120.4	119.3	108.3	111.7	121.1	**125.7**

Table 2. Comparison results on ToonFables-32k dataset across four metrics.

Metric	M-FFN	Haav	PromptCap	CLIPScore	textcap	SCD	Ours
BLEU-1	47.77	54.90	62.11	55.36	62.31	62.52	**67.74**
BLEU-4	19.85	26.05	25.82	25.74	23.77	27.63	**29.31**
METEOR	17.25	17.88	17.28	19.36	19.84	20.23	**22.73**
ROUGE	45.80	49.67	50.24	47.08	50.79	48.84	**53.52**
CIDEr	55.86	57.82	56.49	58.08	60.20	61.34	**65.40**

Table 3. Ablation study results on CartoonCap-9k dataset.

Method	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr
w/o L_{IBCAM}	76.80	37.50	27.84	54.12	120.3
w/o L_{CE}	78.20	38.10	28.24	54.80	122.8
w/o L_{REG}	79.10	38.50	28.54	55.20	123.5
Ours	**80.20**	**39.20**	**28.84**	**56.32**	**125.7**

Table 4. Ablation study results on ToonFables-32k dataset.

Method	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr
w/o L_{IBCAM}	64.52	27.63	21.23	50.84	61.34
w/o L_{CE}	65.74	28.11	21.73	51.52	62.50
w/o L_{REG}	66.24	28.53	22.23	52.12	63.80
Ours	**67.74**	**29.31**	**22.73**	**53.52**	**65.40**

sion models and using the output of the previous layer as input for the next, our method gradually refines the generated captions. This cascaded structure enhances the visual feature guidance and language coherence, resulting in more accurate and natural captions. The diffusion process, combined with the Transformer encoder-decoder structure, allows for efficient noise removal and high-quality text generation, which significantly contributes to the superior BLEU-4 and ROUGE scores compared to other baselines.

3.3. Parameter Analysis

In natural language processing tasks, the choice of hyperparameters significantly impacts model performance. To gain a deeper understanding of the effects of parameters α and β on model performance, we conducted detailed parameter analysis experiments on the *CartoonCap-9k* and *ToonFables-32k* datasets, as shown in Fig. 2. By comparing the parameter analysis results on the two datasets, we find that the optimal values of α and β differ across datasets. This suggests that in practical applications, the choice of hyperparameters should be adjusted according to the specific characteristics of the dataset to achieve the best model performance. Moreover, the parameter analysis also reveals the adaptability of the model to different datasets,

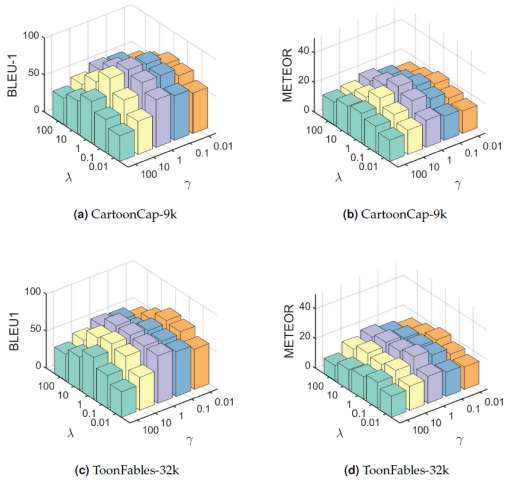

Fig. 2. Parameter analysis of α and β on the *CartoonCap-90k* and *ToonFables-32k* dataset

providing important references for subsequent model optimization and improvement.

3.4. Ablation Study

To validate the contributions of the three components of our loss function, we conducted an ablation study on both the CartoonCap-9k and ToonFables-32k datasets. The results are presented in Tables 3 and 4. From the results, we can draw the following conclusions: (1) The information bottleneck loss L_{IBCAM} is crucial for achieving accurate cross-modal alignment. When L_{IBCAM} is removed, the performance drops significantly across all metrics. This demonstrates that L_{IBCAM} effectively reduces redundant information and enhances the semantic alignment between cartoon images and text representations. (2) The cross-entropy loss L_{CE} plays a vital role in supervising the model's output to match the ground-truth captions. Without L_{CE}, the performance of the model degrades notably. (3) The L2 regression loss L_{REG} is critical for optimizing the denoising capability of the diffusion model. When L_{REG} is removed, the performance of the model also suffers. This indicates that L_{REG} significantly enhances the coherence and naturalness of the generated captions by improving the denoising process. In summary, each component of the loss function contributes uniquely and importantly to the overall performance of the IBA-CD method. The full model, which integrates all three losses, achieves the best performance across all evaluation metrics on both datasets. This ablation study confirms the effectiveness and necessity of each loss component in our proposed framework.

4. Conclusion

We propose IBA-CD, a novel framework combining information bottleneck-driven cross-modal alignment and a cascaded diffusion network for cartoon image captioning. It effectively aligns cartoon images with text descriptions, enhancing semantic correlation and generating coherent captions non-autoregressively. Our model surpasses existing methods in caption accuracy and meaningfulness, marking a significant step forward in cross-modal generation. Future work on IBA-CD will explore expanding its application to video captioning and multi-modal fusion tasks. We also plan to test it on more diverse cartoon datasets and optimize the model architecture for better performance. Enhancing caption diversity and quality, as well as investigating real-world uses in creative industries, are also on our agenda.

5. Acknowledgment

This paper was supported by the Science and technology research projects, name: Research on key technologies of image fusion based on deep learning, project number: 242102210187.

References

[1] J. Prudviraj, C. Vishnu, and C. K. Mohan, (2022) *"M-FFN: multi-scale feature fusion network for image captioning"* **Applied Intelligence *52***(13): 14711–14723. DOI: 10.1007/s10489-022-03463-x.

[2] C.-W. Kuo and Z. Kira. "Haav: Hierarchical aggregation of augmented views for image captioning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, 11039–11049.

[3] Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, and J. Luo, (2022) *"Promptcap: Prompt-guided task-aware image captioning"* **arXiv preprint arXiv:2211.09699**: DOI: 10.48550/arXiv.2211.09699.

[4] J. Gao, P. Li, A. A. Laghari, G. Srivastava, T. R. Gadekallu, S. Abbas, and J. Zhang, (2024) *"Incomplete multiview clustering via semidiscrete optimal transport for multimedia data mining in IoT"* **ACM Transactions on Multimedia Computing, Communications and Applications *20***(6): 1–20. DOI: 10.1145/3625548.

[5] J. Gao, M. Liu, P. Li, A. A. Laghari, A. R. Javed, N. Victor, and T. R. Gadekallu, (2023) *"Deep Incomplete Multiview Clustering via Information Bottleneck for Pattern Mining of Data in Extreme-Environment IoT"* **IEEE Internet of Things Journal *11***(16): 26700–26712. DOI: 10.1109/JIOT.2023.3325272.

[6] J. Gao, M. Liu, P. Li, J. Zhang, and Z. Chen, (2024) *"Deep Multiview Adaptive Clustering With Semantic Invariance"* **IEEE Transactions on Neural Networks and Learning Systems *35***(9): 12965–12978. DOI: 10.1109/TNNLS.2023.3265699.

[7] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi. "CLIPScore: A Reference-free Evaluation Metric for Image Captioning". In: *EMNLP (1)*. 2021. DOI: 10.48550/arXiv.2104.08718.

[8] D. Xu, W. Zhao, Y. Cai, and Q. Huang. "Zero-textcap: Zero-shot framework for text-based image captioning". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, 4949–4957. DOI: 10.1145/3581783.3612571.

[9] J. Luo, Y. Li, Y. Pan, T. Yao, J. Feng, H. Chao, and T. Mei. "Semantic-conditional diffusion networks for image captioning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, 23359–23368.

[10] R. Ramos, B. Martins, D. Elliott, and Y. Kementchedjhieva. "Smallcap: lightweight image captioning prompted with retrieval augmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 2840–2849.

[11] J. Li, D. M. Vo, A. Sugimoto, and H. Nakayama. "Evcap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 13733–13742.

[12] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji. "Dual-level collaborative transformer for image captioning". In: *Proceedings of the AAAI conference on artificial intelligence*. ***35***. 3. 2021, 2286–2293. DOI: 10.1609/aaai.v35i3.16328.

[13] B. Liu, D. Wang, X. Yang, Y. Zhou, R. Yao, Z. Shao, and J. Zhao. "Show, deconfound and tell: Image captioning with causal inference". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 18041–18050.

[14] Z. Zhan, X. Mao, H. Liu, and S. Yu, (2025) *"STGL: Self-Supervised Spatio-Temporal Graph Learning for Traffic Forecasting"* **Journal of Artificial Intelligence Research 2**(1): 1–8. DOI: 10.70891/JAIR.2025.040001.

[15] W. Zhang and J. Wang, (2024) *"English Text Sentiment Analysis Network based on CNN and U-Net"* **Journal of Science and Engineering *1***(1): 13–18. DOI: 10.70891/JSE.2024.100009.

[16] J. Zhang, L. Jain, Y. Guo, J. Chen, K. Zhou, S. Suresh, A. Wagenmaker, S. Sievert, T. T. Rogers, K. G. Jamieson, et al., (2024) *"Humor in ai: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning"* **Advances in Neural Information Processing Systems *37***: 125264–125286.

[17] K. Tanaka, K. Uehara, L. Gu, Y. Mukuta, and T. Harada. "Content-Specific Humorous Image Captioning Using Incongruity Resolution Chain-of-Thought". In: *Findings of the Association for Computational Linguistics: NAACL 2024*. 2024, 2348–2367. DOI: 10.18653/v1/2024.findings-naacl.152.