

## RESEARCH ARTICLE

# Diff-KT: Text-Driven Image Editing by Knowledge Enhancement and Mask Transformer

HONG ZHAO<sup>1</sup>, WENGAI LI<sup>1</sup>, ZHAOBIN CHANG<sup>2</sup>, AND CE YANG<sup>1</sup>

<sup>1</sup>Department of Computing and Communication, Lanzhou University of Technology, Lanzhou, Gansu 730050, China

<sup>2</sup>Department of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu 730000, China

Corresponding author: Wengai Li (liwengai@foxmail.com)

This work was supported in part by the National Natural Science Foundation and Development Program of China under Grant 62166025 and in part by the Science and Technology Project of Gansu Province of China under Grant 21YF5GA073.

**ABSTRACT** Recent advancements in text-to-image generation have demonstrated significant progress, especially with diffusion-based models conditioned on textual prompts, which excel in image quality and diversity. However, these methods often encounter a semantic gap between image and text modalities and suffer from imprecise localization during text-based image editing. To address these challenges, we propose the Diffusion-based Knowledge-enhanced Mask Transformer (Diff-KT) text-to-image model. Diff-KT leverages knowledge enhancement strategies to incorporate fine-grained textual and visual knowledge of key scene elements, thereby improving the fidelity and textual consistency of generated images. Furthermore, it enhances the controllability of textual influences on image generation by using masks to precisely target areas in the image for editing. To facilitate deeper fusion of visual and textual information, we introduce a multimodal pre-trained model CoCa, to extract joint representations of images and text, enhancing the detailed expression in generated images. Diff-KT improves the correlation between text and generated images and enhances image localization precision within the diffusion model, resulting in high-quality images. Experimental results validate the advantages of the Diff-KT model, demonstrating higher correlation between generated images and text prompts, as well as more accurate localization during text-guided image editing, underscoring its practical value.


**INDEX TERMS** Text-driven image editing, diffusion model, knowledge enhancement, mask transformer.

## I. INTRODUCTION

Recent advancements in the field of text-to-image generation have achieved groundbreaking progress, with applications widely adopted in education, art design, game development, and other areas. The goal of text-to-image generation is to create high-quality, realistic images that align with textual descriptions and encompass rich details. Among existing text-to-image generation methods, diffusion models [1], [2], [3] have shown excellent performance in image fidelity and have gained widespread use. However, upon further investigation of diffusion-based text-to-image models, issues have been identified, including low correlation between generated images and textual prompts, as well as inaccurate

localization of editing regions during text-based image editing.

To address the issue of low correlation between generated images and textual prompts in diffusion models, a text-conditioned diffusion model [4] was proposed. This model leverages CLIP text embeddings and the corresponding CLIP image embeddings as generative conditions within the latent space, enhancing the alignment between generated images and textual prompts. Although this method uses a contrastive learning strategy to learn joint text-image representations, it neglects the multimodal text-image representation learning, resulting in generated images that fail to match textual information in fine details. To further mitigate information bias in diffusion models, the VQ-Diffusion model [5] employs a masking mechanism to alleviate error accumulation during the inference process. However, visual scenes often comprise multiple elements of varying importance, and diffusion

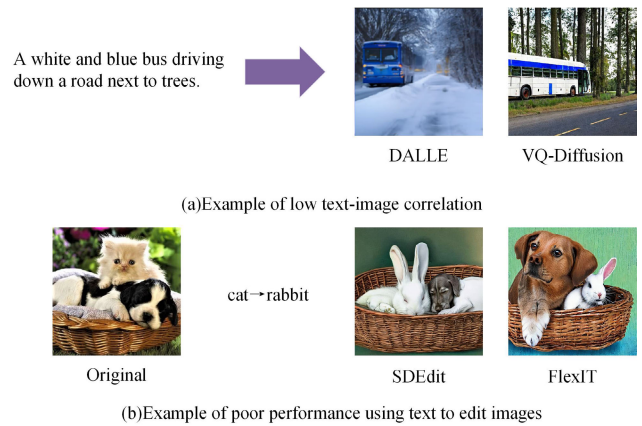
The associate editor coordinating the review of this manuscript and approving it for publication was Antonio J. R. Neves .

models do not prioritize these elements during the denoising process, leading to the omission of key elements and modality interactions. Consequently, the generated images fail to accurately represent the important semantic content of the textual prompts and cannot precisely edit specific regions or attributes of the image, limiting the flexibility and accuracy of the model's editing capabilities. This also poses a risk of text-image misalignment, such as attribute confusion, blurred positional information, and inaccurate entity counts. Moreover, diffusion models often lack attention to the critical semantic information in textual prompts, which hampers their ability to precisely guide the image generation process, thus restricting the controllability of the generated results. For example, the SDEdit method [6] struggles to maintain the color and posture of an object when editing its category attributes, ignoring detailed image features and modifying regions that do not require editing.

In recent years, several Transformer-based models have been introduced into the field of text-to-image generation, including DALL-E [7], Imagen [8], and Parti [9]. DALL-E combines GPT-3 and VQ-VAE to handle multimodal tasks involving both text and images, achieving high-quality text-to-image generation. Imagen integrates Transformers with diffusion models to generate high-fidelity images from textual descriptions, excelling in detail and semantic relevance. The Parti model utilizes the self-attention mechanism of Transformers to further enhance the quality of generated images and their alignment with textual descriptions. Through innovative architecture and algorithm design, these models effectively address the limitations of diffusion models and have made significant progress in the field of text-to-image generation.

We demonstrate the problem of low correlation between generated images and their corresponding text descriptions, as well as the inadequate representation of text-edited images. As shown in figure 1 (a), the DALL-E model [7] erroneously interprets the text prompt "white" as "snow," failing to accurately capture the intended semantics. Additionally, the VQ-Diffusion model generates a region where the "bus" is located that is inconsistent with the text description, highlighting the lack of a robust mapping relationship between text and image information. Consequently, the generated images fail to effectively match the desired text semantics. In figure 1 (b), the goal is for the model to only modify the object attributes of the "cat" according to the textual prompt while preserving the background information. However, the SDEdit model and FlexIT model [10] not only modify regions of the image that do not require editing but also alter the style of the image, indicating poor text editing capabilities.

To alleviate the problems of low correlation between generated images and textual prompts, as well as imprecise positioning of text-edited images, we propose a diffusion text-to-image generation model based on a Knowledge-Enhanced Masked Transformer. First, the model



**FIGURE 1. Problems with the diffusion model-based approach to text-driven image editing.**

employs a knowledge-enhancement strategy to incorporate additional visual scene knowledge during the denoising process of the diffusion model. By enhancing the model's attention to semantically crucial information in the textual prompts, the model selects elements of higher importance to generate the corresponding images, thereby improving the correlation between generated images and textual prompts. Secondly, the multimodal pre-training model CoCa [11] is utilized to learn joint multimodal representations between text and images. This focuses on the detailed visual representation of the generated images and uses the denoising loss function as a supervised loss for the diffusion model, assisting in generating high-quality images with fine-grained features. Finally, the introduction of a masking mechanism involves randomly masking different parts of the image and using multiple Transformer layers to extract features. The cross-entropy loss is calculated between the real labels and the masked labels to improve the model's ability to locate the text editing areas, enabling strong text editing capabilities. During the training process, the model predicts the positions of all masks and iteratively predicts masks during the inference process, resulting in improved image generation quality.

- (1) The knowledge enhancement strategy is introduced for incorporating text and visual knowledge into the diffusion model of text-generated images, which is helpful to improve fine-grained semantic control and mitigate text-image inconsistencies.
- (2) We propose the masking mechanism to edit positional regions in the original image. This can help to reduce information changes in the background region of the image while changing object properties.
- (3) The text-image encoder-decoder is combined with contrast loss and subtitle loss for pre-training, and multimodal image-text features are incorporated into the cross-attention mechanism to reduce semantic differences between text and image.

## II. RELATED WORKS

Denosing Diffusion Probabilistic Models (DDPMs), proposed in 2020, are a type of generative model that uses variational inference to train a hidden Markov chain, simulating a specific distribution from random noise. Diffusion models consist of a forward process and a reverse process. In the forward process, complex data (typically images) are gradually corrupted with noise; in the reverse process, the noise is transformed back into samples from the target distribution.

### A. TEXT-TO-IMAGE

To address the low correlation between generated images and text prompts in diffusion models, Avrahami et al. [12] proposed a text-conditioned diffusion model. This model uses CLIP text embeddings and the corresponding CLIP image embeddings as conditions to enhance the alignment between generated images and text prompts. Although this method uses a contrastive learning strategy to learn joint text-image representations, it overlooks the learning of multimodal text-image representations, leading to generated images that fail to match the text information in detail. To further reduce information bias in diffusion models, the VQ-Diffusion model [5] uses a masking mechanism to mitigate error accumulation during the inference process. However, visual scenes often contain multiple elements of varying importance, and diffusion models do not select these elements during denoising, leading to the omission of key elements and modality interactions. As a result, generated images cannot accurately express the important semantic content of the text prompts, limiting the flexibility and precision of model editing. This also results in risks of text-image misalignment, such as attribute confusion, blurred positional information, and inaccurate entity counts. In complex application scenarios, the Imagen [8] uses a text sequence encoder and a cascade of diffusion models to generate high-resolution images, conditioning on text embeddings returned by the encoder for precise text-to-image mapping. While this method improves the correlation between text prompts and generated images, it still lacks sufficient attention to key semantic information in the text prompts, limiting the controllability of the generated results. For example, the SDEdit [6] cannot ensure that the color and posture of objects remain unchanged when editing their category attributes, ignoring detailed image features and modifying regions that do not require editing. The DALL-E model [7], proposed by Song et al., uses GPT-3 to generate text descriptions and a decoder to generate corresponding images, showcasing the potential of large-scale pre-trained models in text-to-image generation tasks. However, this method still has room for improvement in fine-grained image generation. Ramesh et al. introduced DALL-E 2 [13], which improved the initial model by using a more powerful pre-trained model and an optimized generation process, enhancing the accuracy and detail of text-to-image generation.

### B. TEXT EDITING IMAGES

For the task of text-guided image editing, the precise mapping process from image to noise and back to image is crucial. This mapping process plays a central role in diffusion models, determining how the model progressively transforms the original image by adding noise and eventually generates the edited image through denoising. DDIM [3] is widely used due to its near-perfect inversion [37]. However, the local linearization assumption in DDIM can lead to image reconstruction errors and error propagation [38]. The DiffusionCLIP [14] first uses a pre-trained diffusion model to convert the input image to a latent space, and then fine-tunes the diffusion model during the reverse process. This model employs CLIP loss and consistency loss to align the target image with the text while minimizing changes to the background. However, DiffusionCLIP requires fine-tuning for each new target domain, which increases inference time. To avoid fine-tuning, the LDEdit [12] proposes using a deterministic forward process in the latent space, conditioning the reverse process on the target text. This approach performs well across a wide range of image editing tasks, providing a general framework. To address the issue of simple text modifications leading to different outputs, Hertz et al. [15] employ a cross-attention mapping strategy during the diffusion process to capture the association between each image pixel and word in the text prompt, resulting in consistent and stable image outputs. Additionally, Kwon et al. [16] propose an unsupervised image editing method that successfully disentangles style and content representations. This method can learn attributes from a source image and transfer them to a target image, achieving style transformation and content preservation without requiring additional supervision. Furthermore, DALL-E 3 [17] improves text-to-image generation quality and consistency by enhancing the model's generative capabilities and text understanding. StyleGAN3 [18] improves image fidelity and diversity through network architecture and training strategy enhancements. These methods demonstrate high practicality and flexibility in applications such as creative advertising generation, medical image analysis, and virtual reality scene construction. However, challenges remain in fine-grained image generation and precise matching of textual information. Overall, existing methods still have room for improvement in multimodal joint representation of text and images, semantic extraction of key elements, and localization accuracy.

To address the aforementioned issues, this paper introduces a knowledge-enhanced and masked Transformer model based on the diffusion framework. The model integrates textual and visual knowledge during training to improve its ability to perceive fine details and generate higher-quality images. Additionally, the multimodal pre-training model CoCa is employed to learn joint representations between text and images, enhancing the consistency of generated images with their corresponding text prompts. The model also utilizes a masking mechanism to guide the diffusion model in editing

local regions of the image. The cross-entropy loss between the real and masked labels is used as the learning objective, ensuring that the generated images avoid problems such as text-image misalignment, attribute confusion, and blurry positional information. The Diff-KT model not only enhances text-image consistency but also demonstrates precise image editing capabilities using textual prompts, exhibiting superior performance in both image generation quality and flexibility.

### III. METHODOLOGY

To alleviate the above-mentioned problems, this paper introduces the knowledge-enhanced and masked Transformer model based on the diffusion framework. The model incorporates the textual and visual knowledge during training to enhance its ability to perceive fine details and generate better-quality images. At the same time, the multimodal pre-training model CoCa is used to learn the multimodal joint representation between text and image to improve the consistency of the generated images with the text. Additionally, the model employs a masking mechanism to guide the diffusion model in editing local regions of the image. The cross-entropy loss between the real labels and the masked labels is used as the learning objective, ensuring that the generated images avoid issues such as text-image misalignment, attribute confusion, and blurry positional information. The Diff-KT model not only improves the text-image consistency, but also demonstrates precise image editing capabilities using textual prompts, showcasing superior performance in image generation quality and flexibility.

The Diff-KT model incorporates a knowledge enhancement module, a CoCa pre-trained encoder module, and a masking mechanism module into the underlying diffusion model, as shown in Fig.2. The knowledge enhancement module is responsible for integrating additional information from the visual scene, enhancing the relevance between the generated image and textual cues by selectively generating elements. The encoder module learns multimodal joint representations between text and image, ensuring that the generated images possess visual realism and fine-grained features. The masking mechanism module is employed to accurately locate the editing region in the image, thereby enhancing the model's text editing capabilities.

Firstly, textual prompts and corresponding initial images  $x_0$  are separately input into the text decoder and target detector respectively to extract the key elements of the scene, which are used as a knowledge enhancement strategy to guide the model to select different elements during the learning process to enhance the importance of the semantic representation in the scene, thus avoiding issues such as attribute confusion and text-image misalignment in the generated images. Secondly, the CoCa pre-trained model with frozen weights is used as an encoder to extract text features and image features, making all text markers interact with image regions during the learning denoising step, enhancing the cross-modal interaction of the diffusion model in cascade space. Then, a masked feature map is then obtained by randomly masking different parts of

the image. The masked feature map is fed into the underlying Transformer module together with the text features to further learn the joint multimodal representation between text and image, and the output reconstructed feature map is used to guide the diffusion denoising process. Simultaneously, the cross-entropy loss between the ground truth labels and the masked labels is used with the denoising loss function as a supervised loss for the diffusion model to assist the model in generating images with high-quality and fine-grained features. Finally, Diff-KT uses text conditioning and reconstructed feature maps to infer and localize the regions in the image that require editing during the denoising process, minimizing edits outside the regions of interest. Furthermore, the model focuses on different aspects during each stage of training. In the early stages, the input image is highly noisy, and the model needs to delineate semantic layouts and skeletons from almost pure noise. In the later stages, the model primarily focuses on denoising approximately complete images to improve details and enhance the quality of the generated images.

#### A. DIFFUSION MODEL

In recent years, DDPM have demonstrated remarkable generative performance in the field of text-to-image generation. DDPM is a class of score-based generative models that generate images by iteratively training an inverse diffusion process. DDPM consists of a forward process  $q$  and a reverse process  $p_\theta$ . The forward process  $q$  gradually adds smaller Gaussian noise to the initial data through  $T$ -step iterations until the data structure is completely broken at  $T$ -step. Conversely, the reverse process  $p_\theta$  progressively removes the added noise using a denoising function, ultimately restoring the original data, as shown in Fig.3.

During the training of the forward diffusion process, the model iteratively adds Gaussian noise to the initial image  $x_0$  and transforms the data distribution into an isotropic Gaussian distribution after  $T$  steps, the forward process is defined as:

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1 - \alpha_t}\epsilon \quad (1)$$

where  $t \in \{1, \dots, T\}$ ,  $\epsilon_t \sim \mathcal{N}(0, I)$  is the noise added at each step, and  $\{\alpha_t\}_{1 \dots T}$  is the predefined schedule. The reverse diffusion process is the denoising process, which involves sampling from Gaussian noise to compute  $x_t$ . The reverse process is defined as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon \quad (2)$$

where  $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ , the coefficient  $\alpha_t$  is defined as the noise level decreasing with time step  $t$ ,  $\alpha_t \approx 0$  is the noiseless state and  $\alpha_0 = 1$  is the almost pure noise state. During the training of the denoising process, Song et al. [3] proposed to construct an implicit model of denoising diffusion with a deterministic process through a denoising network  $\epsilon_\theta(\cdot)$ , where a given  $x_t$  is recovered by predicting the noise through the denoising network to  $x_0$ . The denoising network is defined as:

$$\mathcal{L} = \mathbb{E}_{x, \epsilon \sim \mathcal{N}(0, I), t} \left[ \|\epsilon - \epsilon_\theta(x_t, t)\|_2^2 \right] \quad (3)$$

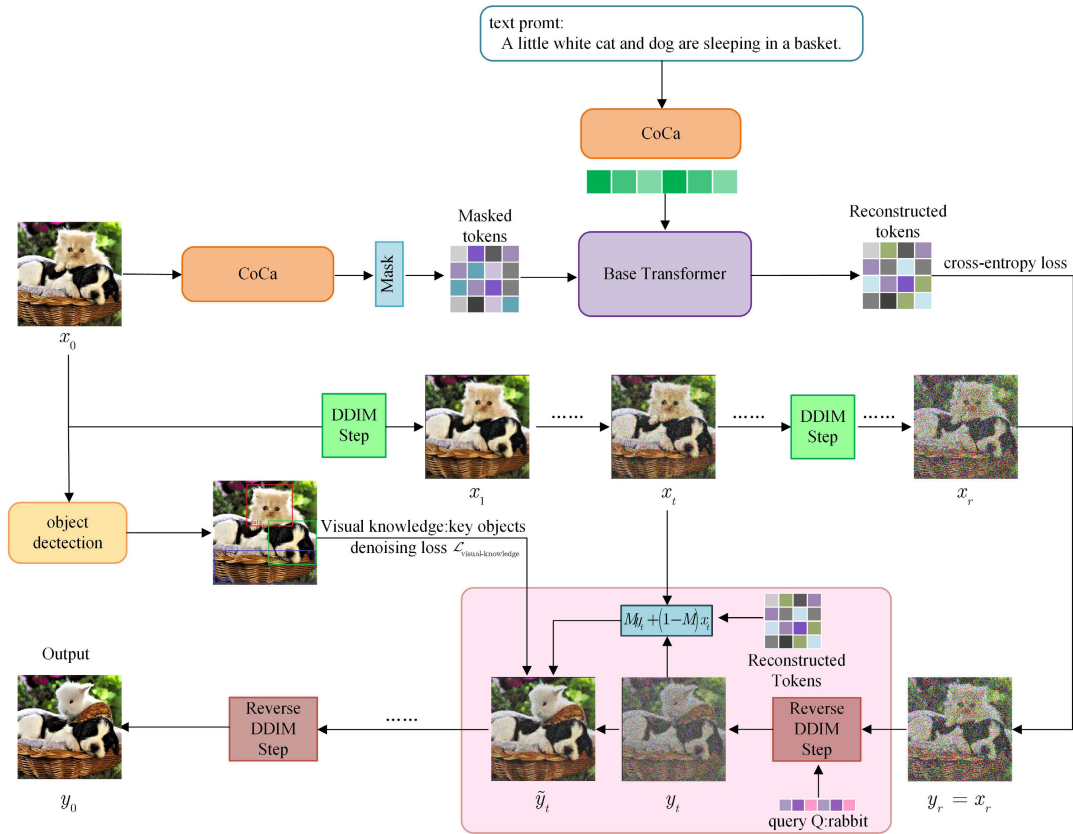


FIGURE 2. Diff-KT model overall structure.

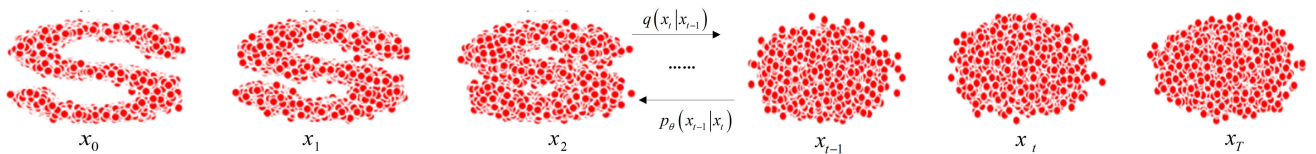


FIGURE 3. Diffusion models consist of both a forward process  $q$  and a reverse process  $p_\theta$ .

In the inference process of the DDPM, variables  $x$  are updated by small iterations in the direction of  $\epsilon_\theta$ . The update equation is defined as:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \left( \frac{x_t - \sqrt{1 - \alpha_t} \epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-1}} \epsilon_\theta(x_t, t) \quad (4)$$

In this paper, we relate the sampling process to a discretized and blurred ordinary differential equation (ODE), using a smaller step size  $T$  compared to the training process. The discretized ODE is defined as follows:

$$du = \epsilon_\theta \left( \frac{u}{\sqrt{1 + \tau^2}}, t \right) d\tau \quad (5)$$

where  $u(t = T) \sim \mathcal{N}(0, \alpha_T I)$ . In the experimental setup of this paper, the time step  $t$  is parameterized between 0 and 1, so which corresponds to the  $T$  diffusion steps in the original equation. Furthermore, during the encoding process, the  $E_r$

function is used in the encoding process to encode  $x_0$  into a latent variable  $x_r$  of time step  $r \leq 1$  using the Eulerian sampling method up to the time step  $r$ , where the variable is called the encoding rate. In the Euler sampling scheme, as long as there are sufficiently small sampling steps, the decoding  $x_r$  can approximately recover the original image  $x_0$ . In the task of text editing images, this method allows for specific region editing and modification of the image based on textual descriptions, significantly improving the controllability of text-based image editing.

### B. KNOWLEDGE ENHANCEMENT MODULE

The goal of text-to-image tasks is to generate high-quality images based on text descriptions that accurately reflect the content and detailed attributes described in the text. Key elements of a visual scene are represented differently in text and images, with keywords in the text corresponding to salient regions in the image. However, conventional

diffusion models do not prioritize the importance of these elements, generating all components indiscriminately during the iterative denoising process. To address this, we propose integrating additional textual and visual knowledge during the training phase of the diffusion model. This integration enhances the model's fine-grained semantic perception and improves the quality of the generated images.

### 1) TEXT KNOWLEDGE

To capture all the key semantic information provided by the textual prompts, this study proposes the utilization of knowledge graph embeddings to represent the relationships between entities mentioned in the textual descriptions, as shown in Fig.4. At the input layer, the triplet embeddings and text embeddings are concatenated into a joint matrix with two channels. This joint matrix is treated as a unified input that is fed into the convolutional layers and gated units. The text features, after passing through the convolutional layers, are fused with the structural features, thereby incorporating the textual information. Through multiple rounds of feature fusion, the model is able to learn the overall impact and interactions between the triplets and entity descriptions.

In the embedding layer, the structural embedding matrix, the textual embedding matrix and the joint embedding matrix are organized such that each column represents either a head entity  $h$ , the relation  $r$  or the tail entity  $t$ . During the convolutional operation, which represents the embedding dimension and  $w$  represents the number of convolutional kernels, the feature map is split into two equal parts. The symbols denote element-wise multiplication and addition, respectively. The operation  $f$  represents the Rectified Linear Unit (ReLU) activation function. After the dot product operation on the feature vectors, the final score for the triplet is obtained.

Knowledge graph embedding is a technique that addresses the knowledge graph link prediction problem by embedding relationships and entities into a low-dimensional continuous vector space. While triplets (entity/subject, relationship, pseudo-entity/object) effectively represent structured data in knowledge bases, they may not fully capture the intrinsic connections between entities and relationships, relationships and relationships, and entities and entities. Therefore, in this work, different convolutional kernels are used to generate distinct feature maps for text embedding  $M_d$ , structural embedding  $M_s$ , and joint embedding  $M_{join}$ . These feature maps are concatenated into a single vector (in shown Fig.4), which is then multiplied by a weight vector  $w$  to compute the score  $(h, r, t)$  for the triplet using a dot product calculation. The scoring function  $f(h, r, t)$  is defined as:

$$f(h, r, t) = \text{concat} \left[ \begin{array}{c} \left( \begin{array}{c} f(M_{join} * W + b) \\ \oplus f(M_{join} * V + c) \end{array} \right) \\ \otimes \left( \begin{array}{c} f(M_s * W + b) \\ \oplus f(M_s * V + c) \end{array} \right) \\ \otimes f(M_d * U + d) \end{array} \right] \cdot w \quad (6)$$

where  $W, V, U$  are different convolution kernels,  $M_s = [h_s, r, t_s]$ ,  $M_d = [h_d, r, t_d]$ ,  $M_{join} = [h, r, t]$  are 3-dimensional matrices,  $(h, r, t)$  is an embedding of a triplet,  $b$  and  $c$  are shared parameters. The loss function of the textual knowledge graph embedding is defined as:

$$\mathcal{L}_{\text{textual-knowledge}} = \sum \log \left( 1 + \exp \left( (l_{(h,r,t)} \cdot f(h, r, t)) \right) + \frac{\lambda}{2} \|w\|_2^2 \right) \quad (7)$$

$$W_a^{ij} = \begin{cases} 1 + w_a & tok_i \in \{x\}, tok_j \in \{x, y_{key}\} \\ 1 & \text{otherwise} \end{cases} \quad (8)$$

where  $W_a^{ij}$  is a scaling factor for the weight of attention between word vectors  $tok_j$ ,  $w_a$  is a hyperparameter,  $x$  is a token for all images, and  $y_{key}$  denotes the most efficient triad embedding. In this paper, the relationships of entities in the text description are reinforced by loss functions, and the text knowledge graph embeddings are fused with text features to extract more effective key semantics, thus improving the consistency of the text with the generated images.

### 2) VISUAL KNOWLEDGE

Similar to textual prompts, there are salient regions in the image that represent attributes of entities, such as people, cats, flowers, and other objects. To fully leverage the visual knowledge from the image, this paper utilizes an object detector provided by Anderson et al. [19] on 80% of the training samples and then employs a heuristic strategy to select objects with significant attributes from the detector's output. As the denoising loss function of the diffusion model operates directly in the image space, this paper modifies the loss function for the denoising process to assign higher weights to the corresponding regions, thereby promoting the model's focus on generating these objects. The loss function for the denoising process is defined as follows:

$$\mathcal{L}_{\text{visual-knowledge}} = \mathbb{E}_{z, \epsilon \sim \mathcal{N}(0, I), t} \left[ W_l \cdot \|\epsilon - \epsilon_\theta(z_t, t)\|_2^2 \right] \quad (9)$$

$$W_l^{ij} = \begin{cases} 1 + w_l & los_{ij} \in \{x_{key}\} \\ 1 & \text{otherwise} \end{cases} \quad (10)$$

where  $W_l \in \mathbb{R}^{n_h \times n_w}$  is the weight matrix,  $n_h$  and  $n_w$  are the heights and weights in image space,  $w_l$  is the hyperparameter,  $los_{ij}$  is the loss term in the  $i$ -th row and  $j$ -th column of image space, and  $x_{key}$  corresponds to the regions of the key entities.

When calculating the denoising loss function  $\mathcal{L}_{\text{visual-knowledge}}$ , the regions corresponding to the key entities are assigned higher weights. For example, as shown in Fig.2, the regions of "cat," "dog" and "basket" receive more attention. During the model training phase, the additional selection strategy is enhanced by randomly selecting a subset of samples. This strategy allows the model to perceive hints from different perspectives, thereby enabling the generation

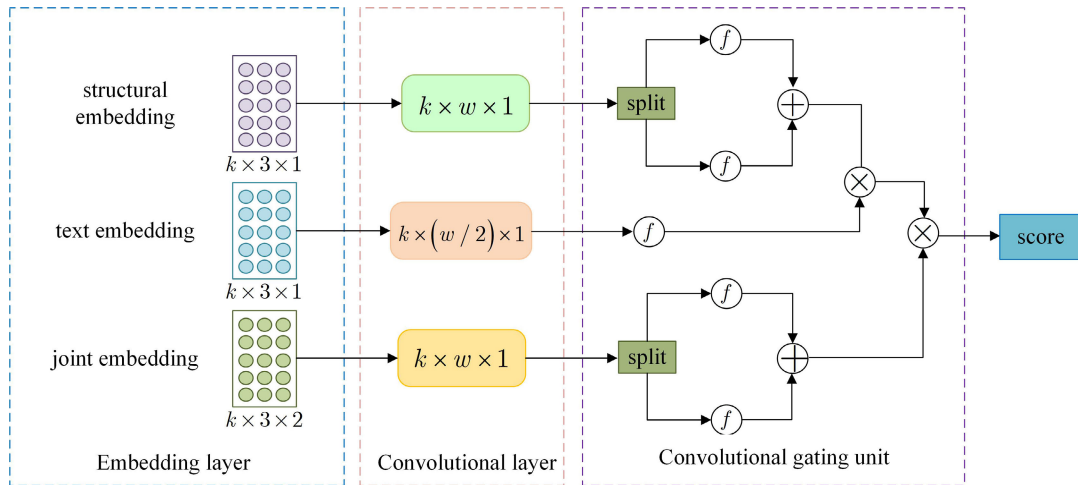


FIGURE 4. Knowledge graph embedding describes the relationships between entities in the textual prompt.

of higher-quality images during the inference stage based on the given rough textual prompts.

C. COCA PRE-TRAINED ENCODER MODULE

The task of text-to-image generation requires a powerful semantic text encoder to extract complex and constructive features from natural language text. The current state-of-the-art methods for extracting text features in this field involve training language models from scratch or using pre-trained models on image-text pairs. Inspired by Yu et al. [11], we utilize the Contrastive Captioner (CoCa) pre-trained model with frozen weights as the encoder. In Appendix A, we present pseudocode for the CoCa pre-training process. In Appendix B, we compare the effects of using CoCa and CLIP as pre-trained encoders on model performance. Experimental results indicate that CoCa emphasizes conditional attention mechanisms in image generation, while CLIP focuses on unified encoding and understanding of image and text information in cross-modal learning. Therefore, using CoCa as a pre-trained encoder has advantages in enhancing the controllability of text-guided image editing tasks. CoCa combines contrastive strategies and generative methods to improve performance, which pre-trains the basic text-image encoder-decoder model by incorporating contrastive and captioning losses. The overall structure is illustrated in Fig.5. By incorporating an approach based on object detection, the model first obtains the object and attribute categories for each region of the image. Then, the corresponding class labels are combined with the original textual prompt, enabling fine-grained descriptions and ensuring that the final input contains both coarse-grained and fine-grained information.

CoCa adopts a structure similar to the standard image-text encoder-decoder model, which encodes the image into latent representations using a Vision Transformer encoder and decodes the text using a Transformer decoder with causal masking. However, CoCa differs from the standard decoder in

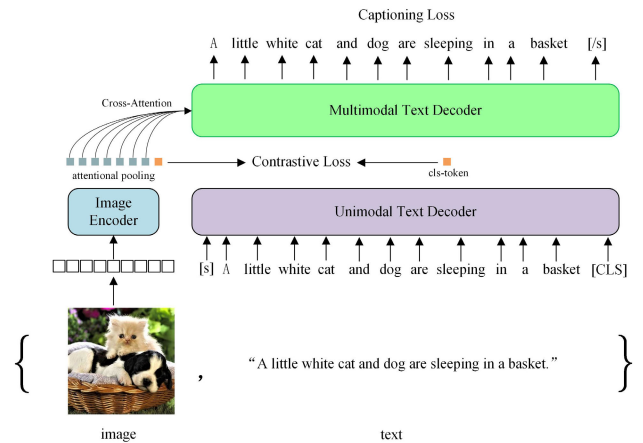


FIGURE 5. CoCa pre-trained model as an encoder with a cascaded decoder structure.

its implementation mechanism. Specifically, CoCa omits the cross-attention mechanism in the first half of the decoding layer and directly uses the unimodal text representation from the encoding layer as the input to the decoder for multimodal image-text tasks. As a result, the CoCa decoder produces both unimodal and multimodal text representations, allowing the model to apply contrastive and generative objectives as the loss functions. The loss function is defined as:

$$\mathcal{L}_{CoCa} = \lambda_{Con} \cdot \lambda_{Con} + \lambda_{Cap} \cdot \lambda_{Cap} \quad (11)$$

where  $\lambda_{Con}$  and  $\lambda_{Cap}$  are hyperparameters for the weighted losses. Contrast loss between image and text embedding is used in training, as well as caption loss, which automatically predicts text labels to mitigate the problem that selected objects may not appear in text prompts, thus promoting consistent alignment between words and objects during learning. Additionally, CoCa adopts a cascaded decoder structure, integrating the image encoder for multimodal

image-text representation into the cross-attention mechanism, enabling a seamless fusion of cross-modal features. Compared to traditional encoder structures, CoCa offers higher computational efficiency and superior performance.

#### D. MASKING MECHANISM MODULE

In the field of text-to-image generation research, the predominant focus is usually on the quality of the generated images, while the usability of text for image editing receives relatively less attention. The goal of text-based image editing is to make changes to specific parts of an image, such as attributes or entities, while keeping other parts unchanged. Previous text-to-image generation methods did not explicitly specify the regions of the image that needed to be edited, resulting in modifications to the entire image. To enhance the text editing capability of the model, this paper proposes a mask-based semantic image editing method that uses text conditioning in the diffusion model to infer masks indicating the regions that need editing. Starting from the DDIM encoding of the input image, the inferred masks guide the denoising process, minimizing edits in the unaltered regions and improving the controllability of text-based image editing.

##### 1) COMPUTING THE EDITING MASK

During the DDIM denoising process, different diffusion models conditioned on different text conditions produce different noise estimates. By using a mixture of Gaussian noise with different proportions, different regions of the image can be associated with the conditional text, while the variation in noise estimates minimally affects the background of the image. For example, in Fig.2, changing “cat” to “rabbit” while keeping the background regions such as “dog” and “basket” unchanged. Therefore, the differences in noise estimates can be used to infer masks that match the regions in the input image that need to be edited. During the experimentation process, adding too much noise to the input image makes it difficult to correctly recognize the visual elements in the input image. Hence, this paper uses Gaussian noise with an intensity of 0.5 in all experiments and removes extreme values from the noise predictions to stabilize the spatial differences in the noise. To ensure the integrity of controlling the editing regions, the masks typically extend slightly beyond the regions that need to be edited, allowing them to be smoothly integrated into the context.

##### 2) ENCODING

The Diff-KT model uses the DDIM encoder  $E_r$  to encode the input image  $x_0$  and then fuses it with the reconstructed token  $M$  that has undergone the masking mechanism to construct the edited image. Since  $x_r$  is the encoding of  $x_0$  in the hidden space at the time step  $r$ , decoding  $x_r$  using DDIM can restore the original image  $x_0$ . Assume that  $\mathcal{X} = \mathbb{R}^d$  is the space of input images and  $p_D$  represents the data distribution of the original image and the text query  $(x_0, Q)$  that edited the

image. For any  $x \in \mathcal{X}$ ,  $t \in [0, 1]$  satisfies the assumptions:

$$\|\epsilon_\theta(x_t, Q, t)\|_2 \leq C \quad (12)$$

where  $\epsilon_\theta(\cdot, \emptyset, t)$  is  $K_1$ -Lipschitz,  $K_2 = \mathbb{E}_{(x_0, Q) \in p_D} \max_{t \in [0, 1]} \|\epsilon_\theta(x, Q, t) - \epsilon_\theta(x, \emptyset, t)\|$ . We have further explanation in Appendix C. Then, for all noise coding ratios  $0 \leq r \leq 1$ , there are the following two bounds:

$$\mathbb{E}_{\substack{(x_0, Q) \sim p_D \\ \epsilon \sim \mathcal{N}(0, 1)}} \|x_0 - D_r(G_r(x_0, \epsilon), Q)\|_2 \leq (C + 1)\tau \quad (13)$$

$$\mathbb{E}_{(x_0, Q) \sim p_D} \|x_0 - D_r(E_r(x_0), Q)\|_2 \leq \frac{K_2\tau}{\sqrt{\tau^2 + 1}} \left( \tau + \sqrt{\tau^2 + 1} \right)^{K_1} \quad (14)$$

where  $\tau = \sqrt{1/\alpha_r - 1}$  increases with the coding ratio  $r$ . The encoding ratio  $r$  determines the strength of editing, and a larger value can better match the text query, giving the model stronger editing capabilities. This parameter’s impact on the model was evaluated in the ablation experiments conducted in this paper.

##### 3) DECODING WITH MASK GUIDANCE

During the image editing process, DDIM with the textual query  $Q$  as a condition is used to decode the image. In this paper, the mask  $M$  is used to guide the diffusion process. The diffusion process is guided by a mask  $M$ . The edited image should be the same as the input image, except for the regions masked by mask  $M$ . This means that while modifying the edited image in the editing region, the background region of the original image should be preserved. After inferring the latent  $x_r$  from the DDIM encoding, the pixel values outside the mask are replaced with  $x_r$ , and the image is finally mapped back to the original pixels through decoding. The mask-guided DDIM update process is defined as follows:

$$\tilde{y}_t = M y_t + (1 - M) x_t \quad (15)$$

Furthermore, both unconditional and conditional noise estimation networks yield similar estimation results. Therefore, when initialized with the same starting point  $x_r$ , the decoding behavior will also be highly similar, which helps to minimize the distance between the edited image and the input image.

## IV. EXPERIMENTS

### A. EXPERIMENTAL SETUP

#### 1) DATASETS

This paper conducts experiments on three datasets: CUB-200 [20], MS-COCO [21], and ILSVRC-2012 [22]. The specific details of these datasets are shown in Table 1.

#### 2) IMPLEMENTATION

In this paper, a basic Transformer model is trained, including self-attention blocks, cross-attention blocks and Multi-Layer Perceptron blocks, which cross-attention between text and images, as well as self-attention within image tokens. For detailed parameter designs, please refer to Table 2.



TABLE 1. Details of the datasets.

Datasets	Number of training set images	Number of test set images	Text prompts/image	Total categories
CUB-200	8855	2933	10	200
COCO	82k	40k	5	80
ILSVRC-2012	1200k	100k	-	1000

TABLE 2. Basic transformer model configuration.

Configuration	Value
Number of Transformer layers	24
Optimizers	AdaFactor
Learning rate	1e-4
Weight decay	0.045
Optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.96$
Batch size	256
Learning rate schedule	cosine decay
Training steps	15k

This experiment was conducted using NVIDIA A100 GPUs for both network training and testing. During the training process, a batch size of 256 was used, and the training was performed for 15,000 steps, which took approximately one week to complete. The Adafactor optimizer was utilized to save memory consumption, allowing the model to fit without model parallelization. Additionally, in DDIM sampling, 50 steps with a fixed schedule were used, and the encoding ratio parameter further reduced the number of updates used for editing. As a result, Diff-KT was able to complete image editing within 10 seconds on a single A100 GPU.

### 3) EVALUATION CRITERIA

#### a: FID

The Fréchet Inception Distance(FID) score is a metric used to measure the distance between the feature representations of generated and real images. A lower FID value indicates that the features of the two sets are closer, indicating that the generated images are closer to the real images and thus more visually realistic. The calculation formula for FID is as follows:

$$FID = \|\mu_r - \mu_g\| + Tr \left( \frac{\Sigma_r + \Sigma_g - 2 \left( \frac{\Sigma_r \Sigma_g}{\Sigma_r + \Sigma_g} \right)^{\frac{1}{2}}}{\Sigma_r + \Sigma_g} \right) \quad (16)$$

where  $\mu_r$ ,  $\mu_g$ ,  $\Sigma_r$ , and  $\Sigma_g$  represent the mean of real image features, the mean of generated image features, the covariance of real image features, and the covariance of generated image features, respectively.

#### b: CLIPSCORE

CLIPScore [23] is an evaluation metric that does not require inference and differs from methods that use a pre-trained Inceptionv3 network to compute similarity scores between image distributions. In contrast, this evaluation criterion is more robust. It measures the alignment between text and generated images by directly calculating the text-image similarity:

$$CLIPScore(c, v) = w * (\cos(c, v), 0) \quad (17)$$

TABLE 3. FID-30K and zero-shot FID-30K on the MS-COCO 256×256 validation set.

Approach	Model Type	FID-30K↓	Zero-shot FID-30K↓
DF-GAN [26]	GAN	21.42	-
LAFITE [27]		8.12	26.94
MSCGAN [28]		23.84	-
LDM [29]	Diffusion	17.01	12.63
GLIDE [30]		32.08	12.24
DALL-E 2 [13]		-	10.39
Stable Diffusion [2]		-	8.32
Imagen [8]		-	7.27
ERNIE-ViLG 2.0 [31]		-	6.75
RAPHAEL [32]		-	6.61
Diffusion(Ours)		7.62	6.38

where  $w$  is the scaling factor,  $c$  is the CLIP embedding of the caption, and  $v$  is the CLIP embedding of the image. CLIPScore requires no additional model inference operations, can process 4k image-text pairs in 1 minute, and has a high similarity to human evaluation results.

#### c: CSFID-LPIPS TRADE-OFF RATIO

CSFID [24] is a conditional FID metric designed to measure the consistency between image authenticity and editing prompts. LPIPS [25] is a perceptual distance metric that quantifies the distance between the original image and the generated image. When performing text-driven image editing, it is essential not only to match the textual query but also to ensure that the generated image closely resembles the input image as much as possible.

### B. QUANTITATIVE EVALUATION

In this study, 30,000 images were randomly generated on the test sets of the COCO and CUB-200 datasets. The quality and diversity of the generated samples were evaluated by calculating the FID score, while the IS score was used to measure the diversity of the generated images. These metrics were compared with those of state-of-the-art text-to-image generation models in the field. The experimental results are presented in Table 3.

In Table 3, the Diff-KT model shows a 3.48% reduction in zero-shot FID-30K scores on the MS-COCO dataset compared to the RAPHAEL [32]. This indicates that the Diff-KT model has improved the realism of generated images and outperforms well-known image generators such as Stable Diffusion, Imagen, ERNIE-ViLG 2.0 [31], and DALL-E 2 [13], achieving the best performance in text-to-image generation. The Diff-KT model's advantage lies in its combination of knowledge enhancement and the masked Transformer mechanism, which helps preserve the details of textual descriptions. This combination is likely a key factor

**TABLE 4. Quantitative evaluation of the COCO dataset.**

Model	Model Type	FID↓	CLIPScore↑
LAFITE [27]	GAN	26.94	0.28
VQGAN [33]		28.86	0.2
RQVAE [34]	Autoregression model	19.6	0.31
CogView [35]		27.1	0.33
Lformer-E [36]		24.11	0.31
VQ-Diffusion [5]	Diffusion model	13.86	0.25
RAPHAEL [32]		6.61	0.3
Diff-KT(ours)		6.38	0.34

in its superior performance, as the knowledge enhancement strategy effectively supplements critical elements in the scene, resulting in images that better align with the textual descriptions.

To further validate the performance of the model, we employed CLIPScore as an evaluation metric to measure the alignment between images and text. We compared the Diff-KT model with other advanced models in this field, and the experimental results are presented in Table 4.

In Table 4, the Diff-KT model achieves outstanding results on the MS-COCO dataset. Specifically, the Diff-KT model achieves an FID score of 17.96, which is an 8.37% improvement compared to the RQVAE [34] using an auto-encoder approach. Additionally, the CLIPScore is 0.01 higher than that of the CogView [35]. This success can be attributed to the knowledge enhancement strategy, which improves the understanding of fine-grained semantics, and the attention of the CoCa pre-trained encoder to multimodal text-image representations, resulting in a higher degree of alignment between the text and generated images and improved text-image correlation. Computational complexity results indicate that our Diff-KT model maintains low parameter count and computational cost while delivering efficient inference speed and excellent performance. These results are detailed in Appendix B.

### C. QUALITATIVE EVALUATION

To provide a visual understanding of the model's performance, this paper presents examples comparing the images generated by Diff-KT with those generated by DALLE [7] and GLIGEN [1]. The comparative results are shown in Fig. 6.

In Fig. 6, we observe that the Diff-KT model exhibits a precise understanding of the number of objects described in the text. For instance, in the first and second columns of the generated images, the Diff-KT model accurately generates the specified number of objects ("two cats" and "three giraffes"), and these objects blend naturally into the background, resulting in more realistic images compared to other methods. The images in columns 3 to 5 highlight the model's ability to understand multi-object composition and positional relationships while maintaining the spatial relationships indicated in the text (e.g., up vs. down, left vs. right, front vs. back). For example, when the text prompt is "A large present with a red ribbon to the left of a Christmas tree," the Diff-KT model successfully comprehends the

relative position of the gift and the Christmas tree, while also accurately capturing the detailed feature of the gift having a red ribbon.

In addition, the images generated by the Diff-KT model excel in representing the integrity and realism of objects compared to the other two methods. For instance, in the last column of the generated images, the Diff-KT model more comprehensively depicts the "bowl" and "fruits," whereas the other methods tend to selectively represent a specific category of objects. This paper attributes the outstanding performance of the Diff-KT model in object positioning, scene understanding, and quantity representation to the injection of knowledge. This knowledge endows the model with the ability to perceive and understand various named entities and detailed descriptions. Additionally, the use of the multimodal pre-training model, CoCa, further enhances the model's capability for semantic and visual alignment.

To further validate the semantic editing capability of the model, this paper compares the Diff-KT model with the SDEdit [6] and the FlexIT [10] using the CSFID-LPIPS trade-off ratio on the ILSVRC-2012 dataset, as shown in Fig. 7. Both SDEdit and Diff-KT are diffusion-based editing methods, while FlexIT is a mask-free editing method based on VQGAN and CLIP. Since the edited images better match the text query, the CSFID score of the edited images decreases. However, the edited images may deviate more from the input image, leading to an increase in LPIPS distance. Therefore, this paper uses the CSFID-LPIPS trade-off ratio to evaluate the editing capability of the models.

To conduct a fair comparison with two other editing methods, this paper excludes the use of image labels in the experiments and employs empty text as a reference during the inference of editing masks. A lower CSFID score indicates that the model better adheres to the textual conditions. Additionally, the desired outcome is to generate images that are closer to the original image, thereby reducing the LPIPS distance. The experimental results in Fig. 7 demonstrate that the Diff-KT model achieves the best trade-off between CSFID and LPIPS compared to the other methods for the same image editing task. Specifically, the Diff-KT model achieves the lowest CSFID value while significantly lowering the LPIPS score, indicating its strong editing capabilities. The paper attributes the superior editing performance of the Diff-KT model to the utilization of the masking mechanism, which guides the denoising process and enables precise localization of the image editing region. The effects of using the Diff-KT model for text-guided image editing are illustrated in Fig. 8.

### D. ABLATION EXPERIMENTS

To validate the performance enhancement of the model components proposed in our study for text-guided image editing tasks, we evaluated the quality of reconstructed images using various components on the CUB-200 dataset. The experimental results are presented in Table 5. Furthermore, to ensure the objectivity and accuracy of the evaluation

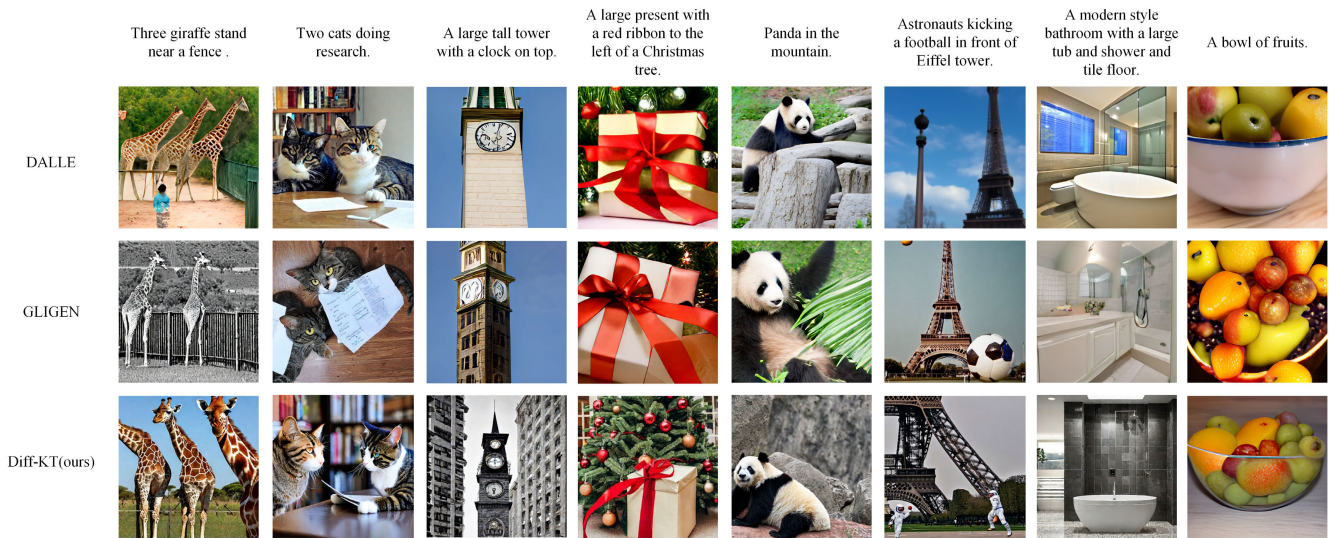


FIGURE 6. The Diff-KT and DALLE, GLIGEN generated images.

TABLE 5. The image reconstruction quality of various methods was evaluated on the CUB-200 dataset.

Base model	Method	MSE↓	LPIPS↓	SSIM↑	PSNR↑	CLIPScore↑
Diffusion	-	0.275442	0.7882	0.4213	6.8092	19.6352
	+K	0.211123	0.6526	0.6037	11.8044	21.2148
	+CoCa	0.016824	0.6413	0.5835	11.8366	21.1808
	+M	0.019924	0.1393	0.5832	11.9722	21.3655
	+K+CoCa+M(Ours)	0.009016	0.1269	0.9045	12.4993	22.4708

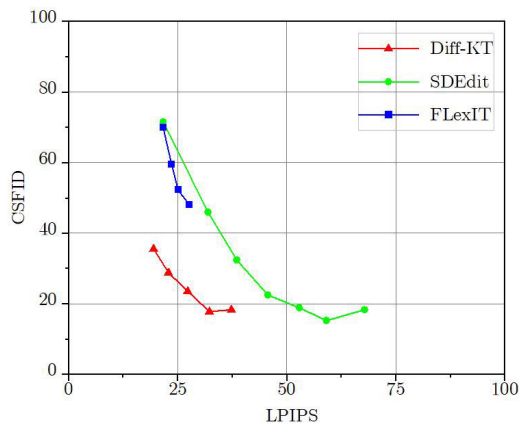


FIGURE 7. Comparison of the Diff-KT model with other image editing methods on the ILSVRC-2012 dataset.

criteria, we detailed the evaluation metrics used and their rationale in Appendix D. In Table 5, “K” denotes the knowledge enhancement module, “CoCa” represents the pre-trained encoder module, and “M” stands for the mask mechanism module. From the experimental results, it is evident that the mask mechanism significantly improves the LPIPS, PSNR, and CLIPScore metrics, leading to a notable enhancement in image quality. This improvement is attributed to the mask mechanism achieving semantic alignment between text and image features in the reconstructed image

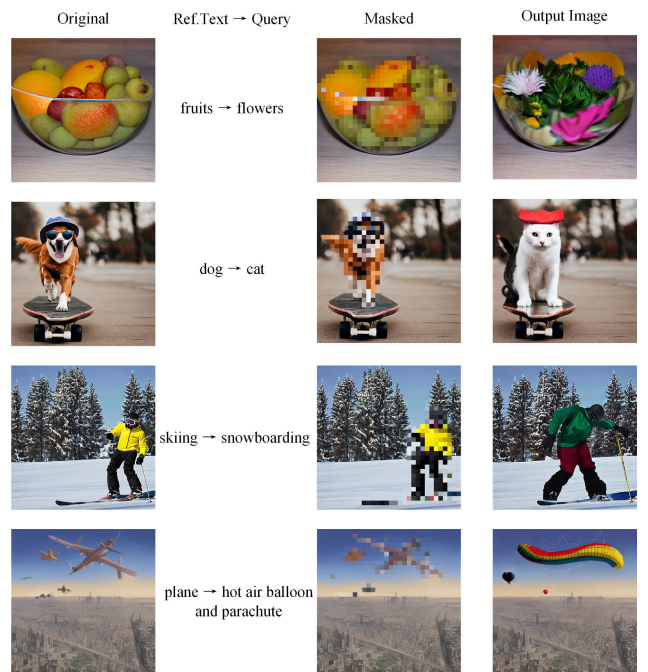
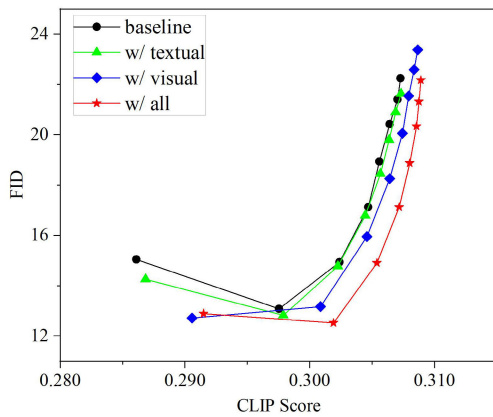
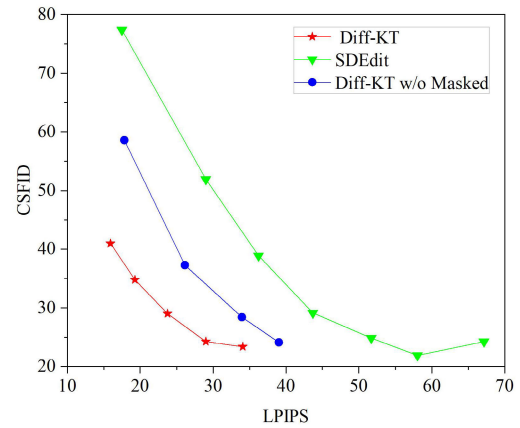


FIGURE 8. Example of image editing.

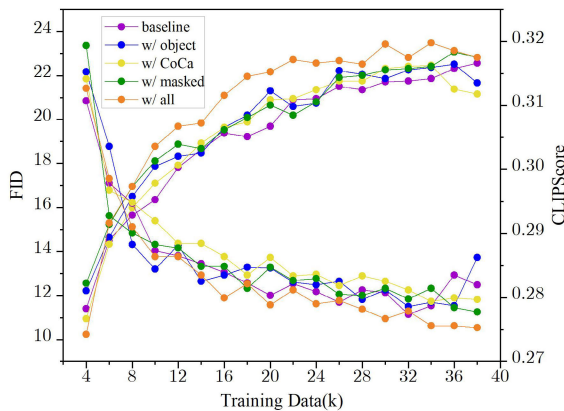
feature space, thereby further enhancing the controllability of text-guided image editing tasks. Moreover, the CoCa pre-trained encoder accurately captures subtle semantic



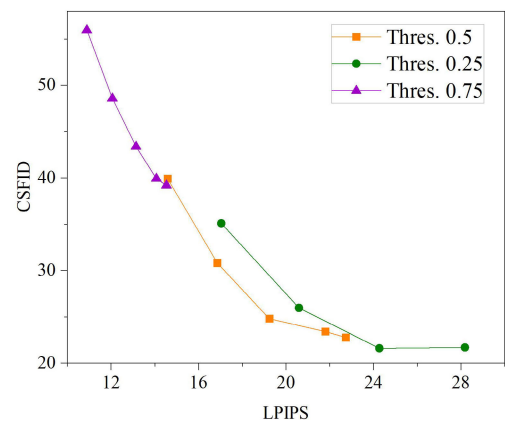
(a) Performance of knowledge enhancement strategies



(a) Effect of masking mechanism



(b) FID, CLIPScore scores for training data of different sizes



(b) Effect of different coding ratios

FIGURE 9. Impact of different components on the model.

relationships between text descriptions and images, ensuring a high semantic match between generated images and text, thereby providing users with more expected editing results. Lastly, the knowledge enhancement strategy makes generated images closer to the original images in terms of brightness, contrast, and structure, significantly increasing the realism and credibility of the images. Therefore, by combining the knowledge enhancement module, pre-trained encoder module, and mask mechanism module, our proposed Diff-KT model demonstrates outstanding performance and extensive application prospects in text-guided image editing tasks.

In this section, multiple variant models were designed to test the impact of text knowledge and visual knowledge enhancement strategies on model performance. The convergence curves in Fig.9(a) show that introducing additional knowledge into the learning process of the diffusion model leads to significant performance gains in terms of image fidelity, image-text alignment, and convergence speed. Specifically, the inclusion of text knowledge (w/ textual) enables the model to have precise fine-grained semantic control over the generated images, while the incorporation of visual knowledge (w/ visual) allows for the natural

FIGURE 10. Effect of masking mechanisms and different coding ratios.

composition of objects and scenes, thereby stabilizing and improving the model performance.

In experiments that incorporate intricate prompts, baseline models encounter significant challenges, primarily manifested as the omission of essential objects and the misattribution of attributes. To delve deeper into these issues, objects or incorrect allocation of attributes. Fig.9(b) presents a meticulous comparison of various components, explicitly showcasing the distinctive influence of each individual strategy. Notably, the harmonious integration of multiple strategies yields a remarkable improvement in model performance, not only guaranteeing a high level of fidelity in image generation but also significantly refining the correspondence between image and text in intricate, fine-grained visual landscapes.

To validate the ability of the mask mechanism and encoding ratio in controlling image editing based on text, this study employed the CSFID-LPIPS trade-off ratio to measure their relative contributions. As shown in Fig.10, compared to the SDEdit, the use of the mask mechanism better preserves the image background, resulting in a lower CSFID-LPIPS trade-off ratio and reduced average editing distance from the input image. Additionally, a lower encoding ratio of

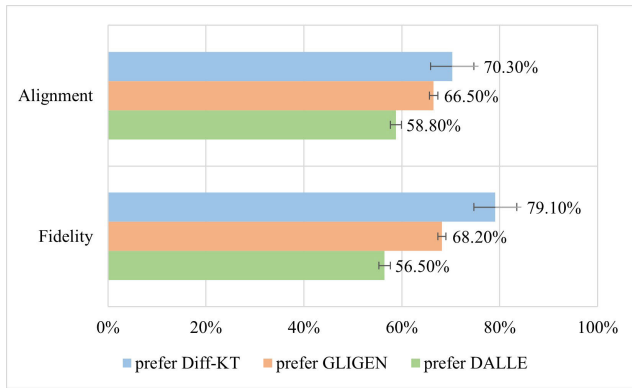


FIGURE 11. Human evaluation of Diff-KT and DALL-E, GLIGEN.

0.25 leads to more image modifications and a poorer CSFID-LPIPS trade-off, while a higher encoding ratio of 0.75 results in overly restrictive masks, causing CSFID scores to plateau around 38 even at larger encoding ratios. Therefore, the encoding ratio used in the Diff-KT model is set to 0.5.

### E. HUMAN EVALUATION

To validate the ability of the Diff-KT model to capture the cross-modal mapping between text and image and enhance the consistency between generated images and text, this paper presented three sets of images generated by the Diff-KT, DALLE, and GLIGEN models to 30 human evaluators. During the evaluation process (in appendix c), the evaluators were asked to compare each set of images based on the alignment of image-text pairs and image fidelity. The evaluators were unaware of which model generated the images. They then selected the set of images they believed performed the best or chose that there was no significant difference in performance among the three sets of images. This paper reports the evaluator preference rate with a 95% confidence interval, as shown in Fig.11.

In Fig.11, it can be observed that the human evaluators showed a preference for the generated images by the Diff-KT model in terms of image fidelity (70.3%) and consistency (79.1%), which were significantly higher than those of the DALLE and GLIGEN models. The outstanding performance of the Diff-KT model in image-text alignment and image fidelity further confirms its ability to generate high-quality images that align with the given text, aided by knowledge-enhancement strategies. Additionally, the mask mechanism and the use of a multimodal pre-training model contribute significantly to capturing the features of both text and image, enabling better alignment of semantic and visual information.

### V. CONCLUSION

To alleviate the issues of low correlation between generated images and text, as well as poor controllability of text editing in text-to-image generation methods based on diffusion models, the diffusion text-to-image generation model

based on Knowledge enhancement and masked Transformer is proposed. The model utilizes knowledge enhancement strategies to learn more important information about key elements in the external world, which is then employed as prior knowledge in the diffusion inference process to enhance the model's fine-grained language control capability. Additionally, a frozen-weight multimodal pre-training model called CoCa is employed to extract multimodal joint representations of complex and constructive language and visual information, mitigating the semantic gap between text and image. It is worth noting that the masking mechanism plays a crucial role in accurately locating the image editing regions, enabling the model to minimize edits in regions of lesser interest. The experimental results in this paper demonstrate that the Diff-KT model not only improves the correlation between generated images and text but also enables more precise localization of image editing regions using textual prompts.

Although the text-to-image generation model proposed in this paper demonstrates superior performance, it still faces issues of insufficient details and semantic inconsistencies in complex scenarios. These issues may be related to the limitations of the training data and the model's ability to process complex textual descriptions. Future research should incorporate more contextual information and multi-modal data, as well as expand the diversity and scale of the dataset, to enhance the model's generative capabilities and generalization performance. Additionally, further testing and optimization in practical applications are necessary to ensure the model's robustness and reliability.

### APPENDIX A COCA PSEUDO-CODE FOR THE PRE-TRAINING PROCEDURE

CoCa is built upon an encoder-decoder architecture, where the text decoder is divided into two parts: the unimodal text decoder and the multimodal text decoder. A cls token is added at the end of the text, and the unimodal text decoder does not engage in cross-attention with image features. This configuration allows the cls token to capture the global features of the entire sentence after passing through the unimodal text decoder. Simultaneously, attention pooling is used to extract global features from the image encoder for image feature extraction. The two global features enable contrastive learning between images and text, with attention pooling functioning as a multi-head attention mechanism where the key and value are features obtained from the image encoder, and the query is a pre-defined trainable embedding. Since only one global feature is needed, only one query is defined.

The multimodal text decoder is used for generation tasks and also employs attention pooling to extract features from the image encoder. However, in this case, the number of queries is set to 256, allowing attention pooling to obtain 256 features to serve as inputs for the cross-attention in

the multimodal text decoder. The pseudocode for CoCa is outlined below:

## APPENDIX B EVALUATING THE IMPACT OF USING COCA AND CLIP AS PRE-TRAINED ENCODERS ON MODEL PERFORMANCE

**CoCa (Conditional Combinatorial Attention):** CoCa is an image generation model designed to achieve more accurate and diverse image generation. It employs a conditional combinatorial attention mechanism by dynamically focusing on local regions during image generation to better capture relationships and contextual information between different objects. This approach enhances the quality and diversity of generated images. The key feature of CoCa is its use of a conditional combinatorial attention mechanism, resulting in more coherent and realistic generated images. It performs well in image generation tasks, particularly in scenarios that require consideration of relationships between objects.

**CLIP (Contrastive Language-Image Pre-training):** CLIP is a cross-modal learning method aimed at unifying image understanding and semantic understanding through contrastive learning of images and text. It leverages large-scale text and image data to train a universal visual-semantic encoder through contrastive learning. The main feature of CLIP is its ability to effectively integrate image and text information for cross-modal semantic understanding. It can be used for tasks such as image classification and text description generation, achieving excellent performance on some benchmark datasets.

**Algorithm:** Pseudocode of Contrastive Captioners Architecture

```
# image, text.ids, text.labels, text.mask: paired image, text data
# con_query: 1 query token for contrastive embedding
# cap_query: N query tokens for captioning embedding
# cls_token_id: a special cls_token_id in vocabulary
# vit_encoder: vision transformer based encoder
# lm_transformer: language-model transformers

def attentional_pooling(features, query):
    out = multihead_attention(features, query)
    return layer_norm(out)

img_feature = vit_encoder(image) # [batch, seq_len, dim]
con_feature = attentional_pooling(img_feature, con_query) # [batch, 1, dim]
cap_feature = attentional_pooling(img_feature, cap_query) # [batch, N, dim]

ids = concat(text.ids, cls_token_id)
mask = concat(text.mask, zeros_like(cls_token_id)) # unpad cls_token_id
txt_embs = embedding_lookup(ids)
unimodal_out = lm_transformers(txt_embs, mask, cross_attn=None)
multimodal_out = lm_transformers(
    unimodal_out[:, :-1, :], mask, cross_attn=cap_feature)
cls_token_feature = layer_norm(unimodal_out)[:, -1:, :][batch, 1, dim]

con_loss = contrastive_loss(con_feature, cls_token_feature)
cap_loss = softmax_cross_entropy_loss(
    multimodal_out, labels=text.labels, mask=text.mask)
```

Method	CLIPScore
With CLIP	0.27
With CoCa	0.34

In summary, CoCa focuses on the conditional attention mechanism in image generation, while CLIP emphasizes unified encoding and understanding of image and text information in cross-modal learning. They can be applied in different scenarios, each with its own advantages and characteristics.

## APPENDIX C SUPPLEMENTARY INFORMATION ON THE ENCODING PROCESS

In Diff-KT, we employ DDIM to encode images before the actual editing steps. In this section, we provide theoretical insights into why this component produces better editing results than adding random noise as in SDEdit [6]. Since  $x_r$  is the encoded version of  $x_0$ , decoding  $x_r$  unconditionally with DDIM will yield the original image  $x_0$ . In Diff-KT, we use DDIM decoding conditioned on a text query  $Q$ , but there is still a strong bias to maintain proximity to the original image. This is because both the unconditional noise estimation network  $\epsilon_\theta$  and the conditional noise estimation network  $\epsilon_\theta(\cdot, Q)$  often produce similar estimations, leading to similar decoding behavior when initialized from the same starting point  $x_r$ . This means that the edited image will have a smaller distance from the input image in terms of key attributes in the context of image editing. We capture this phenomenon with the following proposition, where we compare DDIM encoder  $E_r(x_0)$  with the SDEdit encoder  $G_r(x_0, \epsilon) := \sqrt{\alpha_r}x_0 + \sqrt{1-\alpha_r}\epsilon$ , which simply adds noise to the image  $x_0$ .

## APPENDIX D EVALUATION METRICS AND RATIONALE FOR ABLATION EXPERIMENTS

### A. MSE

Mean Squared Error, commonly used in statistics and machine learning, is a metric that measures the degree of difference between predicted values and actual values. It is computed by averaging the sum of squares of differences between predicted and actual values. The specific formula for calculating MSE is:

$$\text{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - K(i, j)]^2 \quad (18)$$

where  $\Sigma$  represents summation over all samples,  $I(i, j)$  and  $K(i, j)$  are the pixel values at point  $(i, j)$  in the original image and the processed image, respectively,  $m$  and  $n$  are the height and width of the image. A smaller value of MSE indicates a more accurate predictive capability of the model.

### B. LPIPS

Learned Perceptual Image Patch Similarity is a deep learning-based image quality assessment metric designed to mimic human visual perception in measuring similarity

between images. Compared to traditional image quality assessment metrics such as PSNR and SSIM, LPIPS is closer to human perception and can more accurately reflect human evaluation of image quality. When scoring using LPIPS, a pair of input images is first prepared, such as the original image and the processed image. Then, a pre-trained CNN is used to extract features from the input images, and the distance between these feature vectors is calculated using cosine distance to measure the similarity between images. The feature distance calculation method used in this paper employs cosine distance. A lower LPIPS score indicates a greater visual perceptual difference between the processed image and the original image; conversely, a higher LPIPS score indicates a closer visual perceptual similarity between the processed image and the original image.

### C. SSIM

Structural Similarity Index is a metric used to assess image quality by comparing the brightness, contrast, and structural similarity of two images to evaluate their degree of similarity. Specifically, SSIM calculates brightness estimation using mean values, contrast estimation using variance, and structural similarity estimation using covariance. SSIM scores have different implications across different ranges:

- 1) When SSIM is greater than 0.9, it indicates that two images are very similar, with differences barely perceptible, making it suitable for applications with extremely high demands on image quality.
- 2) When SSIM is between 0.7 and 0.9, it indicates a high degree of similarity between two images, where differences are difficult for the human eye to detect.
- 3) When SSIM is between 0.5 and 0.7, it indicates a moderate degree of similarity between two images, where some subtle differences may be present.

### D. PSNR

Peak Signal-to-Noise Ratio is a widely used metric for evaluating image and video quality. It measures the difference between the original image (or video frame) and the processed image (or video frame), specifically in terms of noise intensity. A higher PSNR value indicates better image quality and less distortion. PSNR is calculated primarily based on Mean Squared Error (MSE). First, the Mean Squared Error (MSE) between the original image and the processed image is computed, and then it is converted into decibels (dB) to obtain the PSNR. The specific formula for calculating PSNR is as follows:

$$\text{PSNR} = 20 \cdot \lg(\text{MAX}_i) - 10 \cdot \lg(\text{MSE}) \quad (19)$$

where  $\text{MAX}_i$  represents the possible maximum pixel value in the image. It is important to note that while PSNR is a commonly used evaluation metric, it does not always align perfectly with human visual perception. Sometimes, images with higher PSNR values may not be visually more appealing than images with slightly lower PSNR values. Therefore, in practical applications, it may be necessary to combine

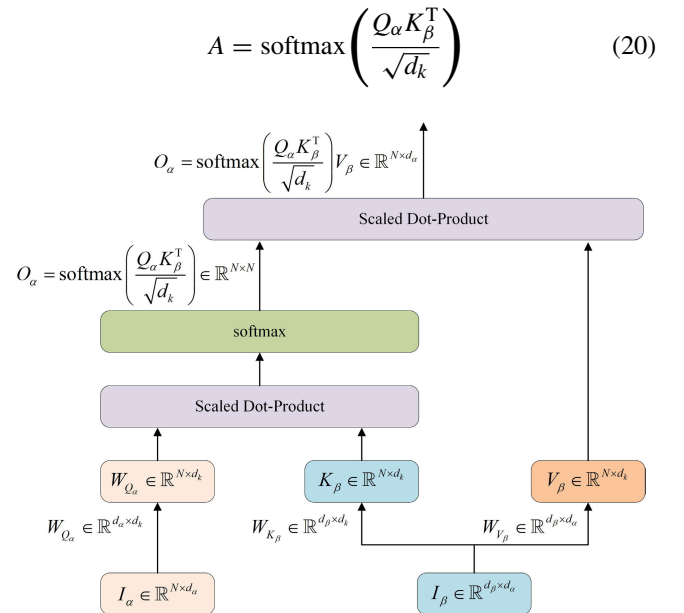
other evaluation metrics (such as SSIM, LPIPS, etc.) for a comprehensive assessment.

## APPENDIX E

### DESIGN OF CROSS-MODAL TRANSFORMER MODULE

Since the number of text tokens  $\alpha$  and image masked tokens  $\beta$  may not be consistent at the same time, the two are filled with 0 vector to the same sequence length  $N$  as the two inputs  $I_\alpha \in \mathbb{R}^{N \times d_\alpha}$  and  $I_\beta \in \mathbb{R}^{d_\beta \times d_\alpha}$  of the basic Transformer module. There are 24 Transformer layers within the base Transformer module. A Transformer layer includes: multi-head attention module, residual and normalization layer, feedforward layer. A multi-head attention module is composed of  $h$  Transformer cross-attention modules to obtain the degree of matching and correlation of multi-mode information under different mappings. The individual cross-attention modules are shown below.

For the two input modes  $I_\alpha \in \mathbb{R}^{N \times d_\alpha}$  and  $I_\beta \in \mathbb{R}^{d_\beta \times d_\alpha}$ , dot multiply with the weight  $W_{Q_\alpha} \in \mathbb{R}^{d_\alpha \times d_k}$  and  $W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_k}$  matrix respectively to generate Query matrix  $Q_\alpha$  and Key matrix  $K_\beta$ ; meanwhile, dot multiply  $I_\beta$  with the weight  $W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v}$  matrix to generate Value matrix  $V_\beta$ . Formula (20) is used to calculate the attention score matrix between the two modes, which represents the correlation degree of the information sequence of the two modes under a certain mapping. The dot product of attention score matrix  $A$  and weight matrix  $V_\beta$  reflects the potential adaptation of mode  $I_\beta$  to mode  $I_\alpha$ .



The  $i$ -th cross-attention module can be expressed as formula (21), where each attention module has different weights  $W_{Q_\alpha}^i$ ,  $W_{K_\beta}^i$ , and  $W_{V_\beta}^i$  to obtain different  $Q_{\alpha i}$ ,  $K_{\beta i}$ , and  $V_{\beta i}$  matrices. These matrices are used to obtain the output  $O_{\alpha}^i \in \mathbb{R}^{N \times d_\alpha}$  of the fusion of the two modes. The outputs of  $h$  single-cross attention modules are concatenated to obtain the output  $O_\alpha \in \mathbb{R}^{hN \times d_\alpha}$  of the multi-head attention module,

where  $h$  is the number of heads.

$$\begin{aligned}
 O_{\alpha}^i &= AV_{\beta i} \\
 &= \text{softmax} \left( \frac{Q_{\alpha} K_{\beta}^T}{\sqrt{d_k}} \right) V_{\beta i} \\
 &= \text{softmax} \left( \frac{I_{\alpha} W_{Q_{\alpha}}^i \left( W_{K_{\beta}}^i \right)^T I_{\beta}^T}{\sqrt{d_k}} \right) I_{\beta} W_{V_{\beta}}^i, \\
 & \quad i \in \{1, 2, 3, \dots, h\} \quad (21)
 \end{aligned}$$

The input mode  $I_{\alpha}$  of the  $n$ -th Transformer layer is the output  $I_{\alpha}^{[n-1]}$  of the  $n-1$  Transformer layer, and mode  $I_{\beta}$  is always the mode  $I_{\beta}^{[0]}$  of the initial layer. The multi-head attention module in the  $n$ -th Transformer layer outputs  $O_{\alpha}^{[n-1]}$  and input mode  $I_{\alpha}^{[n-1]}$  for residuals and normalization operations. As shown in formula (22), the residuals and normalization layers include residuals joining and layer normalization operations.

$$\hat{I}_{\alpha}^{[n]} = \text{LayerNorm} \left( O_{\alpha}^{[n-1]} + I_{\alpha}^{[n-1]} \right) \quad (22)$$

$\hat{I}_{\alpha}^{[n]}$  through the Feed Forward layer and the Add & Norm layer to add nonlinear changes, improve the learning ability of the network. As shown in formula (23), the Feed Forward layer consists of two fully connected layers and ReLU functions.

$$\begin{aligned}
 I_{\alpha}^{[n]} &= \text{LayerNorm} \left( \text{FeedForward} \left( \hat{I}_{\alpha}^{[n]} \right) + \hat{I}_{\alpha}^{[n]} \right) \\
 &= \text{LayerNorm} \left( \max \left( 0, \hat{I}_{\alpha}^{[n]} W_1^{[n]} + b_1^{[n]} \right) W_2^{[n]} + b_2^{[n]} + \hat{I}_{\alpha}^{[n]} \right) \\
 & \quad n \in \{1, 2, 3, \dots, D\} \quad (23)
 \end{aligned}$$

where,  $W_1^{[n]}$ ,  $b_1^{[n]}$  and  $W_2^{[n]}$  and  $b_2^{[n]}$  represent the weight  $W$  and offset  $b$  of the two fully connected layers respectively, and then output the output of the  $n$ -th Transformer layer.

In this paper, the cross-attention mechanism in the Transformer can calculate the similarity between different modal input information, thus achieving adaptive association and fusion of image and text information based on the similarity score. However, the cross-attention mechanism also has some limitations. The calculation of cross-attention is very complex, especially when processing high-resolution images or long sequence of text, which may lead to a large consumption of computing resources. The parameter optimization of cross-attention model may be difficult, and the weight of interaction between different modalities needs to be carefully balanced.

To overcome these limitations, in the subsequent research process, we will consider using more efficient attention mechanisms, designing more robust loss functions, utilizing unsupervised or semi-supervised learning methods to reduce dependence on labeled data, or developing specialized hardware to accelerate the calculation of cross-attention.

**TABLE 6. The model size and inference speed are compared with recent text-generated image methods.**

Model	Type	#Param.	Inf. Time
Parti-3B	AR	3.0B	6.4s
Muse-3B		3.0B	1.3s
Cogview2		6.0B	45.43s
LAFITE	GAN	75.0M	0.02s
GigaGAN		1.0B	0.13s
StyleGAN-T		1.0B	9.62s
GLIDE	Diff	5.0B	15.0s
LDM		1.5B	9.4s
Imagen		3.0B	9.1s
eDiff-I		9.1B	32.0s
LDM		1.5B	4.83s
<b>Diff-KT</b>	<b>Diff</b>	<b>0.9B</b>	<b>4.62s</b>

## APPENDIX F COMPARISON OF MODEL TIME CONSUMPTION

To comprehensively compare the latest text-to-image generation methods, we analyzed them from three perspectives: model type, parameter size, and inference speed. Although there are huge differences in model structure, parameter numbers, and training data, the results in Table 6 cannot be directly compared. However, it is worth noting that GAN-based methods are the most competitive in terms of inference speed in current text-to-image models. Due to the inherent advantage of generating images in one step, GAN methods are faster than autoregressive or diffusion models. Among diffusion-based text-to-image generation methods, our model not only has the smallest number of parameters but also has shorter inference time. Compared with the latest LDM model, we used fewer total images during training.

The results show that compared to LDM, one of the most important open source large-scale pre-trained models, our Diff-KT achieves better performance even with smaller model parameters and data volumes. In addition, our Diff-KT only takes 4.63s to generate an image, which is 0.21s faster than LDM. What's more, the Diff-KT is able to perform fast reasoning on the CPU without other acceleration Settings, greatly reducing the user's hardware requirements. At the same time, the computational cost required to pre-train our Diff-KT is also significantly lower than these large pre-trained autoregressive and diffusion models.

## APPENDIX G HUMAN EVALUATION DESIGN

We designed two main evaluation criteria: textual consistency (how the image content relates to the input text) and visual authenticity (whether the image looks real and natural).

- 1) Reviewer background: We invited 30 reviewers, including experts in computer vision and image processing, as well as graduate students with backgrounds in art and design, to participate in the assessment to ensure diversity and comprehensiveness.
- 2) Evaluation task: Reviewers were asked to vote on each of the three sets of images generated by the Diff-KT, DALLE, and GLIGEN models according to the above criteria.



## 3) Evaluation process:

- a) 100 groups of images were randomly selected from the image set we generated for evaluation to ensure the representativeness of the evaluation results.
- b) Each reviewer independently votes for each image, and the voting criteria are clear and consistent.
- c) Collect the number of votes of all reviewers for the image generated by the Diff-KT, DALLE and GLIGEN models.

## 4) Presentation and interpretation of evaluation results in shown Fig. 11.

It can be observed that the preference degree of human evaluators for the authenticity and consistency of the generated images of the Diff-KT model is 70.3% and 79.1%, respectively, which is significantly higher than that of DALLE and GLIGEN models. The excellent performance of the Diff-KT model in image-text alignment and image fidelity once again proves that the model can generate high-quality images that conform to text with the help of knowledge enhancement strategies. At the same time, mask mechanism and multimodal pre-training model can better capture the features of text and image, and make great contributions to the alignment of semantic information and visual information.

## REFERENCES

- [1] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," 2021, *arXiv:2112.10741*.
- [2] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. G. Lopes, B. K. Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 36479–36494.
- [3] J. Song, C. Meng, and S. Ermon, "Denosing diffusion implicit models," 2020, *arXiv:2010.02502*.
- [4] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18187–18197.
- [5] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10686–10696.
- [6] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "SDEdit: Guided image synthesis and editing with stochastic differential equations," 2021, *arXiv:2108.01073*.
- [7] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 8821–8831.
- [8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.
- [9] J. Yu, Y. Xu, J. Yu Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, B. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge, and Y. Wu, "Scaling autoregressive models for content-rich text-to-image generation," 2022, *arXiv:2206.10789*.
- [10] G. Couairon, A. Grechka, J. Verbeek, H. Schwenk, and M. Cord, "FlexIT: Towards flexible semantic image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18249–18258.
- [11] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "CoCa: Contrastive captioners are image-text foundation models," 2022, *arXiv:2205.01917*.
- [12] B. Li, X. Lin, B. Liu, Z.-F. He, and Y.-K. Lai, "Lightweight text-driven image editing with disentangled content and attributes," *IEEE Trans. Multimedia*, vol. 26, pp. 1829–1841, 2023.
- [13] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.
- [14] G. Kim, T. Kwon, and J. C. Ye, "DiffusionCLIP: Text-guided diffusion models for robust image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2416–2425.
- [15] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," 2022, *arXiv:2208.01626*.
- [16] M. Kwon, J. Jeong, and Y. Uh, "Diffusion models already have a semantic latent space," 2022, *arXiv:2210.10960*.
- [17] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, W. Manassra, P. Dhariwal, C. Chu, and Y. Jiao, "Improving image generation with better captions," *Comput. Sci.*, vol. 2, no. 3, pp. 1–19, 2023. [Online]. Available: <https://cdn.openai.com/papers/dall-e-3.pdf>
- [18] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 852–863.
- [19] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," 2017, *arXiv:1707.07998*.
- [20] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie. (2011). *The Caltech-UCSD Birds-200-2011 Dataset*. [Online]. Available: <http://www.vision.caltech.edu/visipedia/CUB-200.html>
- [21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Cham, Switzerland: Springer, Sep. 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1\_48.
- [22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/S11263-015-0816-Y.
- [23] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIP-Score: A reference-free evaluation metric for image captioning," 2021, *arXiv:2104.08718*.
- [24] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 25–34. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf)
- [25] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," 2018, *arXiv:1801.03924*.
- [26] M. Tao, H. Tang, F. Wu, X. Jing, B.-K. Bao, and C. Xu, "DF-GAN: A simple and effective baseline for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16494–16504.
- [27] Y. Zhou, R. Zhang, C. Chen, C. Li, C. Tensmeyer, T. Yu, J. Gu, J. Xu, and T. Sun, "LAFITE: Towards language-free training for text-to-image generation," 2021, *arXiv:2111.13792*.
- [28] W. Tang, G. Li, X. Bao, F. Nian, and T. Li, "MsCGAN: Multi-scale conditional generative adversarial networks for person image generation," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Aug. 2020, pp. 1440–1445.
- [29] F. Bao, S. Nie, K. Xue, C. Li, S. Pu, Y. Wang, G. Yue, Y. Cao, H. Su, and J. Zhu, "One transformer fits all distributions in multi-modal diffusion at scale," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 1692–1717.
- [30] T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard, and J. L. Banks, "Glade: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening," *J. Medicinal Chem.*, vol. 47, no. 7, pp. 1750–1759, Mar. 2004.
- [31] Z. Feng, Z. Zhang, X. Yu, Y. Fang, L. Li, X. Chen, Y. Lu, J. Liu, W. Yin, S. Feng, Y. Sun, L. Chen, H. Tian, H. Wu, and H. Wang, "ERNIE-ViLg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10135–10145.
- [32] Z. Xue, G. Song, Q. Guo, B. Liu, Z. Zong, Y. Liu, and P. Luo, "RAPHAEL: Text-to-image generation via large mixture of diffusion paths," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 41693–41706.

- [33] J. Yu, X. Li, J. Yu Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldrige, and Y. Wu, "Vector-quantized image modeling with improved VQGAN," 2021, *arXiv:2110.04627*.
- [34] D. Lee, C. Kim, S. Kim, M. Cho, and W.-S. Han, "Autoregressive image generation using residual quantization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11513–11522.
- [35] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, and J. Tang, "CogView: Mastering text-to-image generation via transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 19822–19835.
- [36] J. Li, L. Wei, Z. Zhan, X. He, S. Tang, Q. Tian, and Y. Zhuang, "Lformer: Text-to-image generation with L-shape block parallel decoding," 2023, *arXiv:2303.03800*.



**HONG ZHAO** received the bachelor's degree in computer science and technology from Northwest Normal University, in June 1993, and the Ph.D. degree in computer application technology from Xinjiang University, in June 2010. He is currently a Professor, a Doctor of Engineering, and a Ph.D. Supervisor. He is also a leading talent in Gansu Province. His research interests include parallel and distributed processing, embedded systems, system modeling and simulation, deep learning, natural language processing, and computer vision. He is a member of the System Simulation Professional Committee of the Chinese Society of Automation and the Chinese Computer Society and a China Postdoctoral Science Fund Review Expert.



**WENGAI LI** is currently pursuing the master's degree with the School of Computer and Communication, Lanzhou University of Technology, China. Her current research interests include computer vision, image processing, machine learning, and deep learning.

**ZHAOBIN CHANG** received the master's degree from Lanzhou University of Technology, Lanzhou, China, in 2020. He is currently pursuing the Ph.D. degree with the School of Information Science and Engineering, Lanzhou University, Lanzhou. His research interests include image semantic segmentation and few-shot learning.

**CE YANG** was born in Xianyang, Shaanxi, China, in 1997. He received the master's degree in engineering from Lanzhou University of Technology. His main research interest includes natural language processing.

...