RESEARCH-ARTICLE

HeadRouter: A Training-free Image Editing Framework for MM-DiTs by Adaptively Routing Attention Heads

YU XU

FAN TANG

JUAN CAO

XIAOYU KONG

YUXIN ZHANG

JINTAO LI

View all

Published: 02 March 2026
Accepted: 26 January 2026
Revised: 08 December 2025
Received: 13 May 2025

HeadRouter: A Training-free Image Editing Framework for MM-DiTs by Adaptively Routing Attention Heads

YU XU, Institute of Computing Technology Chinese Academy of Sciences, Beijing, China
FAN TANG*, Institute of Computing Technology Chinese Academy of Sciences, Beijing, China
JUAN CAO, Institute of Computing Technology Chinese Academy of Sciences, Beijing, China
XIAOYU KONG, Beihang University, Beijing, China
YUXIN ZHANG, NLPR, Chinese Academy of Sciences Institute of Automation, Beijing, China and School of Artificial Intelligence, University of the Chinese Academy of Sciences, Beijing, China
JINTAO LI, Institute of Computing Technology Chinese Academy of Sciences, Beijing, China
OLIVER DEUSSEN, Universität Konstanz, Konstanz, Germany
TONG-YEE LEE, National Cheng Kung University, Tainan City, Taiwan

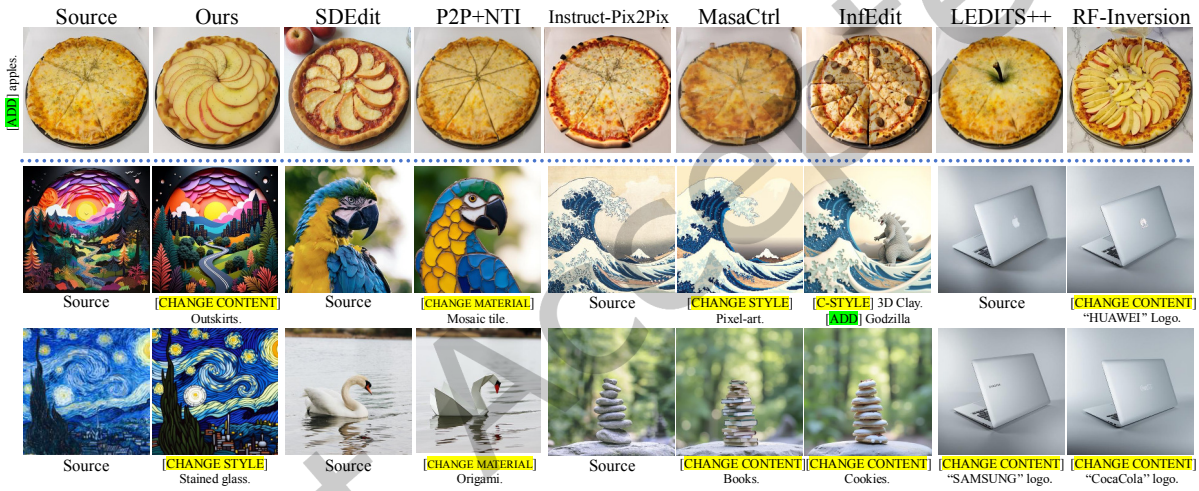

Fig. 1. Results of **HeadRouter** demonstrate accurate text-guided semantic representation while preserving consistency with the source image across diverse editing tasks.

*Corresponding author: Fan Tang.

Authors' Contact Information: Yu Xu, Institute of Computing Technology Chinese Academy of Sciences, Beijing, Beijing, China; e-mail: xuyu21b@ict.ac.cn; Fan Tang, Institute of Computing Technology Chinese Academy of Sciences, Beijing, Beijing, China; e-mail: tfan.108@gmail.com; Juan Cao, Institute of Computing Technology Chinese Academy of Sciences, Beijing, China; e-mail: caojuan@ict.ac.cn; Xiaoyu Kong, Beihang University, Beijing, Beijing, China; e-mail: xiaoykong15@gmail.com; Yuxin Zhang, NLPR, Chinese Academy of Sciences Institute of Automation, Beijing, Beijing, China and School of Artificial Intelligence, University of the Chinese Academy of Sciences, Beijing,

ACM 1557-7368/2026/3-ART
https://doi.org/10.1145/3797956

Diffusion Transformers (DiTs) have exhibited robust capabilities in image generation tasks. However, accurate text-guided image editing for multimodal DiTs (MM-DiTs) still poses a significant challenge. Unlike UNet-based structures that could utilize self/cross-attention maps for semantic editing, MM-DiTs inherently lack support for explicit and consistent incorporated text guidance, resulting in semantic misalignment between the edited results and texts. In this study, we disclose the sensitivity of different attention heads to different image semantics within MM-DiTs and introduce *HeadRouter*, a training-free image editing framework that edits the source image by adaptively routing the text guidance to different attention heads in MM-DiTs. Furthermore, we propose a dual-token refinement module to refine text/image token representations for precise semantic guidance and accurate region expression. Experiments on multiple benchmarks demonstrate HeadRouter's performance in terms of editing fidelity and image quality. The code is available at https://github.com/ICTMCG/HeadRouter.

CCS Concepts: • **Computing methodologies** → **Image processing**.

Additional Key Words and Phrases: Image editing, diffusion transformers, attention heads

1 Introduction

The introduction of the Diffusion Transformers [Peebles and Xie 2023] (DiTs) architecture, which combines diffusion processes with transformers [Dosovitskiy et al. 2020], has substantially augmented the scalability and performance of visual generation models (e.g., Sora [OpenAI 2024], PixArt-α [Chen et al. 2024], etc). Building upon these advances, the Multimodal Diffusion Transformers (MM-DiTs) have been proposed and developed by SD3 [Esser et al. 2024] and Flux [blackforestlabs.ai 2024]. The main difference between MM-DiTs and typical UNet-based [Ronneberger et al. 2015] text-to-image models [Podell et al. 2024; Rombach et al. 2022] is that MM-DiTs employ a joint self-attention architecture, which integrates image and text inputs. This architecture effectively captures inter-feature relationships and generates a more aligned feature space. As a result, MM-DiTs exhibit outstanding performance in diverse downstream tasks, such as controllable generation and personalized generation, with improved scalability, generation quality, and accuracy.

Despite these advancements, the MM-DiTs architecture poses unique challenges to semantic image editing tasks. Traditional UNet-based diffusion models [Huang et al. 2024b; Saharia et al. 2022; Zeng et al. 2024] incorporate text guidance through explicit cross-attention mechanisms. Cross-attention maps modulate the feature generation process by injecting semantic information from the text prompts. Crucially, in these architectures, the text condition remains static and external, enabling consistent and fine-grained control over the synthesized content throughout the denoising process. This mechanism allows previous image editing methods to perform various editing operations effectively.

In contrast, MM-DiTs inject text guidance through a joint self-attention operation, where text and image embeddings are concatenated and processed as a single sequence. While this integration enhances generation quality, it fundamentally alters the dynamics of editing. The text tokens in MM-DiTs are updated alongside image tokens in each layer. In this study (Sec. 3.2), we further reveal that the explicit interaction from text to image tokens naturally diminishes as the joint self-attention block progresses deeper. This "vanishing guidance" leads to a weakened semantic alignment in the deep layers, making it difficult to faithfully maintain the editing intent or preserve the source structure using traditional inversion techniques.

As a pioneering work to accomplish image editing based on MM-DiTs, RF-Inversion [Rout et al. 2024] puts forward an edit-friendly inversion method that uses dynamic optimal control via a linear quadratic regulator. However, even with optimal inversion trajectory, accurately capturing the semantic information to be edited in images and faithfully achieving text-guided editing still pose challenges due to the aforementioned diminishing guidance and entangled representation. Therefore, there persists a requirement for an accurate and high-quality image editing method that can be applied to MM-DiTs.

Beijing, China; e-mail: yuxin.zhazel@gmail.com; Jintao Li, Institute of Computing Technology Chinese Academy of Sciences, Beijing, China; e-mail: jtli@ict.ac.cn; Oliver Deussen, Universität Konstanz, Konstanz, Germany; e-mail: oliver.deussen@uni-konstanz.de; Tong-Yee Lee, National Cheng Kung University, Tainan City, Taiwan; e-mail: tonylee@mail.ncku.edu.tw.

To this end, we first explore the internal structure of MM-DiTs and reveal two valuable observations: the semantic sensitivity within the multi-head attention mechanism and the gradually diminishing feature interaction between text and image. Specifically, different from the previous study [Gandelsman et al. 2024] that disclosed that certain attention heads capture specific image properties in the CLIP-ViT [Dosovitskiy et al. 2020], we, by generating images from diverse text pairs with semantic differences and quantifying the similarity between different semantics and the output features of different heads, highlight that the various image semantics are adaptively distributed across different heads for MM-DiTs. Meanwhile, we explore the feature interaction between text and image tokens within the joint self-attention. We find that, unlike in UNet, due to the absence of explicit text guidance, the interaction would gradually weaken as the joint self-attention block progresses deeper.

In light of these observations, we present *HeadRouter*, a training-free image editing framework for MM-DiTs. Firstly, we propose an instance-adaptive attention head router (**IARouter**). IARouter adaptively activates attention heads according to their semantic sensitivity, thereby enabling more accurate capture of the semantics to be edited in the original images. Secondly, we further propose a dual-token refinement module (**DTR**). DTR employs self-enhancement of image tokens and text tokens to rectify the representation in the deep joint self-attention blocks. Moreover, our method ensures time efficiency that it does so by avoiding from importing model training and complex attention computations. In this way, as shown in Fig. 1, *HeadRouter* achieves satisfactory on text-guided image editing for MM-DiTs. Our contributions are summarized as follows:

- We conduct an analysis of the joint attention in MM-DiTs, and point out that attention heads respond differently to various editing semantics, and the interactions between text and image tokens in the joint attention would be gradually weaken.
- We propose *HeadRouter*, a novel training-free image editing method tailored for MM-DiTs, which includes an instance-adaptive router to enhance key attention heads for semantic representation and a dual-token refinement module for accurate text guidance and editing.
- Experiments on multiple text-guided image editing benchmarks demonstrate that our approach yields more accurate regional, semantic, and attribute-wise editing effects across diverse tasks, surpassing state-of-the-art baseline methods.

2 Related Work

2.1 Diffusion Transformers

The integration of transformers into diffusion models capitalizes on their capacity to model long-range dependencies. Diffusion Transformers (DiTs) [Peebles and Xie 2023] pioneered this integration for class-conditioned image generation, combining the strengths of diffusion processes and transformer architectures. Building on DiT, models such as PixArt-α [Chen et al. 2024], SD3 [Esser et al. 2024], and Flux [blackforestlabs.ai 2024] extended this framework to text-conditioned image generation. Specifically, SD3 and Flux utilize Multimodal Diffusion Transformers (MM-DiTs), entangling text and image modalities during training and inference. This entanglement enhances the interaction between textual and visual features, leading to images that better reflect the input text. Despite these advancements, existing methods treat attention heads uniformly without exploiting their potential for semantic-specific representation. Our approach addresses this gap by assigning different attention heads to specific editing semantics in the MM-DiTs. By enhancing token representations with text prompts, we achieve more precise control over image semantics, resulting in images that more accurately align with the textual descriptions.

2.2 Text-guided Training-free Image Editing

Training-free image editing approaches are considered fast and cost-effective since they eliminate the need for training or finetuning on data throughout the editing process [Huang et al. 2024a; Ma et al. 2024]. DDIM

inversion [Song et al. 2021] is a foundational approach for inversion-based methods [Garibi et al. 2024; Huberman-Spiegelglas et al. 2024; Mokady et al. 2023] that leverages deterministic denoising steps to invert a real image back into noise. InfEdit [Xu et al. 2024] implies a virtual inversion strategy without explicit inversion in sampling. However, MM-DiTs are mainly based on rectified flow models with ordinary differential equation [Albergo and Vanden-Eijnden 2023; Lipman et al. 2023; Liu et al. 2023], which is different from SDE-based diffusion models; thus, using the above method in MM-DiTs for image editing is less effective. RF-inversion [Rout et al. 2024] proposes a dynamic optimal control-based approach for inverting rectified flow models. However, the lack of exploration into the inherent mechanisms and characteristics of editing in MM-DiTs limits its ability.

Besides, attention modification techniques offer a direct and commonly used approach for training-free image editing [Cao et al. 2023; Liu et al. 2024; Parmar et al. 2023; Tumanyan et al. 2023]. Prompt-to-prompt [Hertz et al. 2023] enables prompt-based image editing by aligning spatial relationships in cross-attention layers, ensuring consistency between the edited result and the source image. Guide-and-Rescale [Titov et al. 2024] also uses a modified diffusion sampling process with self-guidance from attention maps. CDS [Nam et al. 2024] extracts the intermediate features of the self-attention layers and calculate loss to regulate structural consistency. However, removing the cross-attention mechanism in MM-DiTs and introducing joint input of text and image embeddings leads to entangled features of image and text, making it hard to explicitly utilize the cross-attention map for text-guided image editing. Therefore, we analyze the influence of text tokens on image tokens in MM-DiTs, identifying and extracting critical regions in the joint attention map to guide image editing.

2.3 Recent Advances in MM-DiT Editing

Following the adoption of Multimodal Diffusion Transformers (MM-DiTs) [Esser et al. 2024], several approaches have been proposed to address their unique editing challenges. **FlowEdit** [Kulikov et al. 2025] introduces an inversion-free framework by constructing a direct Ordinary Differential Equation (ODE) path between source and target distributions. While effective for structural preservation, relying solely on optimal transport trajectories can limit the magnitude of semantic changes, often failing to generate novel textures or shapes requested by the prompt. **Stable Flow** [Avrahami et al. 2025] and **FluxSpace** [Dalva et al. 2024] operate within the internal representations of the model. Stable Flow identifies "vital layers" for hard feature injection, similar to Plug-and-Play [Tumanyan et al. 2023] in UNets. However, such rigid injection often imposes excessive structural constraints, hindering geometry-altering edits. **FluxSpace** posits that joint attention blocks form a linear representation space, employing orthogonal projections to isolate editing directions. However, this approach hinges on the linear representation hypothesis and often necessitates auxiliary attention masking to maintain disentanglement, limiting its flexibility compared to adaptive mechanisms. Concurrently, **Flux.1 Kontext** [blackforestlabs.ai 2024] utilizes in-context learning by concatenating reference image tokens, offering strong generation capabilities but requiring computationally expensive training and often suffering from structural drift. In contrast to these methods, our **HeadRouter** addresses the root cause of editing failures in MM-DiTs—the diminishing text guidance in deep layers. By adaptively routing attention heads based on semantic sensitivity and refining tokens without rigid constraints or additional training, our method achieves a superior balance between semantic alignment and structural fidelity.

3 Analysis on Joint Self-attention

Before delving into the method introduction, we first conduct a detailed analysis of MM-DiTs, aiming to understand why the joint self-attention operation affects the accuracy and effectiveness of text-guided editing. Specifically, we will analyze from two aspects: the semantic sensitivity within the multi-head attention and the text-image token interactions in the joint self-attention calculation.

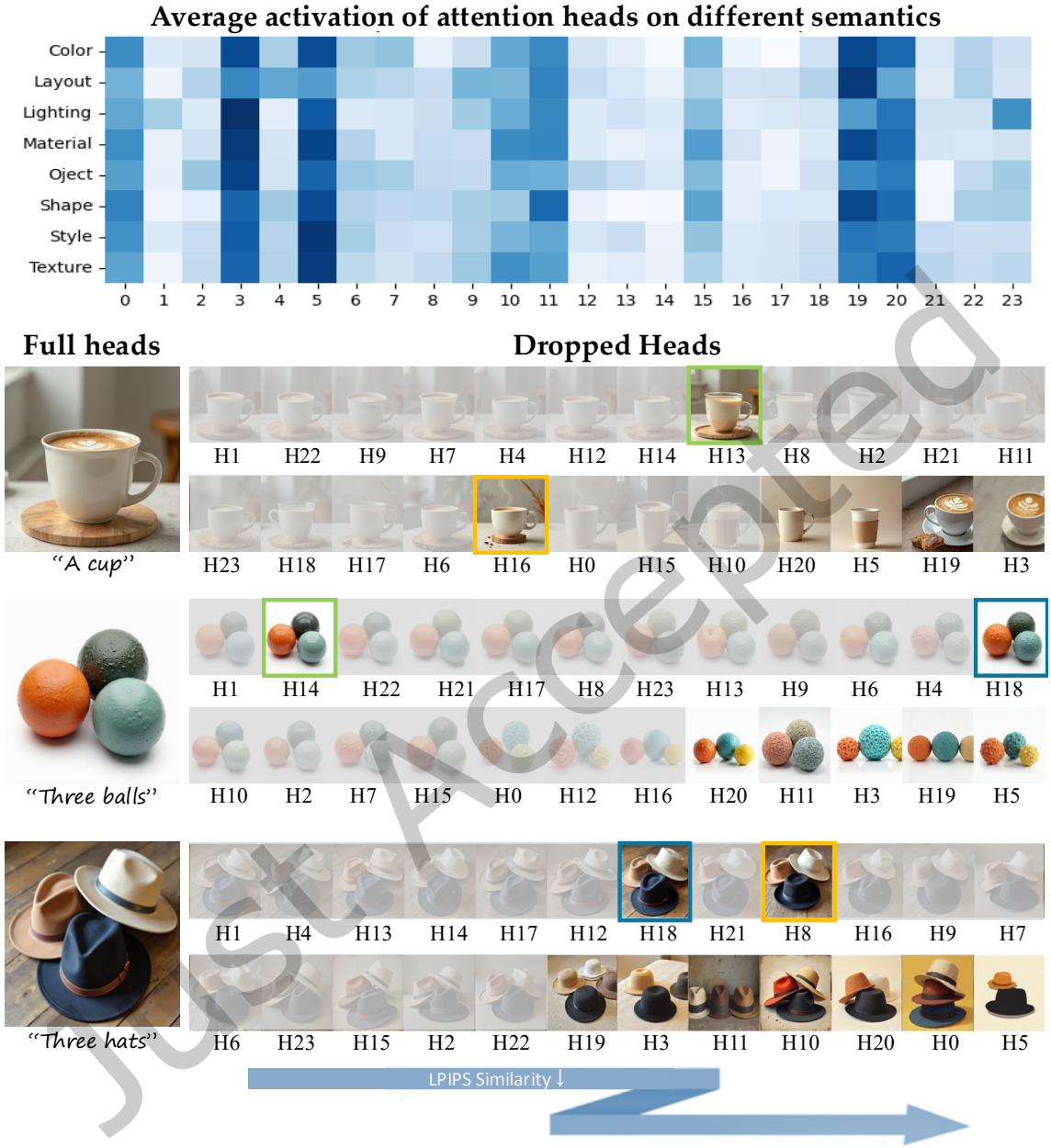

Fig. 2. **Sensitivity of attention heads to different semantics of multi-head attention in MM-DiTs.** (Top) We illustrate the distribution of distinct semantics across attention heads. (Bottom) Inference results by dropping individual heads, ordered by LPIPS similarity to the unmasked results. Dropping common activated heads (e.g., H3, H5, H19, H20) causes changes across multiple semantics, while masking specific heads leads to isolated semantic changes—material (green box), shape (yellow box), and texture (blue box).

3.1 Multi-head Attention Semantic Sensitivity

Attention calculation is the fundamental operation in Transformer architectures [Dosovitskiy et al. 2020; Vaswani 2017]. Particularly, the multi-head attention calculation enhances the representation power by conducting separate attention calculations on different projections, which can be written as:

$$\text{MultiHead}(Q, K, V) = Concat(head_1, ..., head_H)W^o, \tag{1}$$

where

$$head_h = \text{Attention}(QW_h^Q, KW_h^K, VW_h^V). \tag{2}$$

The W^o represents the projection for multi-attention output. Q, K, V indicate the Query, Key, and Value embeddings, respectively. Additionally, W_h^Q, W_h^K, W_h^V denote the projections for the h-th head, where H represents the total number of heads.

Previous studies have explored the function of multiple attention heads in models such as the CLIP encoder [Radford et al. 2021], revealing that different heads capture different correlations between text and images. This finding inspires us to conduct a more in-depth exploration on the influence of each individual attention head with regard to the representation of diverse editing semantics (e.g., shape, color, texture, style, etc.) in MM-DiTs. Specifically, we aim to examine whether particular attention heads exhibit heightened sensitivity to specific attributes and how we can utilize this understanding to achieve more efficient image editing.

To this end, we construct a diverse paired-text dataset $\mathcal{D}$ by GPT-4 to represent eight different editing semantics in the image editing benchmark (e.g., PIE-Bench [Ju et al. 2024], TEDBench++ [Brack et al. 2024] and EditEval [Huang et al. 2024a], etc.). Denoting the eight semantic categories as $S = \{s_1, s_2, \ldots, s_8\}$, for each semantic category $s \in S$, we define its corresponding vocabulary set as W_s. Additionally, we define $W_{-s} = W \setminus W_s$ as the vocabulary set excluding the current semantic category, where W represents the set of all vocabulary words. Then, the corresponding subset D_s is defined as:

$$\begin{aligned} D_s = (p_1, p_2) \mid\ & p_1 = f(w_1, u_1),\ p_2 = f(w_2, u_2), \\ & w_1, w_2 \in W_s,\ w_1 \neq w_2,\ u_1, u_2 \in W_{-s}, \end{aligned} \tag{3}$$

where $f(w, u) = [\text{“a }, w, u\text{”}]$ is the prompt construction function and “a” is the letter, and both p_1 and p_2 contain one word from the same semantic category s and another word form other semantic categories. Here, p_1 and p_2 are prompts containing different words from the same semantic category s and words from other semantic categories. For example, a pair of D_s is like (“a blue cup”, “a red circle”), “blue” and “red” are both color categories but different, “cup” and “circle” are different categories. Finally, we combine all subsets to form the complete dataset $\mathcal{D}$:

$$\mathcal{D} = \bigcup_{s \in S} D_s. \tag{4}$$

Each D_s contains 500 pairs, with the entire $\mathcal{D}$ containing 4000 pairs.

To analyze how different attention heads relate to specific editing semantics, we define a semantic relevance score for each head. To analyze how different attention heads relate to specific editing semantics, we define a semantic relevance score for each head. Specifically, let $\mathcal{D}_s$ denote the set of prompts associated with a specific semantic category $s \in \mathcal{S}$ (e.g., color, material). For the h-th attention head in the final MM-DiT block, let $\mathbf{O}_h(p)$ represent its output feature map given a prompt p. We first compute the raw activation score $a_{h,s}$ by averaging the output features across all prompts in the dataset $\mathcal{D}_s$:

$$a_{h,s} = \frac{1}{|\mathcal{D}_s|} \sum_{p \in \mathcal{D}_s} \mathbf{O}_h(p). \tag{5}$$

To quantify the sensitivity of different heads to the semantic s, we then compute the final semantic relevance score $R_{h,s}$ by applying min-max normalization across all attention heads $h' \in \{1, \ldots, H\}$:

$$R_{h,s} = \frac{a_{h,s} - \min_{h'} a_{h',s}}{\max_{h'} a_{h',s} - \min_{h'} a_{h',s}}. \tag{6}$$

This score $R_{h,s}$ reflects the relative importance of head h for representing semantic s compared to other heads. We visualize these scores as a heat map in the top part of Fig. 2.

Analysis is conducted on the last attention block, which aggregates and refines all prior features, making it the most directly relevant to the final output and providing the most meaningful insights into how heads affect editing different semantics. From the observations in the heat map of Fig. 2, we conclude that: (1) attention heads and semantics exhibit a sparse correlation pattern, and (2) this correlation lacks global consistency across heads. For instance, certain heads such as H3, H5, H19 and H20 are associated with multiple semantics rather than being localized to a single one. Based on this analysis, we further investigate individual samples by dropping each attention head during inference and sorting the results by LPIPS [Zhang et al. 2018] similarity to the full-heads output. As shown in the bottom part of Fig. 2, dropping heads with high activation to multiple semantics (e.g., H3, H5, H19, H20, highlighted in the heatmap) leads to changes across several semantic aspects and yields low similarity with full-heads results. In contrast, dropping other heads can trigger specific semantic changes: for example, masking H13 and H16 alters the material and shape of the "cup", respectively; masking H14 and H18 affects the material and texture of the "balls"; and masking H18 and H8 modifies the texture and shape of the "hats". These results indicate that the attention heads sensitive to specific semantics vary across instances, i.e., different instances rely on different heads for the same semantic attribute. Hence, accurately capturing the corresponding edited semantics is a challenging task.

3.2 Text-image Token Interactions

Previous text-guided image editing methods mainly utilize cross-attention mechanisms to incorporate textual guidance [Hertz et al. 2023]. Formally, in cross-attention, the features from the noisy image $\phi(z_t)$ are projected to query $Q_{ca} = P_Q(\phi(z_t))$ and the text embedding $\psi(p)$ is projected to key $K_{ca} = P_K(\psi(p))$ and value $V_{ca} = P_V(\psi(p))$, where P_Q, P_K and P_V are pre-trained linear projections. Then, an attention map can be explicitly obtained via:

$$\mathcal{M}(Q_{ca}, K_{ca}, V_{ca}) = \text{softmax}\left(\frac{Q_{ca}K_{ca}^T}{\sqrt{d_k}}\right), \tag{7}$$

and can be directly used for various operations.

However, MM-DiTs presents a distinct paradigm that text and image token embeddings are combined into a single embedding and then processed by transformer blocks using joint attention. Formally, the input to the joint attention can be formalized as follows:

$$Q_{ja} = P_Q^I(\phi(z_t)) \odot P_Q^T(\psi(p_t)), \tag{8}$$

$$K_{ja} = P_K^I(\phi(z_t)) \odot P_K^T(\psi(p_t)), \tag{9}$$

$$V_{ja} = P_V^I(\phi(z_t)) \odot P_V^T(\psi(p_t)), \tag{10}$$

where P_Q^I, P_K^I, P_V^I, P_Q^T, P_K^T and P_V^T are pre-trained linear projections for image and text embeddings, p_t represents the text prompt embedding incorporated at timestep t, and $\odot$ represents the concatenation operation. Then Q_{ja}, K_{ja} and V_{ja} are calculated with attention formulation:

$$\text{Attention}(Q_{ja}, K_{ja}, V_{ja}) = \text{softmax}\left(\frac{Q_{ja}K_{ja}^T}{\sqrt{d_k}}\right)V_{ja}. \tag{11}$$

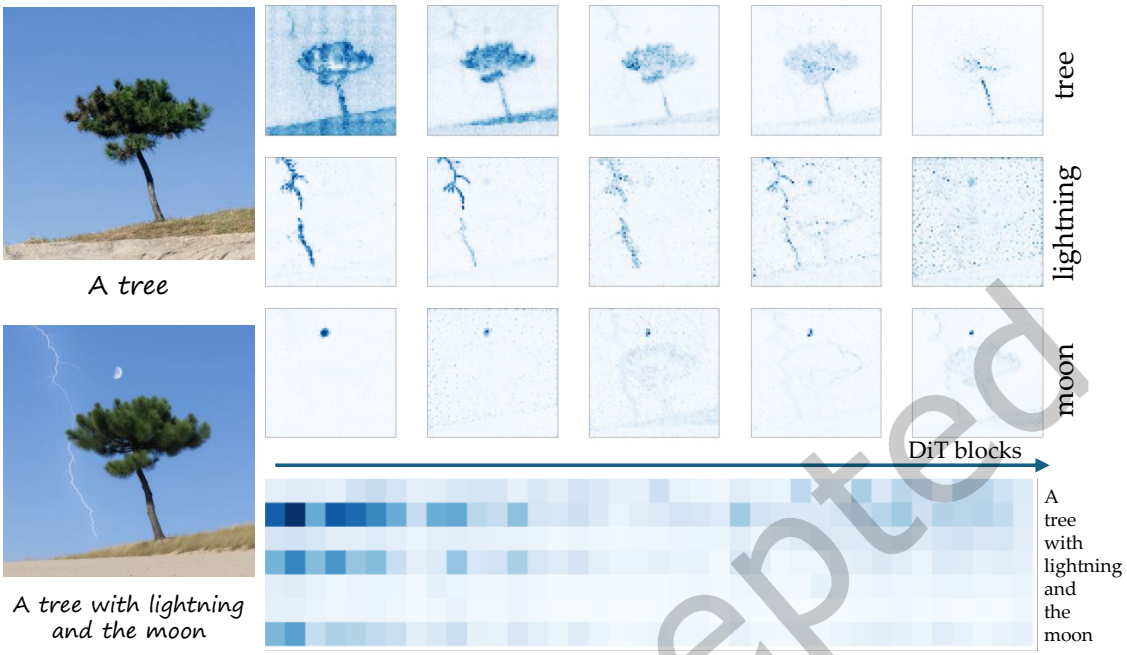

Fig. 3. **Analysis of text guidance on image tokens.** Key image regions influenced by text guidance are identified within the joint self-attention map and visualized. Additionally, we observe that text guidance influence diminishes as attention blocks progress in a single denoising step, leading to weakened semantic representation from text-guided editing.

This joint attention among text and image tokens allows for more intricate interactions and makes influence dynamics less explicit.

We delve into the influence of text tokens on image tokens within the joint attention mechanism. Firstly, we conduct an analysis of the manner in which text tokens directly influence image tokens during the generation process. The influence of each token on the image is weighted by computing the attention weights within the joint self-attention map. Specifically, we count the lower-left area of this map, where each column reveals the impact of every word, and each row shows how the intent of each image token is influenced. Subsequently, we reshape the weights of each row into a 64×64 grid for the purpose of constructing a heatmap. As shown in the top of Fig. 3, the heatmap discloses that image tokens are responsive to relevant textual descriptions, thereby signifying a robust alignment between text and image tokens within this particular region.

However, as the number of DiT blocks passed increases, the activation of each text on the image gradually weakens or becomes chaotic. Actually, unlike UNet-based models where the text embedding remains unchanged once encoded, text embeddings in MM-DiTs are integrated with image embeddings and participate in the joint self-attention calculation. As a result, the text embeddings evolve progressively in tandem with the attention blocks. Unfortunately, when extending this analysis across the entire network, we observe that the text guidance weakens as the block depth increases. To effectively quantify this, we compute the average attention weight of each text token at every block. Specifically, for the j-th text token at the l-th block, we average its attention weights over all attention heads and all spatial image tokens (the 64×64 grid mentioned above). Formally, the

influence score $s_j^{(l)}$ is calculated as:

$$s_j^{(l)} = \frac{1}{H \cdot N_I} \sum_{h=1}^{H} \sum_{i=1}^{N_I} \mathbf{M}_{h,i,j}^{(l)}, \tag{12}$$

where H is the number of heads, N_I is the number of image tokens, and $\mathbf{M}^{(l)}$ represents the text-to-image attention sub-matrix. As shown at the bottom of Fig. 3, visualizing these scores $s_j^{(l)}$ reveals that the text influence on the image-generation process gradually diminishes in deeper blocks. Consequently, the text influence on the image-generation process gradually diminishes. When it comes to text-guided image editing, this problem becomes more severe. Unlike text-to-image generation, which starts from Gaussian noise and allows the model to fully synthesize content aligned with the prompt, image editing begins with an inverted latent code of an existing image. This code already encodes much of the original structure, creating a tension between preserving the original image and making meaningful edits. This difference makes prompt alignment more challenging in editing than in text-to-image generation, and explains why diminishing text influence is more problematic in editing tasks. Therefore, maintaining the activation of the editing text embedding is also a crucial issue for MM-DiTs image editing.

4 Method

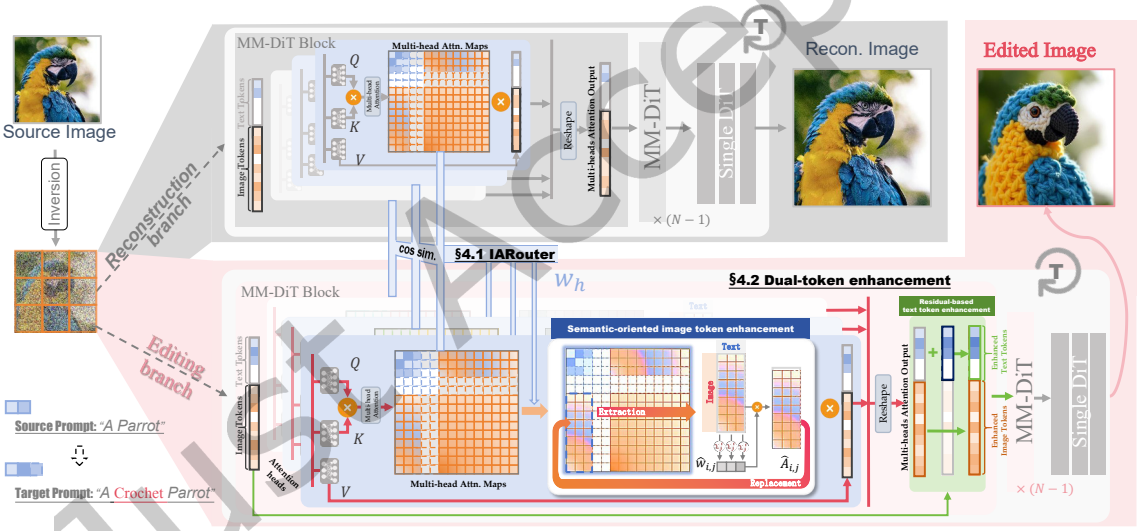

Fig. 4. **Pipeline of our method.** We mainly introduce instance-adaptive attention head router **(IARouter)** to adaptively activate attention heads based on their semantic sensitivity, enabling a more accurate representation of the edited specific images. We also propose dual-token enhancement module for better alignment of edited images and prompts.

As shown in Fig. 4, *HeadRouter* includes two modules: **Instance-adaptive Attention Head Router (IARouter)** (Sec.4.1), which dynamically identifies the attention heads correlted with the target editing semantics and improves the representation by emphasizing the most effective attention heads; and **Dual-token refinement module (DTR)** (Sec.4.2), which refines edits on key image tokens by applying attention weights from text to image tokens and preserves the editing text guidance from dissipating.

4.1 Instance-adaptive Attention Head Router

Building upon our analysis of attention heads' sensitivities to different editing semantics, we aim to identify and emphasize the most effective attention heads for specific editing tasks. By leveraging information from the image reconstruction branch, we guide the image editing branch to focus on the most relevant attention heads, thereby improving editing effectiveness.

We first identify target attention heads that are most sensitive to the desired editing semantics. Given a DiT model with H attention heads, we begin by calculating the cosine similarity between the outputs of corresponding attention heads when generating images with and without a specific semantic, respectively. Let $\mathbf{v}_h^{(r)}$ and $\mathbf{v}_h^{(e)}$ denote the output features of the h-th attention head in the reconstruction branch and the editing branch (i.e Eq. 2), respectively, the cosine similarity s_h for head h is calculated as:

$$s_h = \frac{\mathbf{v}_h^{(r)} \cdot \mathbf{v}_h^{(e)}}{\|\mathbf{v}_h^{(r)}\| \|\mathbf{v}_h^{(e)}\|}, \tag{13}$$

where $\cdot$ denotes the dot product operation and $\|\cdot\|$ represents the Euclidean norm.

To quantify the sensitivity of each attention head to the specific semantics, we use the min-max normalized dissimilarity score $\tilde{d}_h$ for head h:

$$\tilde{d}_h = \frac{s_{\max} - s_h}{s_{\max} - s_{\min}}, \tag{14}$$

where $s_{\min} = \min\{s_1, s_2, \ldots, s_H\}$, $s_{\max} = \max\{s_1, s_2, \ldots, s_H\}$. The normalized score $\tilde{d}_h$ reflects how dissimilar each attention head's outputs are concerning the specific semantics relative to the range of dissimilarities observed across all heads.

To smoothly activate the most semantic-sensitive attention heads, we propose an instance-adaptive attention head router to enhance the representation of editing semantics. IARouter is designed to (1) highlight dissimilar heads: assign high attention to heads with lower $\tilde{d}_h$ to emphasize their importance in representing the desired editing semantics; (2) maintain similar heads: ensure that the contributions of heads that are less relevant to the edits are not excessively altered, thereby maintaining the integrity of other visual aspects in the image; (3) smooth weights: avoid artifacts by preventing sudden weight changes and maintain model stability.

Based on these objectives, IARouter uses soft activation on attention heads. The weight w_h for head h is defined as:

$$w_h = 1 + \gamma \cdot \sigma\left(\tilde{d}_h - \delta\right), \tag{15}$$

where γ is the maximum weight increment, δ shifts the center of the sigmoid, and $\sigma(x)$ is the sigmoid function.

During the image generation process, we multiply the output of each attention head by its corresponding weight to obtain the enhanced output:

$$\mathbf{v}_h^{\text{enhanced}} = w_h \cdot \mathbf{v}_h^{\text{e}}. \tag{16}$$

The proposed IARouter serves as a smooth semantic-specific enhancer. By identifying and emphasizing heads sensitive to specific semantics, IARouter allows for more precise and effective edits. Using a sigmoid function allows for a gradual increase in weights, preventing sudden changes that could introduce artifacts.

4.2 Dual-token Refinement Module

As discussed in Sec. 3, attention weights between text and image tokens reflect the influence of the text prompt on each image token. We make use of these weights to concentrate edits on the key image regions that correspond to the desired semantics for semantic refinement. Furthermore, we put forward the idea of modifying the attention normalization to enhance the influence of significant text tokens on the image tokens.

Semantic-oriented image token enhancement. The joint self-attention mechanism in MM-DiTs produces attention weights that reflect the influence of text tokens on image tokens. To be more specific, for every image token, the attention weights related to text tokens denote the extent to which the image token is "attached to" each text token. We make full use of this characteristic to identify and concentrate on the image tokens that are most strongly influenced by the editing prompt.

Let $\mathbf{A} \in \mathbb{R}^{N \times M}$ denote the attention weight matrix from text tokens to image tokens, where N is the number of image tokens and M is the number of text tokens. The element $\mathbf{A}_{i,j}$ represents the attention weight from text token j to image token i.

We propose **semantic-oriented image token enhancement** to focus the editing on key image tokens, taking into account the impact of text on different image tokens. Formally, the weight mapping is defined as:

$$\hat{w}_{i,j} = \alpha \cdot \sigma\left(v \cdot \frac{e^{\mathbf{A}_{i,j}}}{\sum_{k=1}^{N} e^{\mathbf{A}_{k,j}}}\right), \tag{17}$$

where $\mathbf{A}_{i,j}$ indicates the attention weight of j-th text token to i-th image token. We using a $softmax$ based function that normalizes attention weights of image tokens, and using sigmoid functions to limit the growth of large weights. α is a weight enhancement coefficient, and v is for amplitude adjustment. Further discussion on the influence of α and v can be found in the supplementary materials.

Next, we reweight the attention in the editing branch using the normalized weights $\hat{w}_{i,j}$. The final image tokens $\hat{\mathbf{A}}_{i,j}$ are computed as:

$$\hat{\mathbf{A}}_{i,j} = \hat{w}_{i,j} \cdot \mathbf{A}_{i,j}. \tag{18}$$

This formulation ensures image tokens highly influenced by the text prompt (with higher $\hat{w}_{i,j}$) are assigned high weights, while tokens less influenced remain close to the original.

Residual-based text token enhancement. As attention weights between text and image tokens decay through successive attention blocks, we leverage residual text tokens to retain text guidance in each transformer block. Specifically, this design carries over text guidance from the previous attention block to the current one. As information progresses through deeper blocks, the previous block's input is incorporated as the residual term and combined with the current block's input, enhancing the continuity of text guidance. This mechanism introduces consistent text information into each block, strengthening text guidance and improving the accuracy of image editing.

5 Experiments

Table 1. **Comparison of cosine similarity between DINO features (for image) and CLIP (for text) of the edited images and source images and prompts, respectively**. Our method has the best scores, indicating that our approach successfully edit the image guided by text while maintain consistency with source image and high image quality.

Methods	SDEdit	P2P+NTI	Pix2Pix	MasaCrtl	InfEdit	LEDITS++	RF-Inver.	Ours
Structure-alignment (DINO) (↑)	0.8409	0.8559	0.8722	0.8744	0.8909	0.8963	0.9032	**0.9194**
Prompt-alignment (CLIP) (↑)	0.3051	0.2944	0.2975	0.2955	0.3016	0.3022	0.3109	**0.3203**
Image-quality (LPIPS) (↓)	0.2236	0.2258	0.2975	0.3144	0.2702	0.2796	0.2265	**0.2103**

In this section, we first introduce the experiment settings, then we present qualitative experiments in Sec. 5.2 and quantitative experiments in Sec. 5.3. We then conduct ablation studies on proposed modules, activation weighting formulations, different semantic categories, and hyperparameters in Sec. 5.5. To validate the argument about the challenge of using MM-DiTs for editing, we conduct qualitative and quantitative analysis in Sec. 5.6. Last,

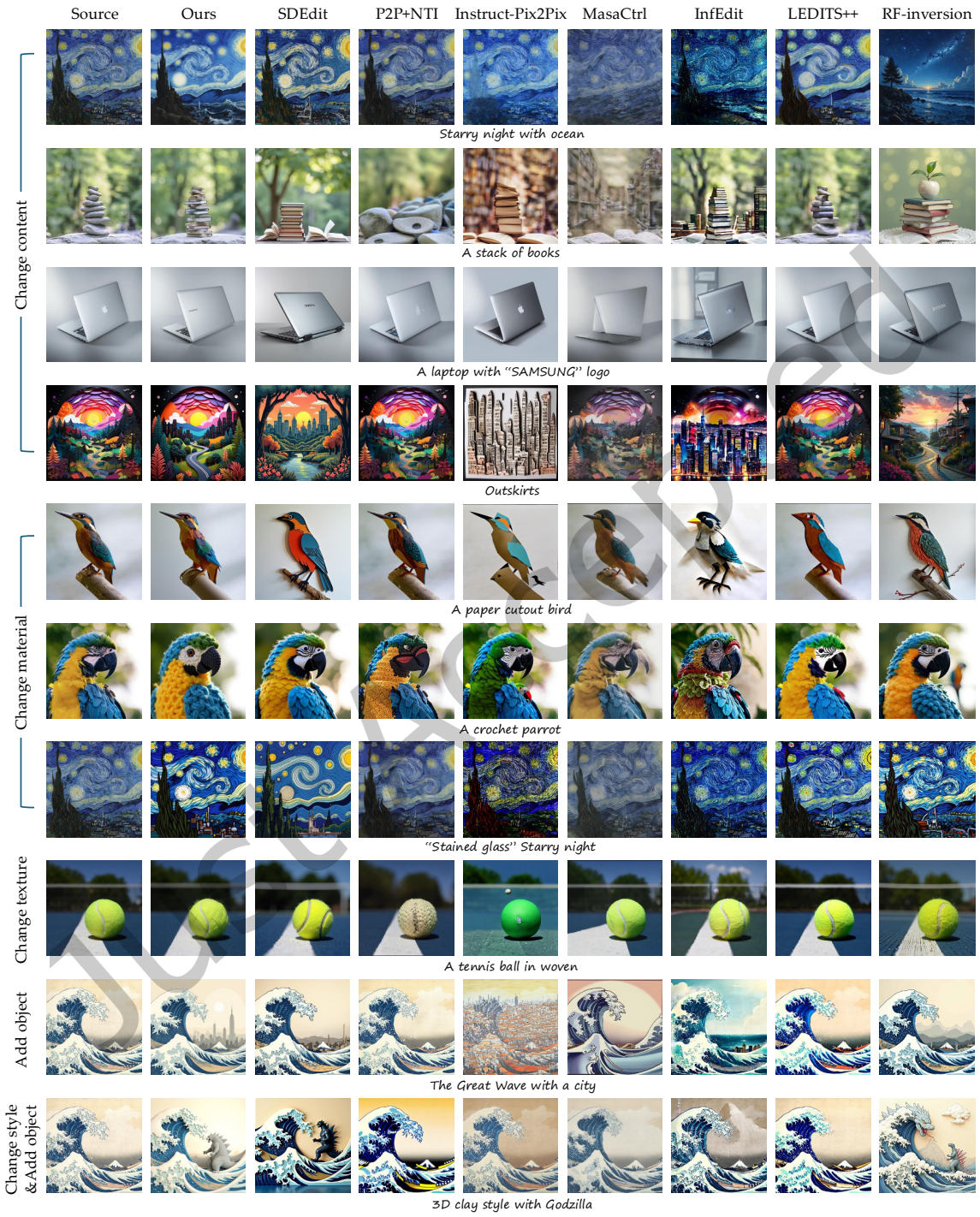

Fig. 5. **Qualitative comparison with baseline methods on various editing tasks.** Our results demonstrate high alignment with the text guidance while keeping consistency with the reference image.

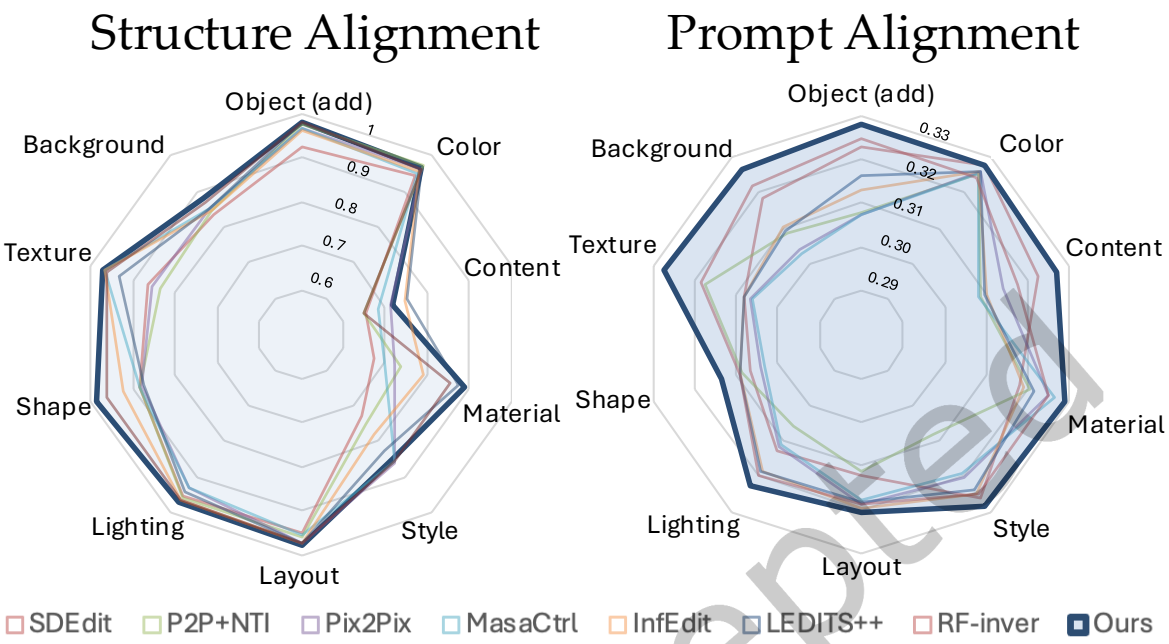

Fig. 6. **Radar chart for evaluating image and prompt alignments in eight editing tasks.** Overall, our approach effectively retains the intrinsic feature of the original image while aligning precisely with the specified text guidance.

we explore the generalization of our method to MM-DiTs variants and adaptability to various inversion-based editing approaches in Sec. 5.7.

5.1 Experimental Setup

Implementation details. In our experiments, we utilize Flux-1.0[dev] [blackforestlabs.ai 2024] with default hyperparameters, we leverage RF-Inversion [Rout et al. 2024] to invert a real image to its latent space, and basic settings are followed with them.

Baselines. We compare our method with seven state-of-the-art text-guided image editing approaches, including two Flux-based approaches: RF-inversion [Rout et al. 2024] and SDEdit [Meng et al. 2021] and five UNet-based approaches: Null-textual Inversion [Mokady et al. 2023] with prompt-to-prompt [Mokady et al. 2023], Instruct-Pix2Pix [Brooks et al. 2023], MasaCtrl [Cao et al. 2023], InfEdit [Xu et al. 2024], and LEDITS++ [Brack et al. 2024], all the approaches are training free.

Datasets. We evaluate our method with baselines on two text-guided image editing benchmarks: TED-Bench++ [Brack et al. 2024], a revised extension of TEDBench [Kawar et al. 2023] that contains 120 entities in total, and PIE-Bench [Ju et al. 2024], which comprises 700 images, each associated with ten distinct editing types.

Evaluation Metrics. Following previous text-guided image editing work [Brack et al. 2024; Huberman-Spiegelglas et al. 2024; Nam et al. 2024; Xu et al. 2024], we evaluate the proposed method on three metrics: results of overall image quality, alignment with text guidance, and structure consistency with source images. Specifically, following baseline image editing methods, we use LPIPS [Zhang et al. 2018] to assess overall quality since it helps assess whether unwanted distortions are introduced and whether the edited semantics align with human perception.

Besides, we use CLIP-T [Ilharco et al. 2021] to measure text alignment, and DINO [Oquab et al. 2023] to evaluate structure consistency with the original image. Additionally, we conduct a user study to further assess performance.

5.2 Qualitative Comparison

In Fig. 5, we present the visual results of different editing types in comparison with baseline methods. SDEdit [Meng et al. 2021] is capable of generating new concepts within textual conditions (in the $2^{nd}/3^{rd}/4^{th}/10^{th}$ rows), but it struggles to change the material, texture and style of the source images (in the $5^{st} \sim 8^{th}$ rows). P2P+NTI [Mokady et al. 2023] is difficult to achieve satisfactory image editing results, often neglecting the information contained within textual conditions (in the $1^{st}/4^{th}/7^{th}/8^{th}$ rows). Instruct-Pix2Pix [Brooks et al. 2023] also struggles with image editing instructions that involve significant changes, leading to semantic loss (in the $1^{st}/2^{nd}/4^{th}/9^{th}$ rows) or inaccurate editing (in the $6^{st}/7^{th}/10^{th}$ rows). MasaCtrl [Cao et al. 2023] and InfEdit [Xu et al. 2024] similarly fail to accurately preserve the semantics of source images (in the $2^{nd} \sim 4^{th}$ rows) and are inaccurate in editing (in the $5^{th} \sim 10^{th}$ rows). LEDITS++ [Brack et al. 2024] achieves editing effects in experiments that alter the material of some images. However, there is still an issue with specific semantic editing (in the $1^{st} \sim 4^{th}/8^{th} \sim 10^{th}$ rows), and the loss of details in source images (in the 5^{th} row). RF-Inversion [Rout et al. 2024] struggles to achieve robust image editing effects, particularly in preserving the structural integrity of source images when modifying content (as seen in the 1^{st}, 2^{nd}, and 4^{th} rows). It also fails to accurately capture the characteristics of materials and textures when changing them. Our approach achieves the best structural preservation and editing effects, surpassing the performance of baseline methods.

5.3 Quantitative Comparison

We reported the quantitative results across the entire dataset, including both TEDBench++ [Brack et al. 2024] and PIE-Bench [Ju et al. 2024]. Tab. 1 presents a comprehensive comparison of image structure alignment between the edited and source images, alignment between the edited images and the textual guidance, and the overall generation quality. We further assess image-text alignment across ten distinct editing types, as illustrated in the radar chart in Fig. 6. In the "change content" category, although InfEdit and LEDITS++ achieve comparable metrics, their text alignment scores are substantially lower than ours, indicating that these methods are less effective at executing meaningful object modification edits. Similar findings are observed in Fig. 5. Additionally, the "change content" score is generally lower than other metrics, as content edits introduce significant alterations in key image regions, thereby reducing structural similarity with the original image. Nevertheless, our method consistently outperforms the baselines across all measures.

User study. We conduct a user study focusing on two primary aspects: alignment with the given prompt and preservation of irrelevant regions in the image. We generated 50 groups of images across various editing tasks, each containing eight images generated by our method and seven generated by baseline methods using the same prompt. 56 participants were presented with each group of images and asked to select the image that best aligned with the prompt and the original image. Results of Fig. 7 indicate that the results of our method closely follow the prompt while retaining the quality of regions not relevant to the editing prompt.

5.4 Experimental Comparison with Concurrent Works

We compare *HeadRouter* with three representative recent works: Flow-Edit [Kulikov et al. 2025], Stable-Flow [Avrahami et al. 2025], and the instruction-based Flux-Kontext [blackforestlabs.ai 2024]. As illustrated in Fig. 8, *HeadRouter* demonstrates superior performance in balancing editing strength and structural preservation. Stable-Flow relies on flow matching inversion but struggles to inject strong semantic changes, often resulting in outputs nearly identical to the source. Flow-Edit achieves better editing but often lacks fine-grained semantic alignment. For instance, in the 3D clay style task, while it adds the object, it fails to transform the overall texture into the

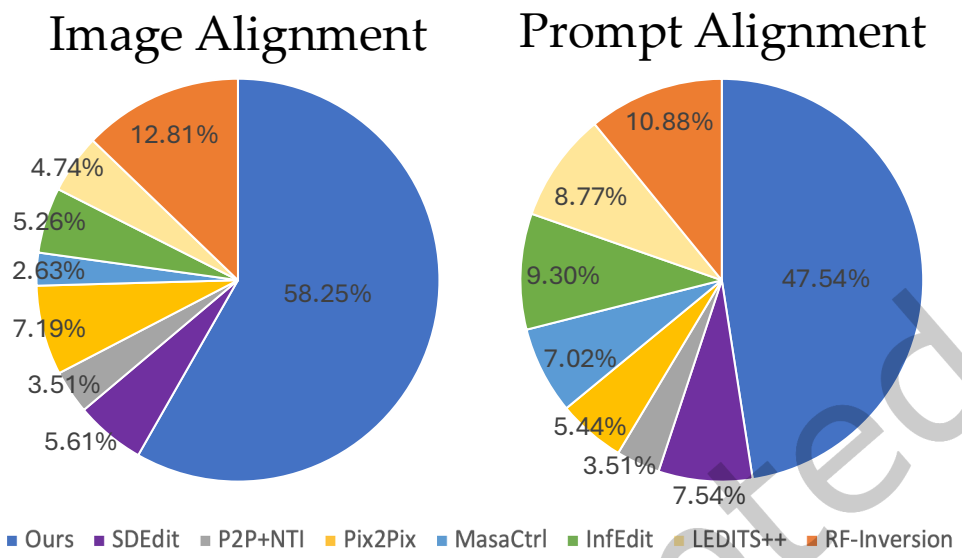

Fig. 7. **Statistics for user study on alignments.** Results of our method achieve the best preferences.

target clay material. Flux-kontext, being an instruction-based model, exhibits limitations in precise instruction following and composite editing. As observed in the "Stained glass" case, it fails to faithfully translate the source into the target style. Furthermore, in the composite task of "3D clay style with Godzilla", it demonstrates partial prompt adherence—successfully applying the clay texture but neglecting to generate the specified object (missing "Godzilla"). In contrast, *HeadRouter* utilizes our proposed IARouter to precisely locate and modulate semantic-relevant heads. This allows for accurate attribute changes (e.g., stained glass style, distinct woven texture, 3D clay style) while strictly maintaining the spatial layout of the source image.

We further conduct a quantitative comparison with the concurrent methods, as reported in Tab. 2. Consistent with our qualitative observations, Stable-flow exhibits the lowest prompt-alignment score, quantitatively confirming its limitation in injecting new semantic content. While Flux-kontext and Flow-edit show competitive prompt alignment, they suffer from lower structure-alignment and degraded image quality (higher LPIPS), reflecting their tendency towards structural drift and visual artifacts. In contrast, ***HeadRouter*** achieves the highest scores across all metrics. This quantitative superiority validates that our head routing mechanism effectively disentangles semantic editing from structural layout, enabling precise text-guided modifications without compromising the integrity of the source image.

Table 2. **Quantitative results of comparing with concurrent works.**

Methods	Flow-edit	Stable-flow	Flux-kontext	Ours
Structure-alignment (↑)	0.8967	0.9055	0.8994	0.9194
Prompt-alignment (↑)	0.3187	0.3101	0.3175	0.3203
Image-quality (↓)	0.2371	0.2231	0.2275	0.2103

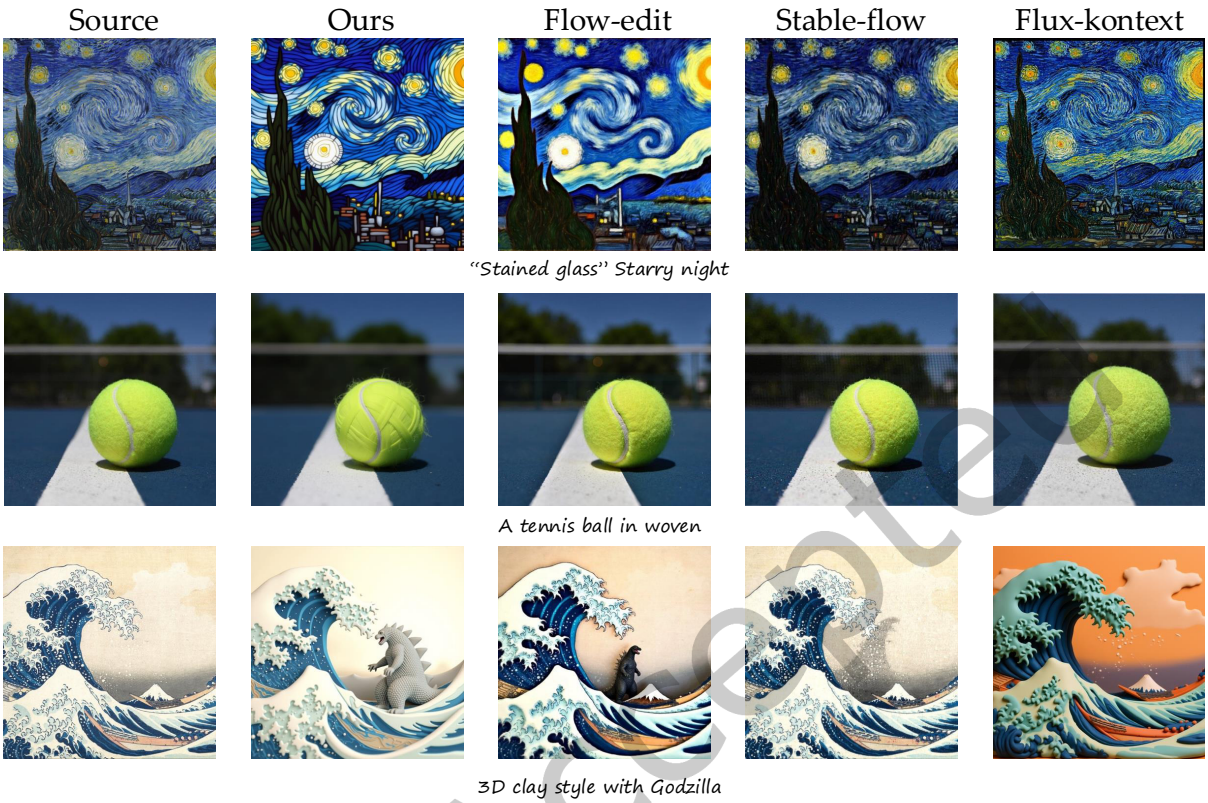

Fig. 8. **Qualitative comparison with concurrent works. Our method allows for accurate and diverse attribute changes.**

5.5 Ablation Study

In this section, we validate the effectiveness of the two key modules (i.e., IARouter and DTR) as well as the formulations designed in our methods.

Ablation on IARouter and DTR.. First, for the IARouter ablation, we remove any head constraints during inference. As shown in the lower left of Fig. 9, this results in weaker semantic expression (e.g., in the "apple" and "origami" examples, although the desired editing semantics are somewhat achieved, residual textures from the original tennis ball image persist). In contrast, IARouter enhances the expression of specific semantics by routing different heads according to semantic content. Next, we ablate DTR, with results shown in the lower right of Fig. 9. The results indicate that by strengthening image tokens and text guidance, our method captures the desired semantics and achieves finer-grained semantic representation in response to detailed text guidance. We also present the quantitative results in Tab. 3, which demonstrates that the ablation of IARouter or DTR leads to significant decreases in both structure alignment and prompt alignment. This indicates that the proposed IARouter and DTR effectively edit images while preserving their structural integrity. Furthermore, the image quality remains largely unchanged, confirming that our method does not introduce artifacts. More ablation studies on IARouter and DTR are provided in the supplementary material.

Table 3. **Quantitative results of ablation study on IARouter and DTR.** Without the proposed IARouter or DTR, the structure alignments, prompt alignments and image quality are all decreased.

Ablations	w/o IARouter	w/o DTR	Ours
Structure-alignment (↑)	0.9035	0.9101	0.9194
Prompt-alignment (↑)	0.3097	0.3117	0.3203
Image-quality (↓)	0.2105	0.2106	0.2103

Ablation on Activation Weighting Formulations. Eq. 15 employs a sigmoid function for two primary reasons: (1) It assigns higher weights to dissimilar heads, thereby activating desired semantic changes, while maintaining weights near 1 for similar heads to preserve original content fidelity. (2) The sigmoid function ensures a smooth transition across different heads, preventing artifacts that might arise from abrupt weight changes. Similarly, Eq. 17 utilizes a sigmoid function to constrain the growth of large weights. To validate our design choice for the attention head activation weighting formulations, we compare our sigmoid-based approach with two alternative mapping formulations of the dissimilarity score $\tilde{d}_h$. Specifically, we first test a linear scaling variant:

$$w_h = 1 + \mu \cdot \tilde{d}_h, \tag{19}$$

where μ is set to 2, and then test an exponential scaling variant:

$$w_h = \exp(\tilde{d}_h). \tag{20}$$

We present qualitative comparison results in Fig. 10 and quantitative comparison results in Tab. 4. The linear variant applies uniform scaling to all attention heads, which fails to sufficiently activate the most semantically relevant ones. As a result, it struggles to highlight key differences across heads, making it less effective in expressing specific editing attributes. The exponential formulation, while sharply boosting dissimilar heads, tends to overly focus on a few top-ranked heads with large weights, which can lead to undesired distribution shifts in the image space and visual artifacts. In contrast, our sigmoid-based formulation softly and smoothly highlights the most dissimilar heads while preserving the contributions of others, offering both instance-adaptivity and model stability. Quantitative results and qualitative comparisons validate that our formulation achieves the best prompt alignment and visual consistency.

Table 4. **Quantitative results of ablation study on activation weighting formulations.** Leveraging linear-based formulation or exponential-based formulation lead to sub-optimal editing results.

Ablations	Linear scaling	Exponential scaling	Ours
Structure-alignment (↑)	0.9018	0.8987	0.9194
Prompt-alignment (↑)	0.3126	0.3075	0.3203
Image-quality (↓)	0.2167	0.2833	0.2103

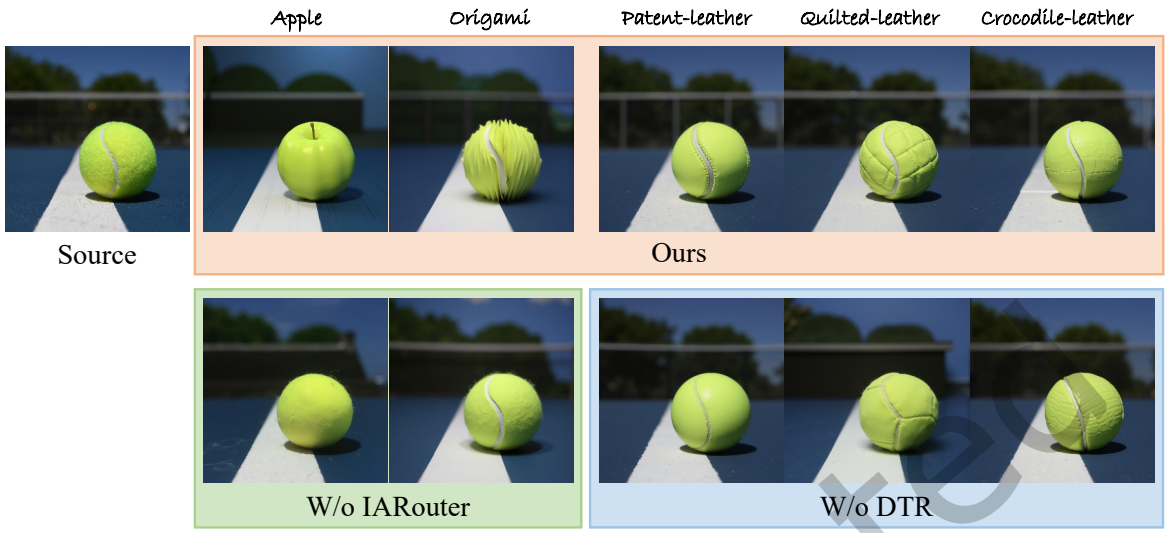

Fig. 9. **Visual analysis for IARouter and DTR.** Without **IARouter**, results tend to weaker semantic representation and retain the original semantics. With **DTR**, we achieve more fine-grained semantic representation.

Ablation on different semantic categories. For more details, we present ablation results for each module on different semantic categories in Tab. 5. Without the IARouter, the image structure alignment and text alignment are decreased, and "color", "material", "style", and "texture" have a more significant decrease, indicating IARouter and DTR have a significant effect on these categories. Besides, image quality remains largely unchanged, confirming our method does not introduce artifacts.

Table 5. **Quantitative ablation results for each module on different semantic categories.**

Ablations	Metrics	Obj(Add)	Color	Content	Material	Style	Layout	Lighting	Shape	Texture	Background
W/o IAR	Structure-alignment(↑)	0.9643	0.9493	0.6921	0.8885	0.8402	0.9533	0.9601	0.9721	0.9613	0.8538
	Prompt-alignment(↑)	0.3185	0.3107	0.3148	0.3128	0.3118	0.3062	0.3067	0.2904	0.3095	0.3156
	Image-quality(↓)	0.2085	0.2097	0.2133	0.2084	0.2156	0.2082	0.2071	0.2066	0.2108	0.2168
W/o DTR	Structure-alignment(↑)	0.9702	0.9577	0.7041	0.8887	0.8492	0.9629	0.9674	0.9755	0.9627	0.8626
	Prompt-alignment(↑)	0.3201	0.3135	0.3173	0.3136	0.3129	0.3083	0.3094	0.2918	0.3131	0.3174
	Image-quality(↓)	0.2083	0.2098	0.2136	0.2081	0.2159	0.2081	0.2074	0.2065	0.2109	0.2174
Ours	Structure-alignment(↑)	0.9815	0.9641	0.7198	0.8893	0.8522	0.9786	0.9715	0.9857	0.9722	0.8789
	Prompt-alignment(↑)	0.3264	0.3258	0.3233	0.3281	0.3265	0.3118	0.3156	0.2973	0.3252	0.3234
	Image-quality(↓)	0.2082	0.2101	0.2135	0.2089	0.2152	0.2078	0.2066	0.2071	0.2095	0.2161

Ablation on γ for attention heads weights. In Eq. 15, γ controls the activation weights extend on attention heads. We include ablation results on γ in Tab. 6. The results indicate that $\gamma > 0.5$ offers minor image alignment and prompt alignment gains but significantly decreases image quality (higher LPIPS). $\gamma = 0.5$ optimally balances editing accuracy and image quality.

5.6 UNet-based Image Editing Methods for MM-DiTs

To better understand the challenges of applying UNet-based image editing techniques to MM-DiTs architectures, we conduct a comparative study with two representative editing methods originally designed for UNet-based denoisers, such as those in Stable Diffusion 1, 2, and XL with explicit cross-attention: Prompt2Prompt [Hertz

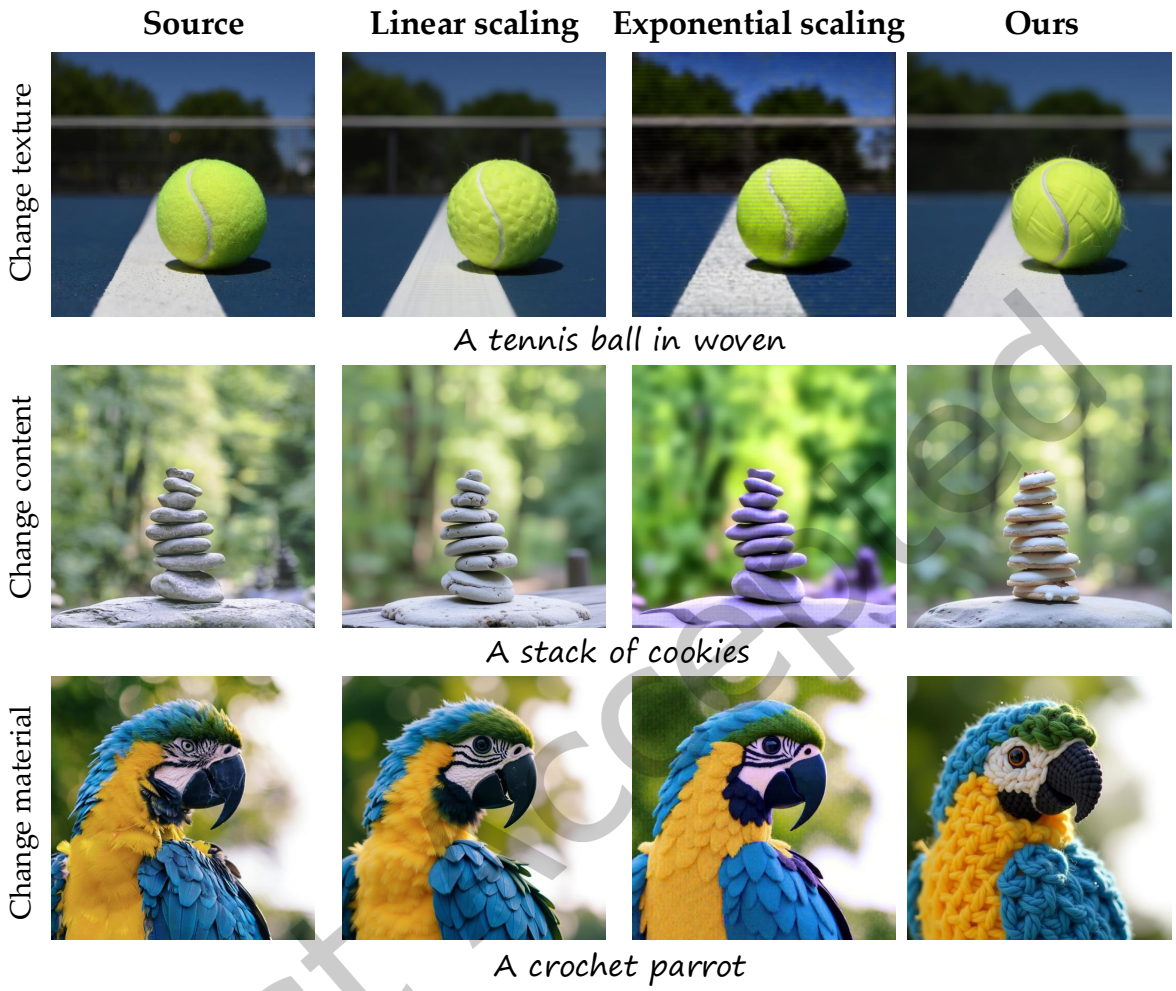

Fig. 10. **Ablation study on activation weighting formulations.** The linear formulation fails to highlight semantically relevant distinctions, while the exponential formulation tends to cause visual artifacts in the output images.

et al. 2023] and self-attention: MasaCtrl [Cao et al. 2023]. When applied to MM-DiTs, however, both methods exhibit significant limitations in terms of image fidelity and prompt-based controllability.

We present qualitative comparisons in Fig. 11, where the editing prompts aim to introduce noticeable semantic changes (e.g., adding "lightning" and "Moon", changing "stone" to "cookies", altering material to "crochet"). When using Prompt2Prompt, the editing effect is significantly weakened. This degradation can be attributed to the architectural difference between MM-DiTs and UNet-based models: Prompt2Prompt operates by manipulating cross-attention maps, which explicitly represent prompt-to-image associations through fixed text guidance in the denoising process. However, in MM-DiTs, cross-attention is no longer isolated or clearly defined. Instead, joint attention mechanisms are used, where prompt and image tokens are fused into a shared space. As the transformer layers deepen, the model progressively entangles visual and textual tokens in a latent space, leading to an implicit

Table 6. **Quantitative ablation results on gamma.**

Metrics	0	0.5	1	1.5
Structure-alignment(↑)	0.9035	0.9194	0.9198	0.9202
Prompt-alignment(↑)	0.3097	0.3203	0.3217	0.3221
Image-quality(↓)	0.2105	0.2103	0.2988	0.3279

and diluted prompt-to-image influence, as presented in Fig. 3. Consequently, edits guided by cross-attention manipulation become less effective. As illustrated in Fig. 11, this results in failures such as missing lightning, incomplete transformation of stone to cookies, and only partial material change as presented.

With MasaCtrl, which injects the key and value tokens of the self-attention from the source image into the edited image, the results retain strong priors from the source image. While this mechanism helps preserve structure, it severely restricts the extent of editing. In our experiments, edits guided by MasaCtrl appear largely unchanged compared to the source image, suggesting that the injection overpowers the intended textual modifications. Additionally, we observe slight blurriness in the generated results, likely due to a distribution shift when applying this method to joint attention framework of MM-DiTs.

These experiments demonstrate that UNet-based editing methods are not directly transferable to MM-DiTs, due to fundamental differences in attention structure. This underscores the need for editing approaches specifically designed for the joint attention and token fusion of MM-DiTs architectures. We also conduct quantitative experiments, results in Tab. 7 show that our method achieves the best alignment with source image structure, prompt consistency, and overall image quality.

Furthermore, we apply our method to the UNet-based Stable Diffusion 2 [Rombach et al. 2022] without the text token enhancement module, as text influence in SD2 is fixed via cross-attention. As shown in Fig. 11 and Tab. 7, our method achieves promising editing results while preserving the source image structure. These findings suggest that the proposed IARouter and image token enhancement methods are effective for UNet-based architecture, demonstrating strong generalization capability.

Table 7. **Quantitative comparison of applying UNet-based methods to MM-DiTs.**The results confirm the limited editing effect and structure preservation of UNet-based methods applying on MM-DiTs.

Methods	MM-DiTs+PtP	MM-DiTs+MasaCtrl	UNet+Ours	Ours
Structure-alignment (↑)	0.8441	0.8554	0.8958	**0.9194**
Prompt-alignment (↑)	0.2927	0.2932	0.3019	**0.3203**
Image-quality (↓)	0.2802	0.3257	0.2733	**0.2103**

5.7 Generalization to MM-DiTs Variants and Inversion-based Editing Methods

To further assess the robustness and adaptability of our approach, we extend our experiments to additional MM-DiTs-based architectures and inversion-based editing pipelines. Specifically, we evaluate our method on SD3.5 [Esser et al. 2024] and integrate it with two popular inversion-based editing techniques: Fireflow [Deng et al.

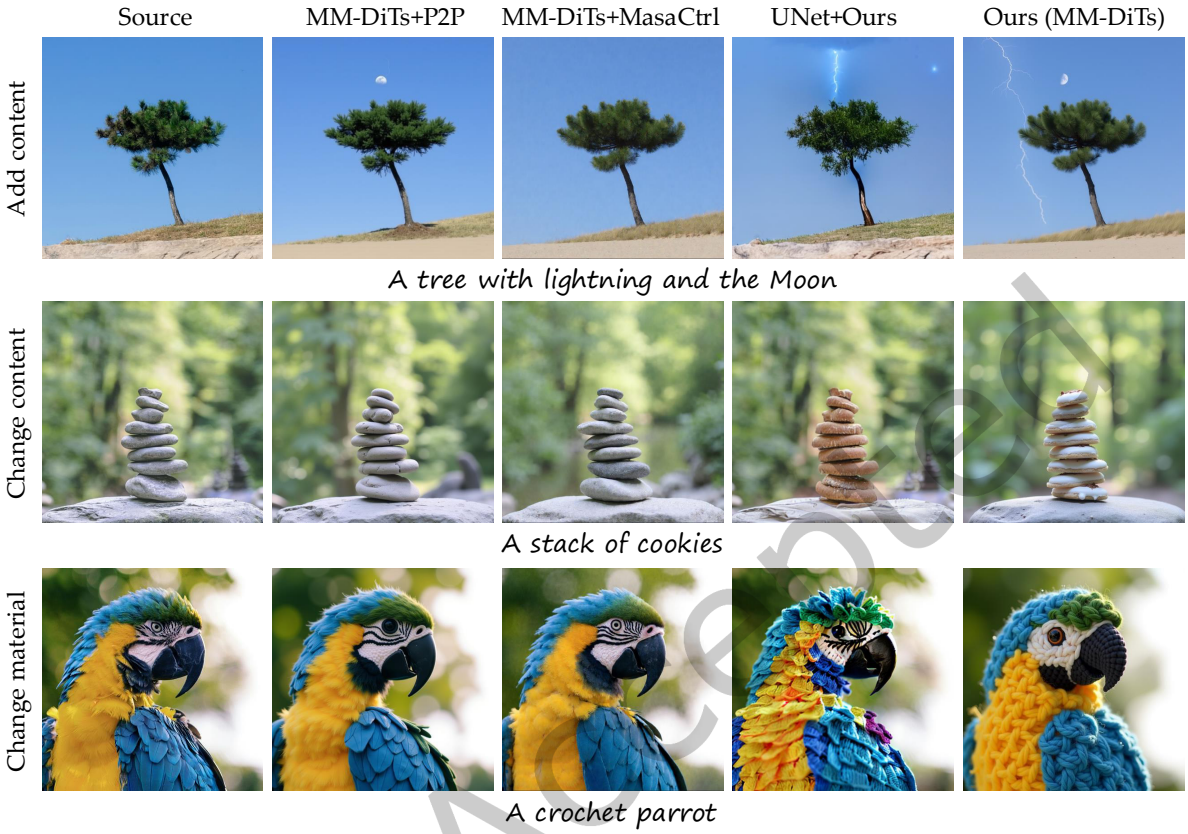

Fig. 11. **Qualitative comparison of applying UNet-based methods to MM-DiTs and our approach applying to UNet-based models.** Our method outperforms UNet-based approaches while also generalizes well to UNet-based models, demonstrating strong generalization capability.

2024] and RF-solver [Wang et al. 2024]. As shown in Tab. 8, these results demonstrate that our method generalizes well to other MM-DiTs variants and remains compatible with diverse inversion-based editing approaches, underscoring its robustness and broader applicability.

Table 8. **Quantitative results on Stable Diffusion 3.5 and adaptation to various inversion-based editing approaches.** The results confirm the effectiveness on different MM-DiTs variants and adaptability to various inversion-based editing approaches.

Methods	SD3.5+Ours	Fireflow+Ours	RF-solver+Ours	Ours (RF-inversion)
Structure-alignment (DINO) (↑)	0.9185	0.9228	0.9247	0.9194
Prompt-alignment (CLIP) (↑)	0.3175	0.3251	0.3266	0.3203
Image-quality (LPIPS) (↓)	0.2128	0.2076	0.2058	0.2103

5.8 Hyper-parameters Analysis

In the IARouter, we set $\gamma = 0.5$, which defines the domain of the sigmoid function, and $\delta = 0.5$ to ensure symmetry in the mapped range. Additionally, the editing strength and style can be modulated through α and σ in semantic-oriented image token enhancement. We present results of different α and σ in Fig. 12. We can observe that α significantly controls the strength of editing semantics, while σ has a relatively weak influence to the results. Adjusting α enables fine-grained control over image editing outcomes to accommodate diverse user preferences. For all our experiments, we use fixed values of $\alpha = 2$ and $v = 1$. While these default parameters work effectively, users can adjust them for specific samples if needed to achieve desired editing effects.

5.9 Limitation

As shown in Fig. 13, due to the multimodal text-image priors in the pre-trained models, when editing common elements like the "Eiffel Tower" with "a <description> Eiffel Tower" may yield limited results, as these prompts already encode specific visual details. Additionally, our approach requires image inversion to the latent space, so the alignment between the edited result and the original image depends on the accuracy of the inversion process.

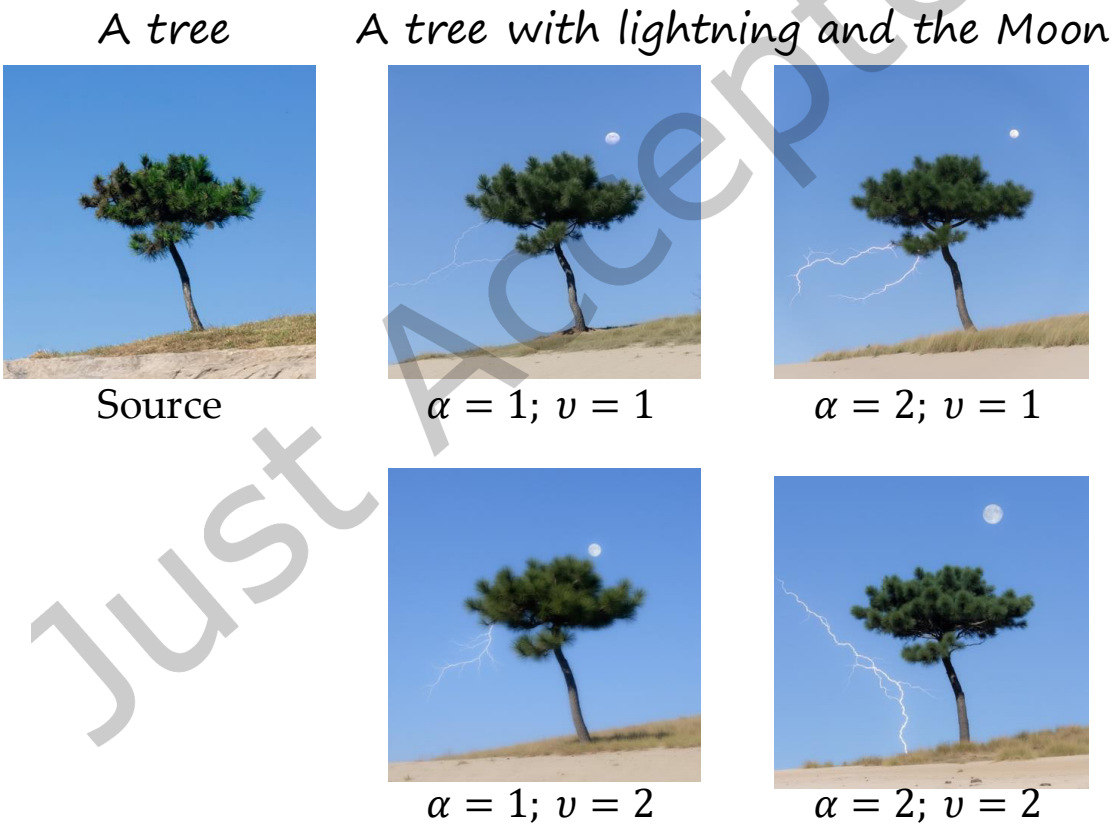

Fig. 12. **Influence of hyper parameters.** Observing that α controls the strength of editing semantics, while σ has a relatively weak influence on the results. Adjusting α enables fine-grained control over image editing outcomes to accommodate diverse user preferences.

Fig. 13. **Analysis of badcase and limitation**. Objects with complete semantics (such as “Eiffel Tower") may yield fewer effects, and the background and certain details may be lost during reconstruction.

6 Conclusion

In this work, we investigate multi-head attention within MM-DiTs for image editing, revealing the distribution of distinct image semantic information across heads. Additionally, we analyze text-to-image token guidance, observing that text influence diminishes in deeper attention blocks. Building on these insights, we introduce the instance-adaptive attention head router to enhance the representation of key attention heads for targeted editing semantics and propose the dual-token refinement module to ensure precise text guidance and emphasis on key regions. Extensive quantitative and qualitative evaluations, as well as user studies, demonstrate the superiority of our approach over existing state-of-the-art methods.

Furture work. Future exploration could focus on leveraging images as guidance. Images inherently provide richer information, spanning from coarse-grained to fine-grained details. To achieve this, adapters for DiTs can be employed to effectively capture image-based information, enabling more consistent and semantically coherent image editing.

Acknowledgments

We thank the reviewers for their insightful comments and suggestions. This work was supported in part by the National Natural Science Foundation of China under Nos. 62572458, Beijing Science and Technology Plan Project under No. Z251100008125009, National Science and Technology Council, Taiwan under Grant No. 113-2221-E006-161-MY3, and the German Research Foundation (DFG) Project No. 508324734.

References

Michael Samuel Albergo and Eric Vanden-Eijnden. 2023. Building Normalizing Flows with Stochastic Interpolants. In The Eleventh International Conference on Learning Representations.

Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. 2025. Stable flow: Vital layers for training-free image editing. In Proceedings of the Computer Vision and Pattern Recognition Conference. 7877–7888.

blackforestlabs.ai. 2024. FLUX, offering state-of-the-art performance image generation. https://blackforestlabs.ai/. Accessed: 2024-10-07.

Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. 2024. Ledits++: Limitless image editing using text-to-image models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8861–8870.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 18392–18402.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 22560–22570.

Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. In The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=eAKmQPe3m1

Yusuf Dalva, Kavana Venkatesh, and Pinar Yanardag. 2024. Fluxspace: Disentangled semantic editing in rectified flow transformers. arXiv preprint arXiv:2412.09611 (2024).

Yingying Deng, Xiangyu He, Changwang Mei, Peisong Wang, and Fan Tang. 2024. FireFlow: Fast Inversion of Rectified Flow for Image Semantic Editing. arXiv preprint arXiv:2412.07517 (2024).

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In International Conference on Learning Representations.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In Proceedings of the 41st International Conference on Machine Learning. 12606–12633.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. 2024. Interpreting CLIP's Image Representation via Text-Based Decomposition. In The Twelfth International Conference on Learning Representations.

Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2024. ReNoise: Real Image Inversion Through Iterative Noising. arXiv preprint arXiv:2403.14602 (2024).

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In The Eleventh International Conference on Learning Representations.

Xingchang Huang, Corentin Salaun, Cristina Vasconcelos, Christian Theobalt, Cengiz Oztireli, and Gurprit Singh. 2024b. Blue noise for diffusion models. In ACM SIGGRAPH 2024 Conference Papers. 1–11.

Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. 2024a. Diffusion model-based image editing: A survey. arXiv preprint arXiv:2402.17525 (2024).

Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. 2024. An edit friendly ddpm noise space: Inversion and manipulations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 12469–12478.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. OpenCLIP. doi:10.5281/zenodo.5143773

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. 2024. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In The Twelfth International Conference on Learning Representations.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6007–6017.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. 2025. Flowedit: Inversion-free text-based editing using pre-trained flow models. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 19721–19730.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow Matching for Generative Modeling. In The Eleventh International Conference on Learning Representations.

Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. 2024. Towards Understanding Cross and Self-Attention in Stable Diffusion for Text-Guided Image Editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7817–7826.

Xingchao Liu, Chengyue Gong, et al. 2023. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In The Eleventh International Conference on Learning Representations.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. 2024. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In ACM SIGGRAPH 2024 Conference Papers. 1–12.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In International Conference on Learning Representations.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6038–6047.

Hyelin Nam, Gihyun Kwon, Geon Yeong Park, and Jong Chul Ye. 2024. Contrastive denoising score for text-guided latent diffusion image editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9192–9201.

OpenAI. 2024. Sora: Creating Video from Text. https://openai.com/sora.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research (2023).

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. 2023. Zero-shot image-to-image translation. In ACM SIGGRAPH 2023 Conference Proceedings. 1–11.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 4195–4205.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In The Twelfth International Conference on Learning Representations.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning. PMLR, 8748–8763.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer, 234–241.

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. 2024. Semantic Image Inversion and Editing using Rectified Stochastic Differential Equations. (2024).

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems 35 (2022), 36479–36494.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In International Conference on Learning Representations.

Vadim Titov, Madina Khalmatova, Alexandra Ivanova, Dmitry Vetrov, and Aibek Alanov. 2024. Guide-and-Rescale: Self-Guidance Mechanism for Effective Tuning-Free Real Image Editing. arXiv preprint arXiv:2409.01322 (2024).

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1921–1930.

A Vaswani. 2017. Attention is all you need. Advances in Neural Information Processing Systems (2017).

Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. 2024. Taming Rectified Flow for Inversion and Editing. arXiv preprint arXiv:2411.04746 (2024).

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. 2024. Inversion-Free Image Editing with Language-Guided Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9452–9461.

Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. 2024. Rgbx: Image decomposition and synthesis using material-and lighting-aware diffusion models. In ACM SIGGRAPH 2024 Conference Papers. 1–11.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition. 586–595.

Received 13 May 2025; revised 8 December 2025; accepted 26 January 2026