IET Computer Vision

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Panoptic Scene Graph Grounded Training-Free Image Editing With Mutually Exclusive Attention Manipulation

Yunqing He | Ruichao Hou | Jia Bei | Tongwei Ren

State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

Correspondence: Ruichao Hou (rchou@nju.edu.cn)

Received: 22 August 2025 | **Revised:** 13 April 2026 | **Accepted:** 30 April 2026

Keywords: conditional generation | image editing | image generation | image processing | image reconstruction

ABSTRACT

Prevailing image editing methods heavily rely on user-provided bounding boxes or pixel-level masks to ensure visual consistency in nonedited regions. Although some attention-based approaches eliminate the need for manually annotated input, they often unexpectedly alter nontarget areas due to semantic leakage between objects. Our goal is to address this semantic inconsistency challenge with minimal user input by leveraging the mutual exclusion of scene graph nodes, thereby enhancing both editability and background preservation without additional training costs. To address the challenge of semantic inconsistency, we propose a Scene graph-based ImaGe editing method with Mutually exclusive Attention manipulation, namely SIGMA, to leverage the inherent semantic mutual exclusion properties between scene graph nodes for attention map distribution manipulation. Specifically, we propose a semantic decoupling module to disentangle the desired and nontarget editing areas. We also introduce a semantic injection module to facilitate both foreground editing and background preservation. We validated the effectiveness of SIGMA on the widely used image editing dataset PIE-Bench. The experimental results demonstrate that SIGMA significantly outperforms existing approaches without any additional training cost.

1 | Introduction

Recent advancements in text-based image editing have witnessed significant progress with the emergence of diffusion models [1–3]. Nonetheless, precise control over semantic concepts in composite images remains a challenge, especially when object-level editing is required while preserving the overall coherence of the scene [4, 5]. Scene graphs, which serve as structured representations of objects and their relationships, present a promising solution to bridge this gap in controllable image editing [6–8]. Compared to text-based image editing strategies, scene graph offers a higher-level and finer-grained approach to image manipulation, enabling object-level semantic control [9].

Existing scene graph-based image editing methods enhance controllability based on various user inputs. One explicit control strategy involves users defining the editing area by providing bounding boxes or pixel-level masks. Although this method allows for precise object positioning, it increases user interaction costs. Conversely, an implicit approach requires no additional input, instead leveraging the attention maps to identify different semantic concepts. Attention-based methods reduce interaction costs and improve overall image quality but are more prone to causing semantic inconsistencies in nonedited areas. As shown in Figure 1, when an original image containing both a ‘cat’ and a ‘dog’ is edited to replace the ‘cat’ with a ‘rabbit’, existing methods significantly alter the appearance of the ‘dog’ and the background. The semantic leakage between objects primarily arises from the semantic coupling that occurs during image generation. In other words, the semantics of objects not only affect their visual appearance but also influence the overall visual presentation of the entire image. Semantic coupling, which

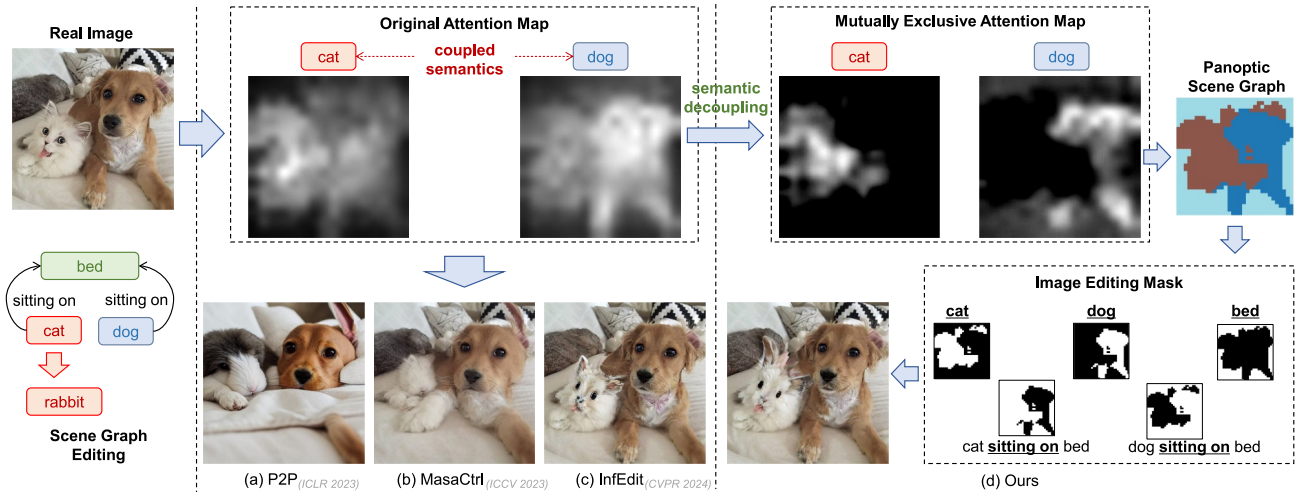

FIGURE 1 | Motivation of our method. We leverage the mutual exclusion between scene graph nodes to decouple object semantics for more accurate image editing.

is beneficial for maintaining global coherence in image generation, becomes a critical barrier in editing tasks. Because the semantics of each object are intertwined, modifying one object inevitably disturbs others. Existing attention-based methods fail to decouple these semantic associations, whereas scene graph-based methods lack efficient ways to utilise semantic structure without extra training. In practical scenarios such as product image editing, portrait retouching and scene reconstruction, users often require precise object-level edits (e.g., replacing a product while preserving the background or modifying facial expressions without altering hairstyle) with minimal interaction. Current methods either demand complex manual annotations or produce inconsistent results, failing to meet these demands.

In this paper, we propose a Scene graph-based ImaGe editing method with Mutually exclusive Attention manipulation, SIGMA, to facilitate more precise image editing. The motivation for SIGMA is to deal with the trade-off between minimal user intervention and semantic consistency in nonedited regions. The primary advantage of our proposed method is its training-free nature, which eliminates the need for additional training or fine-tuning. This approach ensures broad applicability and efficiency without sacrificing performance. The core concept of SIGMA is to differentiate between edited and nonedited areas through scene graph nodes. Given an edited scene graph, we generate a pixel-level panoptic scene graph as an image editing mask for mutually exclusive semantic localisation. Specifically, the SIGMA method is divided into two stages. Initially, during the semantic decoupling stage, SIGMA decouples the original attention map into foreground and background attention maps. We define that the foreground attention maps of a specific object primarily define its appearance, whereas the background attention maps more significantly influence its co-relationships with other objects. We propose a mutually exclusive attention manipulation algorithm to decouple the attention maps between objects in the spatial–temporal dimension. Following semantic decoupling, in the semantic injection stage, SIGMA recombines the foreground and background attention maps with various semantics to produce the edited image. A significant strength of SIGMA is its mutually exclusive attention manipulation strategy, which effectively separates foreground and background semantics. This separation allows for precise target editing while maintaining the consistency of nonedited regions, a crucial improvement over existing attention-based methods that often suffer from semantic leakage. Furthermore, SIGMA combines scene graph structural guidance with attention manipulation, achieving more refined object-level control compared to text-based editing methods. It also demonstrates superior efficiency, capable of editing a 512 × 512 image in under 1.5 s on a single RTX 4090 GPU, making it well-suited for real-world applications. We validate the effectiveness of the proposed SIGMA method on the widely used PIE-Bench image editing benchmark. The experimental results demonstrate that the SIGMA method performs both improved editability in target regions and improved consistency in nonedited areas.

Our main contributions are summarised as follows:

- We propose a training-free scene graph-based image editing method with mutually exclusive attention manipulation, which enables semantic decoupling between different objects in the image generation process, thereby achieving high semantic consistency in image editing.
- We propose a semantic decoupling module based on spatial–temporal attention mutual exclusion, which decouples the attention maps of different objects into foreground and background attentions; we also propose a semantic injection module based on decoupled attention maps recombination, which combines semantic information of object nodes in the scene graph into the visual representation to achieve consistent image editing.

2 | Related Works

2.1 | Training-Free Image Editing

From the perspective of the model parameters modification, image editing methods can be classified into training-based and training-free approaches [10–12]. Training-based methods are further divided into two subtypes: those trained on specific datasets for general editing tasks and those fine-tuned on

individual target images for personalised adaptation [1–3, 13]. In contrast, training-free image editing techniques significantly reduce time costs and offer broader applicability [14, 15]. These training-free approaches can be implemented through various paradigms. Some researchers focus on the inversion process of image editing [16, 17]. DDIM inversion provides the ability for diffusion-based image generation models of the unconditional natural image editing task [18, 19]. Direct inversion addresses the challenge of the conditional image editing with text-guided diffusion models [20]. Additionally, some utilise masks to refine the inference-time sampling process, ensuring that localised modifications align precisely with specific regions of interest [21–23]. Some recent works introduce multi-noise redirection, which predicts multiple noises in a single sampling step and subsequently redirects them into a single noise [24]. Some researchers have observed that manipulating attention maps can significantly influence image generation results and have applied this in image editing tasks [5, 25, 26]. P2P introduces an intuitive image editing framework by manipulating the attention map distribution in cross-attention layers [4]. MasaCtrl further extends the editability of attention maps to self-attention layers [27].

Our approach follows the attention map manipulation paradigm for training-free image editing. We integrate both cross-attention and self-attention manipulation operations to enhance the comprehensiveness of our image editing capabilities. Within the cross-attention layer, we also introduce a semantic decoupling process to extend semantic alignment from the image level to the object level.

2.2 | Scene Graph-Based Image Editing

In recent years, the field of scene graph-based image editing has seen significant advancements. The concept of scene graphs is defined as a combination of objects and their inter-relationships in a scene [6]. It offers a structural representation that bridges the gap between visual and semantic content, demonstrating advantages in creating realistic and semantically coherent images [7, 11, 12, 28]. SIMSG is the first attempt to employ scene graphs for image editing, using a layout predictor to estimate bounding box regions for each node in the scene graph, thereby enabling targeted object editing through a GAN-based network [8]. Building on this, Su et al. assume that bounding boxes provide insufficient precision and replace the layout predictor with a pixel-level mask predictor for more refined image editing [29]. With the rapid advancement of diffusion models, Zhang et al. present the first diffusion-based image editing approach with scene graph manipulation [9]. They redefine the editing task as an inpainting problem, where target regions are masked and regenerated using a re-trained diffusion model. SGEdit advances this paradigm by incorporating a large language model alongside a re-trained diffusion model, thereby extending scene graph-based image editing research from a close-domain focus to open-vocabulary manipulation [30].

Compared to existing scene graph-based image editing approaches, our method offers two significant advantages. Firstly, although our method also predicts mask regions for each scene graph node, it fundamentally differs by utilising the intrinsic image generation capabilities of diffusion models through an image reconstruction branch, thereby eliminating the need for training a specific layout predictor. Secondly, instead of retraining the diffusion model, we employ attention map manipulation to directly generate edited images, resulting in superior efficiency and enhanced editability while maintaining the structural guidance of scene graphs.

3 | Methodology

3.1 | Problem Formulation

Preliminaries: The foundation for attention manipulation stems from a previous discovery that the attention map can significantly influence the layout of a synthesised image [4]. The attention mechanism in the diffusion model can be expressed as follows:

$$Attn(Q,K,V) = \underbrace{A}_{\text{relevant to layout}} \cdot \overbrace{V}^{\text{relevant to semantics}} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{1}$$

where Q is the query features projected from the visual representations. K and V are the key and value features projected from the visual representations (in self-attention layers) or semantic embeddings (in cross-attention layers). $\mathbf{A}$ is the attention map, which is the target of attention manipulation. The process of attention manipulation in image editing can be formulated as follows:

$$x_0 = A_{src} \cdot V_{src} \rightarrow A_{src}^{unedit} \cdot V_{src} + A_{src}^{edit} \cdot V_{tgt} = x_0', \tag{2}$$

where x_0 is the original image, and x_0' is the edited image. The edited region is created by aggregating the target semantics V_{tgt} with the original layout A_{src}^{edit}, whereas the unedited region is intended to retain its original appearance through $A_{src}^{unedit} \cdot V_{src}$.

Task Definition: The fundamental task of our proposed approach is scene graph-based image editing, wherein the input is a natural image x_0, an original scene graph G_{src} and an editing operation E on G_{src}, and the output is an edited image x_0', as shown in Equation (3).

$$G_{src}, E, x_0 \rightarrow x_0', \tag{3}$$

where G_{src} can be described as a directional graph $G_{src} = (O, R^{\Omega})$, wherein the nodes are objects O in an image, and edges are their relationships R. Ω indicates that the set of relationships can be empty. For editing operation E, as shown in Table 1, we follow previous research to define the image editing operation set we supported [10].

Specifically, our approach further decomposes the image editing task described in Equation (3) into several subtasks, as illustrated in Equation (4).

$$\begin{aligned} G_{src}, E &\rightarrow G_{tgt}, \\ G_{src}, G_{tgt}, x_0 &\rightarrow PSG_{mask}, \\ PSG_{mask}, G_{src}, G_{tgt}, x_0 &\rightarrow x_0', \end{aligned} \tag{4}$$

where G_{tgt} indicates the target scene graph, and PSG_{mask} represents for the pixel-level panoptic scene graph on x_0 generated by the union of G_{src} and G_{tgt}.

TABLE 1 | Editing operations supported by our method.

Semantic editing	Stylistic editing	Structural editing
Object replacement	Colour	Object addition
Background change	Texture	Object removal
Expression modification	Style	Action/pose change

Robustness analysis for pre-processing errors: In a fully automated end-to-end pipeline, the main errors in text-to-scene graph parsing progress can be categorised into two types: concept omission and concept merging, and SIGMA shows good robustness to both types. Concept omission means the generated scene graph misses some objects, relationships or attributes. If the omitted concept belongs to a nonedited region, the concept that does not appear in the scene graph will be treated as background and thus will not be modified; therefore, SIGMA still performs well. If the omitted concept belongs to the target editing region, this error can be easily detected before the editing process; therefore, the editing pipeline will not proceed to the image generation step. Concept merging means multiple objects, or an object and its attribute, are incorrectly merged into a single node in the scene graph. If the merged node is in a nonedited region, it will still be preserved as background; therefore, the robustness of SIGMA is not affected. If the merged node is in the editing region, it may cause the editing region to be larger than expected, which may affect the appearance of the surrounding nonedited regions. However, because T-to-SG errors only affect object localisation but do not change the semantic understanding, SIGMA still maintains decent robustness in this case. Therefore, SIGMA is able to handle most common parsing errors in the upstream Text-to-SG step with good robustness.

Signals Definition: Finally, we provide definitions and explanations for the main symbols used in the main content to enhance the readability, as presented in Table 2.

3.2 | Overview

Our primary motivation for this work is to tackle semantic inconsistency issues in training-free, low-interaction image editing. As diffusion models grow more powerful, the need for precise control over semantic concepts is also increasing. In this context, we propose utilising the mutual exclusion of scene graph nodes to integrate attention graphs into foreground and background components. By transferring the mutual exclusion characteristics between objects from the visual level to the attention level, we can accurately localise the edited and unedited regions in the iterative image generation process.

As illustrated in Figure 2, we present the framework of our training-free image editing method. As depicted on the top–left, the general image generation process integrates semantics and visual representations for conditional image creation. The image editing process follows a similar paradigm to aggregate the semantics of target objects into the layout of the source visual representations. The source visual representations are obtained through an inversion process. The target visual representations are produced by merging the source visual representations with the target semantics in the regeneration process. The editing operation is conducted in both self-attention and cross-attention layers. In self-attention layers, inspired by MasaCtrl [27], we introduce a mutual self-attention control technique to query visual contents from the source image. The editing operation in self-attention layers can be formulated as follows:

$$Attn(Q_{src}, K_{src}, V_{src}) \rightarrow Attn\big(Q_{tgt}, K_{src}, V_{src}\big), \tag{5}$$

which means the edited image is querying visual content from source semantics V_{src} for controllable image editing. In cross-attention layers, we propose a semantic decoupling module and a semantic injection module to disentangle the foreground and background attentions. The editing operation in cross-attention layers can be formulated as follows:

$$Attn(Q_{src}, K_{src}, V_{src}) \rightarrow Attn\big(Q_{src}, K_{src}, V_{tgt}\big). \tag{6}$$

The details about the semantic decoupling module and semantic injection module are presented in Figure 3. Given two objects o_1 and o_2, their original attention maps at time step t_e are denoted by $A^{o1}_{t_e}$ and $A^{o2}_{t_e}$. Initially, ST-MEAN employs spatial attention manipulation M_s to obtain mutually exclusive attention maps. These attention maps provided by M_s are then used to maintain the attention map cache for the temporal attention manipulation M_t, as well as to generate foreground regions through pixel-level attention manipulation M_p.

3.3 | Semantic Decoupling Module

In the semantic decoupling module, we introduce a Spatial–Temporal Mutually Exclusive Attention maNipulation algorithm (ST-MEAN) to effectively disentangle the semantics of different nodes within scene graphs. ST-MEAN separates object attentions into two components: foreground attentions and background attentions. Foreground attention physically represents the appearance of individual objects and requires adjustment when the target object is being edited. In contrast, background attention primarily captures the contextual influence from surrounding objects and should remain unchanged when other objects are unmodified, ensuring scene consistency. As shown in Figure 4, ST-MEAN operates through three dimensions: pixel-level exclusion constraints between attention maps, spatial mutually exclusive attention manipulation, and temporal mutually exclusive attention enhancement.

Pixel-level attention manipulation: The pixel-level attention manipulation M_p serves as an atom operation for pixel classification to construct panoptic scene graphs. Specifically, given an attention matrix $\mathbf{A}$ with dimensions (N, N, D) that contains D object nodes, M_p performs sparsification by retaining only the maximum attention value at each pixel location while suppressing all others. This process can be formulated as follows:

$$\mathbf{A}_p(i,j,k) = \begin{cases} 1, & \text{if } \mathbf{A}(i,j,k) = \max\limits_{1 \le m \le D} \mathbf{A}_{i,j,m} \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where the maximum attention position at each pixel is set to 1 and others to 0. The attention mask $\mathbf{A_p}$ is subjected to a Hadamard product with the original attention map to determine

TABLE 2 | Definitions and explanations of main symbols.

Symbol	Explanation
Basic variables and image-related	
x_0	Original input image
x_0'	Edited output image
G_{src}	Scene graph for original input image, formulated as a directed graph (O, R^{Ω})
G_{tgt}	Scene graph for target editing image
PSG_{mask}	Pixel-level panoptic scene graph mask (for edited region localization)
O	Set of object nodes in the scene graph
R^{Ω}	Set of object relationships in the scene graph
E	Image editing operation (see Table 1)
d	Dimension of attention features
N	Dimension related to image size
D	Total number of object nodes in the scene graph
Attention mechanism-related	
$Attn(Q, K, V)$	Attention calculation function
Q	Query feature in the attention mechanism
K	Key feature in the attention mechanism
V	Value feature in the attention mechanism
A	Attention map calculated by Q and K
$A(i, j, k)$	Attention map value of k-th object at pixel location (i, j)
A_{src}	Attention map corresponding to the original image
A_{src}^{edit}	Part of the attention map corresponding to the edited region
A_{src}^{unedit}	Part of the attention map corresponding to the unedited region
A_{fg}^{k}	Foreground attention map of the k-th object (representing the object's own appearance)
A_{bg}^{k}	Background attention map of the k-th object (representing contextual appearance)
A_p	Pixel-level attention mask (obtained via M_p operation)
A_s	Spatial mutually exclusive attention map (obtained via M_s operation)
A_s^{norm}	Normalised result of the spatial attention map
A_t	Accumulated attention map cache in the temporal dimension
A_t'	Updated temporal attention map via EMA
A_{union}^{k}	Union attention map of the k-th object (for foreground enhancement)
$A_{bg}^{\sim k}$	Background attention maps of all unedited objects except the k-th object
$A_{fg}^{\sim k}$	Foreground attention maps of all unedited objects except the k-th object
Attention operation-related	
M_p	Pixel-level attention operation (for attention map sparsification and mask generation)
M_s	Spatial mutually exclusive attention operation (resolving semantic ambiguity)
M_t	Temporal mutually exclusive attention operation (attention accumulation based on EMA)
α	Decay coefficient for EMA update (set to 0.9999 in the paper)
$p_{i,j}^{k}$	Payoff value of the k-th object at pixel location (i, j)
$p_{i,j}^{\prime k}$	Context-aware payoff value of the k-th object at pixel location (i, j)

(Continues)

TABLE 2 | (Continued)

Symbol	Explanation
$p_{i,j}^{\sim k}$	Sum of the payoff value at pixel location (i,j) of all objects except the k-th object
$\mathcal{N}$	Set of neighbouring pixels around location (i,j)

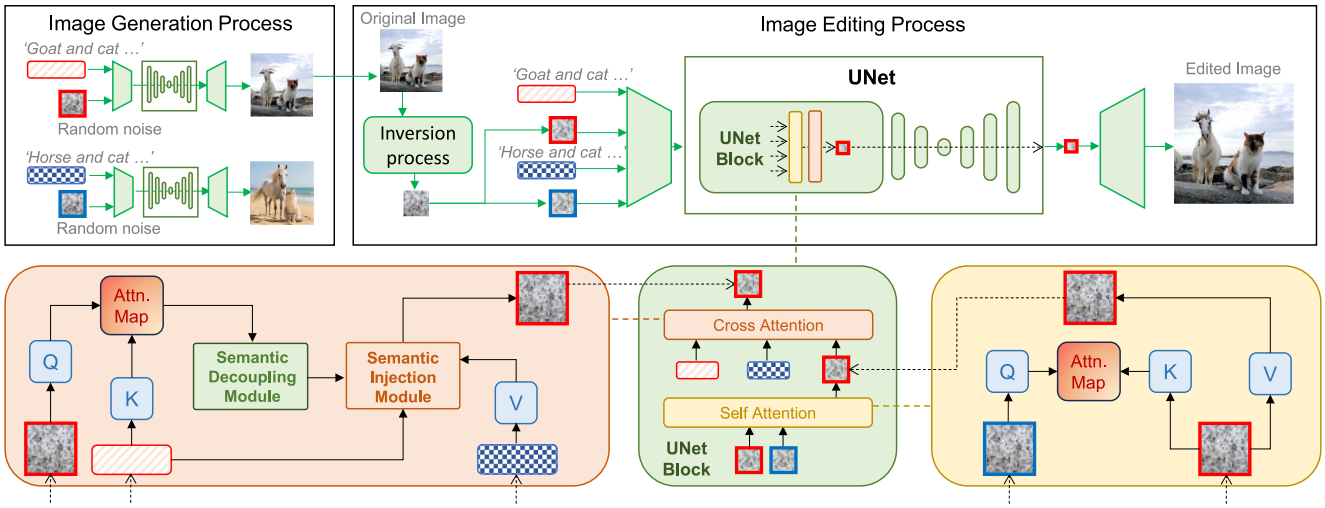

FIGURE 2 | The framework of our training-free image editing method. Real image or generated image is inversed to latent vector as the input of image editing. The image editing process is like a regeneration progress, with the target semantics are mixed into the progress. In each block in UNet, the information from source semantics, target semantics and target visual representations are mixed into the source visual representations, and finally the updated source visual representations are used for edited image generation by a VAE decoder.

the region to be edited. It is important to note that the computation of cross-attention remains approximately independent across all cross-attention layers, as the derivation of attention relies solely on the Q and K values. This implies that the hard masks generated by M_p in preceding layers minimally interfere with the intrinsic attention calculation of subsequent layers. This characteristic ensures that while M_p modifies the attention distribution through a rough strategy, it does not alter the fundamental attention computation that governs image generation, thereby preserving the model's capacity for controllable editing while enforcing structural preservation.

Spatial attention manipulation: The challenge of manipulating spatial attention lies in addressing semantic ambiguity when multiple objects exhibit similar attention values at the same spatial coordinates. We conceptualise the spatial attention manipulation algorithm M_s as a complete-information cooperative game among multiple players to tackle this semantic ambiguity. Within this game-theoretic framework, we begin the manipulation process by considering the original attention matrix $\mathbf{A}$ as representing individual pixel-level payoffs for each of the N object nodes in the scene graph. Our primary goal is to construct payoff matrices that meet Nash equilibrium conditions in a single-round game, governed by two fundamental constraints: (1) Players naturally prioritise pixels with higher attention values due to their inherent utility maximisation tendency; (2) Spatial coherence is enforced through cooperative rewards when adjacent pixels are jointly occupied, reflecting the perceptual continuity of visual perception. To mitigate the potential of an object occupying too many pixels (i.e., the 'tragedy of the commons' scenario), we implement a pre-game normalisation procedure that equitably distributes stakes among all players, as depicted in Equation (8).

$$\mathbf{A}_s^{\text{norm}}(i,j,k) = N^2 \cdot \frac{\mathbf{A}(i,j,k)}{\sum_{j=0}^{N}\sum_{i=0}^{N}|\mathbf{A}(i,j,k)|}, \tag{8}$$

where $\mathbf{A} \in R^{N,N,D}$ represents the attention metric, and $\mathbf{A}_s^{\text{norm}}$ indicates that each object node exhibits similar competitiveness.

Furthermore, to define the two fundamental constraints, we design a payoff calculation process to get the payoff matrix. We define the payoff value $p_{i,j}^k$ of object k in location (i,j) as its attention value $A(i,j,k)$, then its contextual payoff can be formulated as follows:

$$p'^{k}_{i,j} = p_{i,j}^k + \sum_{j=0}^{\mathcal{N}}\sum_{i=0}^{\mathcal{N}} p_{i,j}^k - \left(p_{i,j}^{\sim k} + \sum_{j=0}^{\mathcal{N}}\sum_{i=0}^{\mathcal{N}} p_{i,j}^{\sim k}\right), \tag{9}$$

where $\mathcal{N}$ denotes all the neighbouring pixels around (i,j), and $p_{i,j}^{\sim k}$ represents the payoff of all objects in location (i,j) except for object k. The direct computation of the payoff matrix has a time complexity of $O(N^{2\times D})$. However, we show that by employing a sliding window strategy for neighbourhood selection, this computation simplifies to a basic convolution operation. By leveraging a fixed 3×3 convolution kernel filled with ones, the process achieves an acceptable time complexity.

Temporal attention manipulation: The proposed temporal attention manipulation algorithm M_t addresses the observed

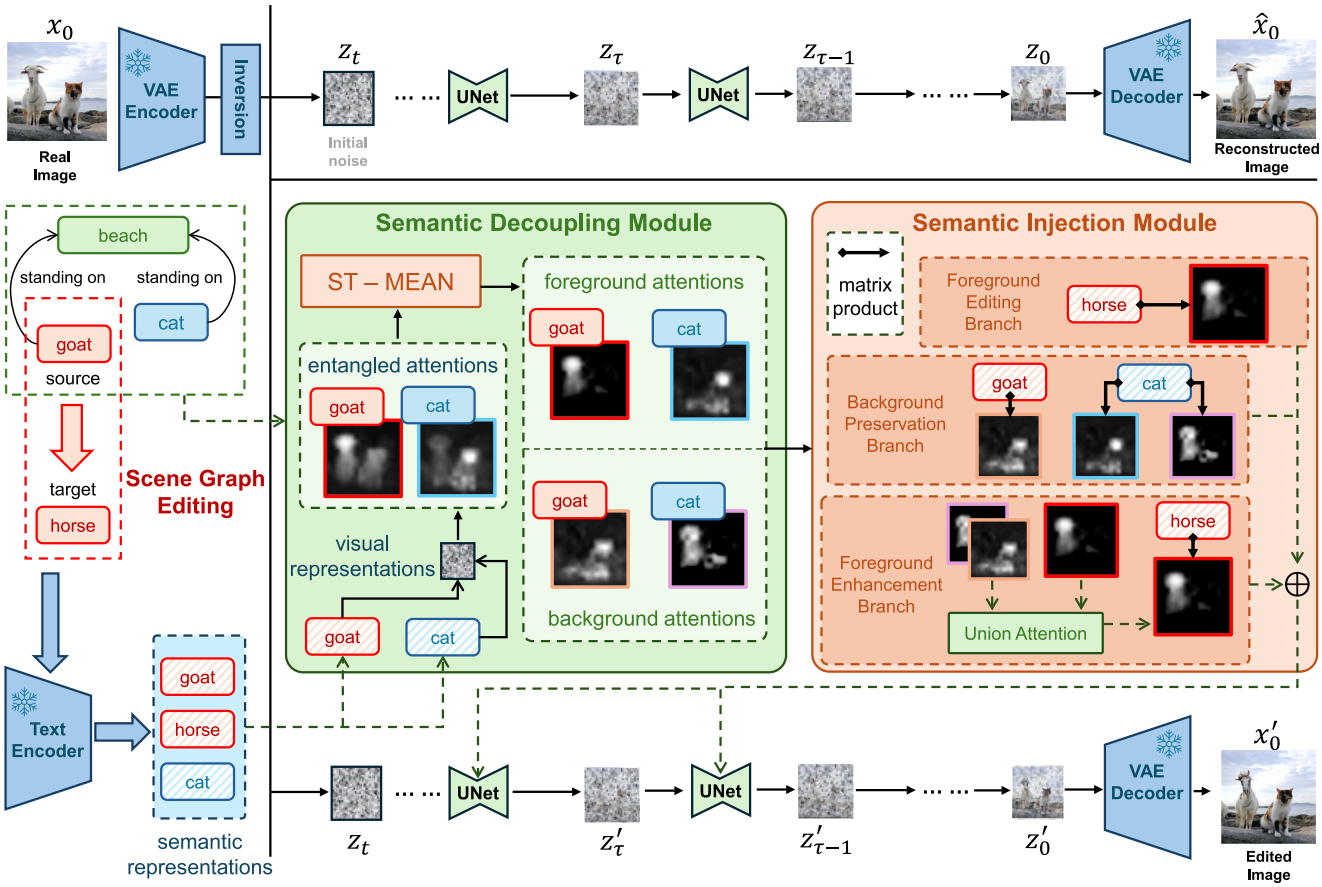

FIGURE 3 | The overview of our method in the cross-attention layer. The left depicts the inputs. The upper section indicates the overall image reconstruction branch, which is the same as the image editing branch in the bottom, except for the intervention of target semantics. The green box in the centre represents the semantic decoupling module, wherein we propose an ST-MEAN block to disentangle original attention maps into foreground attentions and background attentions. The red box illustrates how the semantic injection module recombines foreground and background attentions for precise semantic editing and preservation.

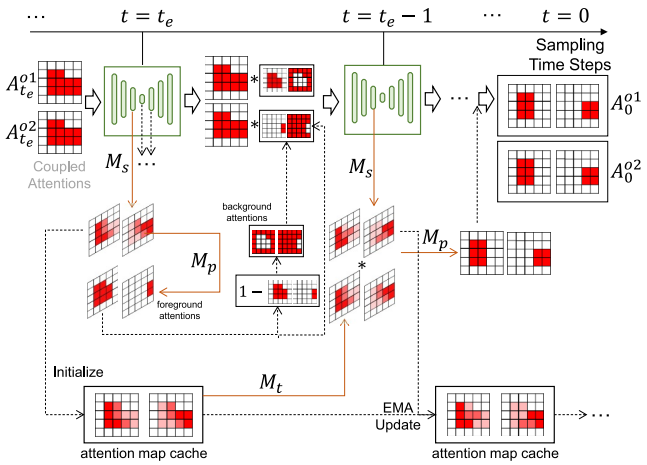

FIGURE 4 | The calculation process of ST-MEAN. ST-MEAN extracts the foreground region of each object, and the integration of these regions forms the panoptic scene graph within the current cross-attention layer, which serves as a mask to identify the image area that requires editing.

variance in attention map contributions to object appearance across different cross-attention layers and timesteps. Recognising that static spatial attention $\mathbf{A}_s$ derived solely from the M_s operation fails to capture temporal coherence, we introduce a dynamic exponential moving average algorithm (EMA) to progressively aggregate temporal attention maps. Specifically, the EMA with a decay rate of $\alpha = 0.9999$ continuously updates the attention map cache as follows:

$$\mathbf{A}_t' = \alpha \mathbf{A}_t + (1-\alpha)\mathbf{A}_s, \tag{10}$$

where $\mathbf{A}_t$ represents the raw temporal attention stored from preceding steps, while $\mathbf{A}_t'$ denotes the updated attention map cache. The matrices $\mathbf{A}_t$ and $\mathbf{A}_s$ are fused using a Hadamard product for the subsequent M_p operation. Before this multiplicative fusion, both $\mathbf{A}_s$ and $\mathbf{A}_t$ undergo a normalisation process, as defined in Equation (8), to ensure numerical stability. The final fused attention distribution effectively combines instantaneous spatial saliency with accumulated temporal evidence, demonstrating superior robustness.

3.4 | Semantic Injection Module

The semantic injection module consists of two essential components designed for precise visual content modification with controlled contextual preservation. The first component recombines decoupled attention maps to dynamically differentiate between modifiable and invariant regions. The second

component merges these attention maps with high-level semantic information, facilitating the targeted injection of semantic attributes while preserving the consistency of the visual context in unedited regions.

The foreground attention mask $\mathbf{A}_p$ is derived following the M_p operation within the semantic decoupling module. For a particular object k, its attention map $\mathbf{A}^k$ is split into two components as follows:

$$\mathbf{A}^k = \mathbf{A}^k \cdot \mathbf{A}_p^k + \mathbf{A}^k \cdot \left(1 - \mathbf{A}_p^k\right), \tag{11}$$

where $\mathbf{A}^k \cdot \mathbf{A}_p^k$ is identified as the foreground attention $\mathbf{A}_{fg}^k$ of object k, and $\mathbf{A}^k \cdot \left(1 - \mathbf{A}_p^k\right)$ as its background attention $\mathbf{A}_{bg}^k$. Importantly, because the sum of foreground and background attentions equals the original attention, our attention decoupling operation preserves robustness without altering the original attention computation process, adhering to the general attention calculation process as follows:

$$\mathbf{A}_{fg}^k \cdot V + \mathbf{A}_{bg}^k \cdot V = A^k \cdot V. \tag{12}$$

In the context of image editing, existing methods typically inject target semantics directly into the entire attention matrix as follows:

$$\mathbf{A}^k \cdot V_{src}^k \rightarrow \mathbf{A}^k \cdot V_{tgt}^k, \tag{13}$$

where k is the sole object node requiring editing in the scene. We note that, as illustrated in Figure 1, the attention matrix of the target object can capture not only its visual representation but also its inter-relationships with other objects. Therefore, we redefine the semantic injection process as follows:

$$\mathbf{A}_{fg}^k \cdot V_{src} + \mathbf{A}_{bg}^k \cdot V_{src} \rightarrow \mathbf{A}_{fg}^k \cdot V_{tgt} + \mathbf{A}_{bg}^k \cdot V_{src}. \tag{14}$$

Here, only the foreground region of the target object is edited, and the background attentions are used for contextual preservation. Additionally, recognising that the attention maps of other objects may include relevant components related to the target editing object, we further introduce a foreground enhancement component that identifies these correlated parts through a straightforward mutual cross-attention computation as follows:

$$\mathbf{A}_{union}^k = \text{Softmax}\left(\frac{\mathbf{A}_{fg}^k \cdot \left(\mathbf{A}_{bg}^{\sim k}\right)^T}{\sqrt{d}}\right)\mathbf{A}_{bg}^{\sim k}, \tag{15}$$

where $\mathbf{A}_{bg}^{\sim k}$ represents all the unedited objects' background attentions. Finally, the complete semantic injection module can thus be expressed as follows:

$$x_0' = \left(\mathbf{A}_{fg}^k + A_{union}^k\right) \cdot V_{tgt} + \left(\mathbf{A}_{bg}^k + \mathbf{A}_{fg}^{\sim k} + \mathbf{A}_{bg}^{\sim k}\right) \cdot V_{src}, \tag{16}$$

where $\mathbf{A}_{bg}^{\sim k}$ denotes all the unedited objects' foreground attentions, and x_0' is the edited image.

4 | Experiments

4.1 | Dataset and Evaluation Metrics

We primarily employ the widely-used image editing benchmark, PIE-Bench [20], to evaluate our proposed method. Additionally, we incorporate the FlowEdit dataset [31] to verify its generalisability. PIE-Bench consists of 700 images across four distinct scenarios, featuring 10 unique editing types. The evaluation primarily focuses on assessing translation quality and translation consistency. Therefore, we employ six distinct metrics as follows:

- Structure distance [32] measures the structural similarity between images, commonly used to assess structural differences between two images.
- Peak signal-to-noise ratio (PSNR) assesses image quality by determining the mean square error between the edited and original images, with higher values signifying less distortion.
- Structural similarity index measure (SSIM) [33] comprehensively evaluates the similarity between the original image and edited image from the perspective of brightness, contrast, and structure.
- Learnt perceptual image patch similarity (LPIPS) [34] uses deep learning features to simulate human visual perception differences.
- Mean squared error (MSE) directly calculates the average squared difference in pixel values between two images.
- CLIP score [35, 36] introduces a visual-language aligned model to evaluate editing performance. We employ CLIP_{Whole} to assess the entire image's similarity to target semantics, whereas CLIP_{Edited} is used to evaluate only the edited region's alignment with target semantics.

Finally, to offer a comprehensive evaluation, we also introduce a $score_{avg}$ matrix. The $score_{avg}$ is calculated using the harmonic mean of seven distinct metrics. For indicators where lower numerical values indicate better performance, that is, structure distance, LPIPS and MSE, their reciprocals are added in the numerator during the harmonic mean calculation.

4.2 | Implementation Details

We apply the proposed method to the latent consistency model using publicly available checkpoints [43]. This model offers performance comparable to stable diffusion v1.4 but operates significantly faster [44]. During the image inversion process, we utilise the DDCM algorithm to obtain the initial random noise for image reconstruction and editing [37]. Given that we achieve Nash equilibrium directly in one round of computation at each diffusion sampling step, we do not need extra iterations within a single sampling step. This does not introduce significant additional computational complexity beyond the original diffusion sampling process, as depicted in Table 3; therefore, it can preserve the ultra-fast inference advantage of LCM. With 12 denoising steps, we can edit a single image with a resolution of 512 × 512 in less than 1.5 s on a single RTX 4090 GPU.

4.3 | Comparison With the State-of-the-Arts

We primarily compare the proposed method with existing attention manipulation-based image editing methods, such as

TABLE 3 | Comparison of inference efficiency on a single A40 GPU.

Method	Time (sec/per image)
StyleDiffusion Wang et al. [1]	382.98
Imagic Kawar et al. [45]	349.98
P2P-Zero Parmar et al. [5]	56.78
EDICT Wallace et al. [17]	35.48
DiffEdit Couairon et al. [23]	27.65
P2P Hertz et al. [4]	2.60
InfEdit Xu et al. [37]	2.22
Ours	2.70

P2P [4], P2P-Zero [5], PnP [25], MasaCtrl [27], InfEdit [37], Stable Flow [38] and CoLan [39]. Additionally, we consider other scene graph-based editing methods, including SIMSG [8] and SGEdit [30]. We also include a latest retraining-based image editing method SwiftEdit [40]. The comparison results are shown in Figure 3.

Our method demonstrates significant advantages over existing scene graph-based image editing approaches, primarily due to our ability to provide more precise target localisation. In current image editing methods, when users do not specify editing regions, there are mainly two strategies for locating target objects: one employs inversion technology to identify each object's attention map in a nonrigid during image reconstruction, whereas the other introduces external locators to differentiate objects rigidly, such as separately trained layout predictors or segmentation models such as SAM. Given that research on scene graph-based image inversion and generation is still underdeveloped, existing scene graph-based editing methods adopt the second strategy for image-level object localisation. In contrast, we utilise the first strategy by leveraging text-to-scene graph technology to convert text-encoded attention maps into scene graph attention maps, thereby fully exploiting the advantages of existing inversion techniques to achieve layer-level object localization.

Compared to existing attention manipulation-based approaches, our method demonstrates significant superiority in background preservation metrics. In terms of CLIP-based foreground editing evaluation, our proposed approach also performs comparable results. Even though some methods achieve higher CLIP scores, this does not necessarily indicate better editing performance. As shown in Table 4, we can achieve a CLIP_{Whole} score of 25.46, which is better than those in Table 5. The trade-off between background preservation and CLIP score is primarily due to the biased evaluation of the CLIP model. Because the CLIP model is not specifically designed for image editing, it merely evaluates the similarity between the image and target semantics. Some editing methods directly inject full semantic information into target regions, artificially inflating CLIP scores without ensuring consistent image contexts. These comprehensive evaluation results $score_{avg}$, combining both foreground editing and background preservation performance, demonstrate the effectiveness of the proposed method. As illustrated in Table 6, we also compare our results with some of the latest baselines that utilise the Stable Diffusion 3.0 backbone. Our approach continues to perform well in LPIPS while achieving a comparable CLIP score.

4.4 | Efficiency Analysis

As shown in Table 3, we present the inference time cost of our method. On a single A40 GPU, the proposed method demonstrates a comparable efficiency with P2P and InfEedit, which are also evaluated with the LCM baseline and DDCM inversion technique. In contrast to these methods, our approach achieves a balance between efficiency and performance through semantic decoupling. We also validate the SIGMA on an RTX 4090 GPU, achieving a speed of up to 1.35 s per image, showcasing its potential for real-world applications.

4.5 | Ablation Study

As shown in Table 4, we present a component analysis of each module in the proposed framework. We initially evaluated the core components of our image editing framework by separately examining the roles of self-attention and cross-attention mechanisms. Experimental results demonstrate that relying solely on either cross-attention or self-attention is insufficient, as both tend to inadvertently alter the contextual semantics of non-edited regions without explicit guidance. Notably, the cross-attention branch achieves the highest score of both CLIP_{Whole} and CLIP_{Edited}. While this trade-off sacrifices background preservation, the edited regions exhibit stronger alignment with target semantics, as reflected in competitive CLIP scores.

Then we evaluate the performance of the semantic injection module. As illustrated in Figure 3, the ForeGround editing component (FG), which is calculated as the product of the foreground attention and semantics of the target region, serves as the foundational baseline for editing. We incrementally integrate FG with the BackGround preservation component (BG) and foreground enhancement component (FE). Although FG+BG shows marginal improvements in individual metrics, it delivers significant overall gains. This is because attention maps in the background region tend to have lower magnitudes, meaning that manipulation in this area exerts only a subtle influence. In contrast, FG+FE dramatically boosts foreground editing capability at the cost of background retention, as FE overamplifies target semantics. Importantly, BG can effectively mitigates the negative side effects of FE due to their overlapping regions, highlighting their complementary roles.

Finally, we validate the effectiveness of the semantic decoupling module. Using the mask generation operation M_p as a baseline, we notice only limited improvements in object localization from raw attention maps. Interestingly, neither the spatial (M_s) nor the temporal (M_t) attention manipulation operations alone significantly enhance editing performance. However, their combined application results in a breakthrough. Although M_s reduces intra-layer noise in attention maps and M_t utilises historical data to filter out unreliable layers, their synergy can be multiplicative. Therefore, their joint deployment is crucial for precise object-aware editing.

4.6 | Qualitative Results

Qualitative comparison with SOTAs: We present the comparison results of image editing with other SOTA approaches, as illustrated in Figure 5. We select several recent attention

TABLE 4 | Ablation study results on the PIE-Bench dataset.

Components			Structure Distance$^{\downarrow}_{\times 10^3}$	PSNR$^{\uparrow}$	SSIM$^{\uparrow}_{\times 10^2}$	LPIPS$^{\downarrow}_{\times 10^3}$	MSE$^{\downarrow}_{\times 10^4}$	CLIP$^{\uparrow}_{Whole}$	CLIP$^{\uparrow}_{Edited}$	$score_{avg}$
Only self-attention			27.79	21.30	80.26	88.23	131.34	24.73	21.75	0.38
Only cross-attention			32.78	26.55	83.95	68.47	74.50	**25.46**	**22.77**	0.44
S+C	FG	$M_p + M_s + M_t$	10.74	<u>29.15</u>	<u>86.53</u>	<u>43.44</u>	28.47	24.62	21.81	1.14
	FG + BG		<u>9.55</u>	28.86	86.35	44.33	<u>28.37</u>	24.70	21.88	<u>1.23</u>
	FG + FE		13.79	28.28	85.93	48.27	33.87	<u>25.17</u>	<u>22.30</u>	0.93
	FG + BG + FE	M_p	10.98	28.71	86.24	45.43	30.15	24.82	21.99	1.11
		$M_p + M_s$	10.98	28.73	86.26	45.22	30.05	24.80	22.00	1.11
		$M_p + M_t$	11.00	28.68	86.22	45.52	30.22	24.85	22.01	1.10
		$M_p + M_s + M_t$	**9.43**	**29.36**	**86.72**	**41.76**	**26.71**	24.49	21.71	**1.26**

Note: The bolds stand for the optimal values. The underlined indicate the sub-optimal values.

TABLE 5 | Comparison results on the PIE-Bench dataset.

Methods	Structure distance$^{\downarrow}_{\times 10^3}$	PSNR$^{\uparrow}$	SSIM$^{\uparrow}_{\times 10^2}$	LPIPS$^{\downarrow}_{\times 10^3}$	MSE$^{\downarrow}_{\times 10^4}$	CLIP$^{\uparrow}_{Whole}$	CLIP$^{\uparrow}_{Edited}$	$score_{avg}$
SIMSG Dhamo et al. [8]	>80	19.48	70	>400	>100	20.40	—	—
SGEdit Zhang et al. [30]	>60	22.45	79	>100	>100	24.19	—	—
P2P-Zero Parmar et al. [5]	49.22	21.53	77.05	138.98	127.32	23.31	21.05	0.24
PnP Tumanyan et al. [25]	24.29	22.46	79.68	106.06	80.45	**25.41**	**22.62**	0.45
MasaCtrl Cao et al. [27]	24.70	22.64	81.33	87.94	81.09	24.38	21.35	0.44
P2P Hertz et al. [4]	<u>11.65</u>	27.22	84.76	54.55	32.86	25.02	22.10	1.01
InfEdit Xu et al. [37]	13.78	<u>28.51</u>	<u>85.66</u>	<u>47.58</u>	32.09	<u>25.03</u>	<u>22.22</u>	0.95
Stable Flow Avrahami et al. [38]	22.29	22.41	85.65	93.97	87.81	23.68	20.93	0.46
CoLan Luo et al. [39]	13.97	28.46	85.12	53.04	—	24.94	22.45	0.69
SwiftEdit Nguyen et al. [40]	—	23.33	76.34	89.69	**6.60**	25.16	21.25	<u>1.14</u>
Ours	**9.43**	**29.36**	**86.72**	**41.76**	<u>26.71</u>	24.49	21.71	**1.26**

Note: The bolds stand for the optimal values. The underlined indicate the sub-optimal values.

TABLE 6 | Comparison results on the FlowEdit dataset.

Methods	LPIPS$^{\downarrow}$	CLIP$^{\uparrow}$
SDEdit Meng et al. [41]	<u>0.316</u>	**0.34**
iRFDS Yang et al. [42]	0.376	0.335
ODE Inv Kulikov et al. [31]	0.318	<u>0.337</u>
Ours	**0.273**	<u>0.337</u>

Note: The bolds stand for the optimal values. The underlined indicate the sub-optimal values.

manipulation-based approaches for comparison, including P2P [4], P2P-Zero [5], PnP [25], MasaCtrl [27] and InfEdit [37]. In the first row, although P2P, P2P-Zero and our method all achieve accurate target semantic editing, our approach significantly excels in preserving the semantics of nonedited regions. The second row demonstrates that although PnP, InfEdit and our method all maintain the image layout well, our edited targets display more complete and realistic semantics. For style transfer in the third row, P2P, P2P-Zero, and MasaCtrl lose layout control, whereas InfEdit fails the editing task entirely. For object removal in the fourth row, only P2P-Zero, InfEdit and our method succeed, but P2P-Zero and InfEdit introduce noticeable alterations to the background and foreground, respectively. When dealing with objects with complex poses in the fifth row, our method uniquely achieves both object replacement and original pose preservation. The final challenging case with less prominent edit targets shows PnP, MasaCtrl and InfEdit retaining original house details and failing full semantic replacement. P2P-Zero significantly loses layout control, and P2P produces inferior image quality compared to ours. These comprehensive comparisons validate the superior performance of our method across diverse editing scenarios.

Evaluation on each editing operation: We validate the effectiveness of SIGMA on nine different editing operations defined in Table 1. Initially, as illustrated in Figure 6, we evaluate the proposed method on object-level manipulations and background change. In scenarios involving object addition, both standalone objects and subordinate objects are examined. The method demonstrates robust foreground editing while preserving background consistency, although the positional ambiguity of newly added objects occasionally caused subtle interference, as seen in

FIGURE 5 | Compared to other image editing methods, our method produces both high-quality edited regions and highly consistent unedited backgrounds. The red boxes/texts in scene graphs are the source & target semantics to be edited.

the second row. Object removal includes both singular and multiple object eliminations. It is a particularly challenging task given the absence of auxiliary inpainting techniques and user intervention. Although background semantics are effectively utilised to reconstruct edited regions, residual object contours persist due to the inherent challenge of modifying attention layout without explicit guidance. In semantic replacement tasks including both object and background change, the method achieves notable fidelity.

Subsequently, we assess the performance of SIGMA in stylistic editing and expression semantic modification tasks, as depicted in Figure 7. For expression modification, we conduct comprehensive validation across three scenarios: close-up portraits, distant portraits and anthropomorphic animal depictions. Our approach consistently demonstrates efficacy in all test conditions. In colour manipulation experiments, we focus on three modification scales: face details, global adjustment and accessory alterations. Both texture editing and artistic stylisation are executed at the holistic level, as in usual cases. The proposed method demonstrates superior performance in both global and local semantic image editing, remaining effective across various editing paradigms while consistently preserving the image context.

Finally, as illustrated in Figure 8, we validate the effectiveness of our method in editing both pose and interaction semantics. For pose editing, we select three common target poses: holding out an arm for a handshake, folding arms, and waving hands. In terms of interaction editing, we select a typical two-person conversation scenario, controlling the target semantics through three usual interaction behaviours: handshaking, dancing and arguing. Although pose and behaviour editing operations also entail structural adjustments, our method effectively preserves the original context while transforming edited objects into the target semantics.

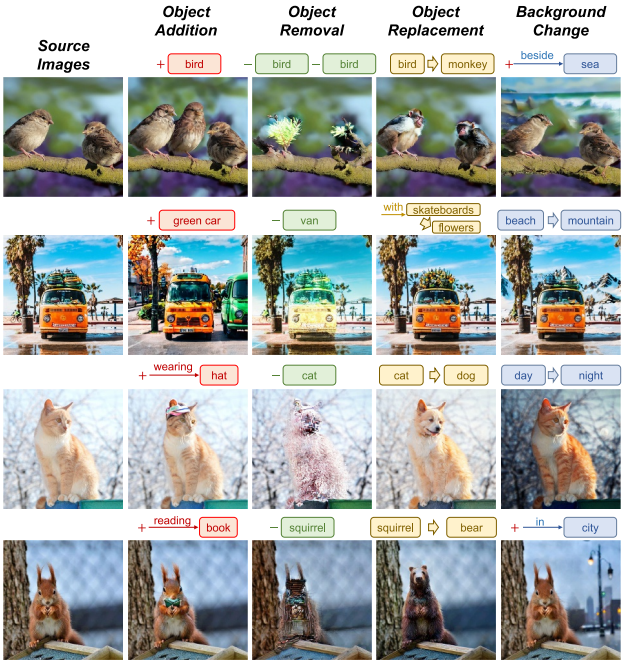

FIGURE 6 | Qualitative results on object-level modifications and background change.

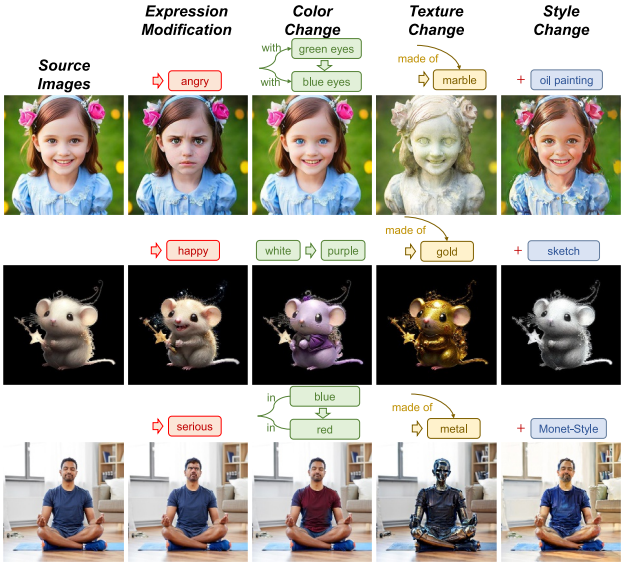

FIGURE 7 | Qualitative results on stylistic editing and expression modification.

4.7 | Limitations and Future Work

Limitations: Finally, we analyse the limitations of our proposed approach. Similar to other user-localisation-free methods, SIGMA shows limited performance when dealing with structural modifications. It excels in managing nonstructural edits where attention distribution remains stable, including object replacement, background change, style, colour and texture change. Although SIGMA remains functional for structural edits involving significant attention shifts, that is, object addition/removal and pose changes, its performance becomes sub-optimal compared to nonstructural edits. The residual contour artefacts mainly arise because in image editing process directly removing attention from the object-removed region will prevent the corresponding area from properly generating semantic content, thereby resulting in noisy artefacts. Mainstream methods address this issue by following inpainting strategies, which inject background semantics into the attention region of the removed object. As shown in Figure 9, our method maintains the attention map distribution consistent with the source image to preserve semantic consistency in cross attention layers, and achieves object removal by injecting background semantics. Furthermore, if we regenerate the entire layout from scratch based on the target scene graph without retaining the original attention distribution, it will cause significant changes to the background, leading to inconsistent visual content in nonedited regions before and after editing. Therefore, the residual artefact is actually a trade-off between the semantic consistency of nonedited regions and the editing quality of the target region. Layout modification without user intervention remains an open challenge in image editing. SIGMA addresses this by enabling the semantic exchange between foreground and background. For example, in the object removal

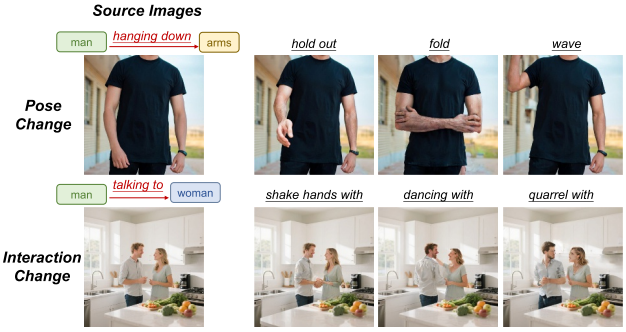

FIGURE 8 | Qualitative results on pose and interaction modification.

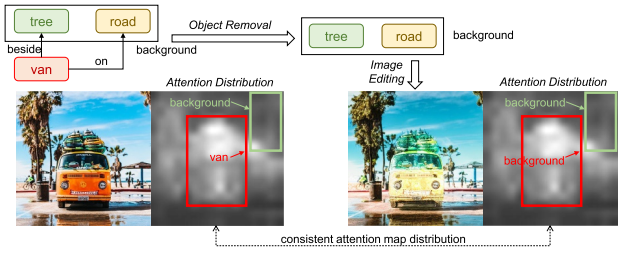

FIGURE 9 | SIGMA improves semantic consistency by maintaining consistent attention map distribution in cross attention layers, but it is prone to producing artefacts in the object removal scenario.

scenario depicted in Figure 6, background semantics are integrated into target objects to aid in object removal. Although this strategy is effective, it results in artefact contours due to the fixed attention maps.

Future Work: In our future work, we plan to explore several promising directions to further enhance the capabilities of image editing. First, we aim to extend the mutually exclusive attention manipulation mechanism to accommodate more complex structural edits, such as dynamic scene transformations and multi-object interaction rearrangements. This will involve integrating explicit geometric reasoning and spatial relation modelling into the ST-MEAN algorithm, which could address the current limitations in object removal task. Second, we intend to investigate the integration of large language models with our framework to enable more flexible and natural editing. This would allow users to input high-degree-of-freedom textual descriptions and automatically convert them into precise editing instructions. Third, we will explore the generalisation of SIGMA to arbitrary resolution image editing by optimising the attention manipulation pipeline and addressing potential performance degradation in large-size image processing. These extensions are expected to advance the state-of-the-art in controllable image editing, making the technology more versatile and practical for real-world scenarios.

5 | Conclusion

In this paper, we proposed SIGMA, a novel method for training/tuning-free and semantic-coherent image editing using scene graphs. SIGMA adopts a mutually exclusive attention manipulation strategy to decouple the semantic distribution of different objects. It comprises two key components: a semantic decoupling module based on spatial–temporal attention mutual exclusion to locate edited and unedited semantics, and a semantic injection module to conduct image editing and background preservation. We conducted comprehensive experiments to demonstrate the effectiveness of SIGMA in the PIE-Bench dataset. The results show that SIGMA not only maintains semantic consistency in backgrounds but also embraces the high-quality foreground editing capability with diffusion models, facilitating more controllable image editing.

Author Contributions

Yunqing He: investigation, methodology, software, writing – original draft. **Ruichao Hou:** project administration, validation. **Jia Bei:** supervision. **Tongwei Ren:** funding acquisition, resources, supervision.

Funding

This research was supported by the National Natural Science Foundation of China (Grants 62072232 and 92582103) and the Collaborative Innovation Center of Novel Software Technology and Industrialisation.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data availability is not applicable as all dataset used in the article are public available: PIE-Bench: https://github.com/cure-lab/PnPInversion; FlowEdit dataset: https://github.com/fallenshock/FlowEdit.

References

1. Z. Wang, L. Zhao, and W. Xing, "StyleDiffusion: Controllable Disentangled Style Transfer via Diffusion Models," in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE, 2023), 7643–7655, https://doi.org/10.1109/ICCV51070.2023.00706.

2. S. Sheynin, A. Polyak, U. Singer, et al., "Emu Edit: Precise Image Editing via Recognition and Generation Tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2024), 8871–8879, https://doi.org/10.1109/CVPR52733.2024.00847.

3. B. Yang, S. Gu, B. Zhang, et al., "Paint by Example: Exemplar-Based Image Editing With Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2023), 18381–18391, https://doi.org/10.1109/CVPR52729.2023.01763.

4. A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-Prompt Image Editing With Cross-Attention Control," in *Proceedings of the International Conference on Learning Representations* (OpenReview.net, 2023), 14369–14404, https://openreview.net/forum?id=_CDixzkzeyb.

5. G. Parmar, K. K. Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu, "Zero-Shot Image-to-Image Translation," in *Special Interest Group on Computer Graphics and Interactive Techniques* (2023), 1–11.

6. J. Johnson, R. Krishna, M. Stark, et al., "Image Retrieval Using Scene Graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2015), 3668–3678, https://doi.org/10.1109/CVPR.2015.7298990.

7. J. Johnson, A. Gupta, and Li Fei-Fei, "Image Generation From Scene Graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), 1219–1228, https://doi.org/10.1109/CVPR.2018.00133.

8. H. Dhamo, A. Farshad, I. Laina, et al., "Semantic Image Manipulation Using Scene Graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2020), 5212–5221, https://doi.org/10.1109/CVPR42600.2020.00526.

9. Z. Zhang, H. He, B. A. Plummer, Z. Liao, and H. Wang, "Complex Scene Image Editing by Scene Graph Comprehension," in *Proceedings of the British Machine Vision Conference* (2023), 451, http://proceedings.bmvc2023.org/451/.

10. Yi Huang, J. Huang, Y. Liu, et al., "Diffusion Model-Based Image Editing: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47, no. 6 (2025): 4409–4437, https://doi.org/10.1109/tpami.2025.3541625.

11. F. Wang, T. Zhang, Y. Wang, X. Zhang, X. Liu, and Z. Cui, "Scene Graph-Grounded Image Generation," in *Proceedings of the AAAI Conference on Artificial Intelligence* (2025), 7646–7654, https://doi.org/10.1609/aaai.v39i7.32823.

12. Wang, J., J. Hu, X. Ma, H. Ma, X. Wei, and E. Wu, "Image Editing With Diffusion Models: A Survey," CoRRabs/2504.13226 (2025).

13. R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-Text Inversion for Editing Real Images Using Guided Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2023), 6038–6047, https://doi.org/10.1109/CVPR52729.2023.00585.

14. S. Lu, Y. Liu, and A. W.-K. Kong, "TF-ICON: Diffusion-Based Training-Free Cross-Domain Image Composition," in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE, 2023), 2294–2305, https://doi.org/10.1109/ICCV51070.2023.00218.

15. H. Lee, M. Kang, and B. Han, "Conditional Score Guidance for Text-Driven Image-to-Image Translation," in *Advances in Neural Information Processing Systems* (2023), 38685–38708, http://papers.nips.cc/paper_files/paper/2023/hash/799f81cfa0611f93586c007024041460-Abstract-Conference.html.

16. C. H. Wu and F. De la Torre, "A Latent Space of Stochastic Diffusion Models for Zero-Shot Image Editing and Guidance," in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE, 2023), 7344–7353, https://doi.org/10.1109/ICCV51070.2023.00678.

17. B. Wallace, A. Gokul, and N. Naik, "EDICT: Exact Diffusion Inversion via Coupled Transformations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2023), 22532–22541, https://doi.org/10.1109/CVPR52729.2023.02158.

18. P. Dhariwal and A. Q. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2021), 8780–8794, https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

19. J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models," in *Proceedings of the International Conference on Learning Representations* (2021), 14205–14224, https://openreview.net/forum?id=St1giarCHLP.

20. X. Ju, A. Zeng, Y. Bian, S. Liu, and Q. Xu, "PnP Inversion: Boosting Diffusion-Based Editing With 3 Lines of Code," in *Proceedings of the International Conference on Learning Representations* (2024), 6136–6163, https://openreview.net/forum?id=FoMZ4ljhVw.

21. O. Avrahami, O. Fried, and D. Lischinski, "Blended Latent Diffusion," *ACM Transactions on Graphics* 42, no. 4 (2023): 149:1–149:11, https://doi.org/10.1145/3592450.

22. Z. Yu, H. Li, F. Fu, X. Miao, and B. Cui, "Accelerating Text-to-Image Editing via Cache-Enabled Sparse Diffusion Inference," in *Proceedings of the AAAI Conference on Artificial Intelligence* (2024), 16605–16613, https://doi.org/10.1609/aaai.v38i15.29599.

23. G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, "DiffEdit: Diffusion-Based Semantic Image Editing With Mask Guidance," in *Proceedings of the International Conference on Learning Representations* (2023), 14324–14345, https://openreview.net/forum?id=3lge0p5o-M-.

24. Z. Yang, G. Ding, W. Wang, H. Chen, B. Zhuang, and C. Shen, "Object-Aware Inversion and Reassembly for Image Editing," in *Proceedings of the International Conference on Learning Representations* (2024), 20036–20059, https://openreview.net/forum?id=dpcVXiMlcv.

25. N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "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* (IEEE, 2023), 1921–1930, https://doi.org/10.1109/CVPR52729.2023.00191.

26. Or Patashnik, D. Garibi, I. Azuri, H. Averbuch-Elor, and D. Cohen-Or, "Localizing Object-Level Shape Variations With Text-to-Image Diffusion Models," *Proceedings of the IEEE International Conference on Computer Vision* (IEEE, 2023), 22994–23004, https://doi.org/10.1109/ICCV51070.2023.02107.

27. M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng, "MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing," in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE, 2023), 22503–22513, https://doi.org/10.1109/ICCV51070.2023.02062.

28. S. Wu, H. Fei, H. Zhang, and T.-S. Chua, "Imagine That! Abstract-to-Intricate Text-to-Image Synthesis With Scene Graph Hallucination Diffusion," in *Advances in Neural Information Processing Systems* (2023), 79240–79259, http://papers.nips.cc/paper_files/paper/2023/hash/fa64505ebdc94531087bc81251ce2376-Abstract-Conference.html.

29. S. Su, L. Gao, J. Zhu, J. Shao, and J. Song, "Fully Functional Image Manipulation Using Scene Graphs in A Bounding-Box Free Way," in *Proceedings of the ACM International Conference on Multimedia* (Association for Computing Machinery, 2021), 1784–1792, https://doi.org/10.1145/3474085.3475326.

30. Z. Zhang, D. Chen, and J. Liao, "SGEdit: Bridging LLM With Text2 Image Generative Model for Scene Graph-Based Image Editing," *ACM Transactions on Graphics* 43, no. 6 (2024): 195:1–195:16, https://doi.org/10.1145/3687957.

31. V. Kulikov, M. Kleiner, I. Huberman-Spiegelglas, and T. Michaeli, "FlowEdit: Inversion-Free Text-Based Editing Using Pre-Trained Flow Models," in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE, 2025), 19721–19730, https://doi.org/10.1109/ICCV51701.2025.01834.

32. N. Tumanyan, O. Bar-Tal, S. Bagon, and T. Dekel, "Splicing ViT Features for Semantic Appearance Transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2022), 10738–10747, https://doi.org/10.1109/CVPR52688.2022.01048.

33. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing* 13, no. 4 (2004): 600–612, https://doi.org/10.1109/tip.2003.819861.

34. R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 586–595, http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

35. A. Radford, J. W. Kim, C. Hallacy, et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the International Conference on Machine Learning* (2021), 8748–8763, http://proceedings.mlr.press/v139/radford21a.html.

36. J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A Reference-Free Evaluation Metric for Image Captioning," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2021), 7514–7528, https://doi.org/10.18653/v1/2021.emnlp-main.595.

37. S. Xu, Y. Huang, J. Pan, Z. Ma, and C. Joyce, "Inversion-Free Image Editing With Language-Guided Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2024), 9454–9461, https://doi.org/10.1109/CVPR52733.2024.00903.

38. O. Avrahami, Or Patashnik, O. Fried, et al., "Stable Flow: Vital Layers for Training-Free Image Editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2025), 7877–7888, https://openaccess.thecvf.com/content/CVPR2025/html/Avrahami_Stable_Flow_Vital_Layers_for_Training-Free_Image_Editing_CVPR_2025_paper.html.

39. J. Luo, T. Ding, K. Ho R. Chan, H. Min, C. Callison-Burch, and R. Vidal, "Concept Lancet: Image Editing With Compositional Representation Transplant," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2025), 28502– 28512, https://openaccess.thecvf.com/content/CVPR2025/html/Luo_Concept_Lancet_Image_Editing_with_Compositional_Representation_Transplant_CVPR_2025_paper.html.

40. T.-T. Nguyen, Q. Nguyen, K. Nguyen, A. T. Tran, and C. Pham, "SwiftEdit: Lightning Fast Text-Guided Image Editing via One-Step Diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2025), 21492–21501, https://openaccess.thecvf.com/content/CVPR2025/html/Nguyen_SwiftEdit_Lightning_Fast_Text-Guided_Image_Editing_via_One-Step_Diffusion_CVPR_2025_paper.html.

41. C. Meng, Y. He, Y. Song, et al., "SDEdit: Guided Image Synthesis and Editing With Stochastic Differential Equations," in *Proceedings of the International Conference on Learning Representations* (2022), 13781–13813, https://openreview.net/forum?id=aBsCjcPu_tE.

42. X. Yang, C. Chen, X. Yang, F. Liu, and G. Lin, "Text-to-Image Rectified Flow as Plug-and-Play Priors," in *Proceedings of the International Conference on Learning Representations* (2025), 80086–80110, https://openreview.net/forum?id=SzPZK856iI.

43. Luo, S., Y. Tan, L. Huang, J. Li, and H. Zhao, "Latent Consistency Models: Synthesizing High-Resolution Images With Few-Step Inference," CoRR (2023).

44. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis With Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2022), 10674–10685, https://doi.org/10.1109/CVPR52688.2022.01042.

45. B. Kawar, S. Zada, O. Lang, et al., "Imagic: Text-Based Real Image Editing With Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2023), 6007–6017, https://doi.org/10.1109/CVPR52729.2023.00582.