ObjMST: Object-focused Multimodal Style Transfer[⋆]

Chanda Grover Kamra[a,*,1], Indra Deep Mastan[b] and Debayan Gupta[a]

[a] *Ashoka University, Rajiv Gandhi Education City, Sonipat, 131029, Haryana, India*

[b] *Indian Institute of Technology, BHU, , Varanasi, 221005, Uttar Pradesh, India*

ARTICLE INFO

Keywords:
Multimodal
Style Transfer
CLIP
Segmentation
Harmonization

ABSTRACT

We propose ObjMST, an object-focused multimodal style transfer framework that provides separate style supervision for salient objects and surrounding elements while addressing alignment issues in multimodal representation learning. Existing image-text multimodal style transfer methods face the following challenges: (1) generating non-aligned and inconsistent multimodal style representations; and (2) content mismatch, where identical style patterns are applied to both salient objects and their surrounding elements. Our approach mitigates these issues by: (1) introducing a Style-Specific Masked Directional CLIP Loss, which ensures consistent and aligned style representations for both salient objects and their surroundings; and (2) incorporating a salient-to-key mapping mechanism for stylizing salient objects, followed by image harmonization to seamlessly blend the stylized objects with their environment. We validate the effectiveness of ObjMST through experiments, using both quantitative metrics and qualitative visual evaluations of the stylized outputs. Our code is available at: https://anonymous.4open.science/r/ObjMST-4DC0.

1. Introduction

Classical Image-guided Image Style Transfer (IIST) methods primarily rely on a style image to supervise the stylization process [1, 11, 10, 6]. In contrast, multimodal learning [12] integrates additional modalities, such as textual descriptions [16, 15] or audio cues [7] to guide or augment the stylization [10, 6]. There also exist uni-modal (image or text) [11, 10, 6] and multimodal (image and text) [19] guided methods for IST.

IIST methods that rely solely on image data are limited in practical use when style images are unavailable [10] or when unique styles are lacking [19]. Consequently, there has been a shift towards Text-Guided Image Style Transfer (TIST) methods [10, 6, 15]. However, TIST approaches often introduce textual artifacts [10] due to challenges in achieving effective style alignment and content preservation [10, 6].

In a recent study, Wang et al. [19] proposed a novel cross-modal GAN inversion-based style generation mechanism. This mechanism is guided by both style text and style image, ensuring consistency across modalities.

We observed that MMIST [19] encounters difficulties in preserving semantics in the style transfer. For instance, Fig. 1, top row (a-d) shows that MMIST [19] fails to capture color and texture tones corresponding to the style-text description "Copper plate engraving". This limitation is primarily due to the difficulty in aligning text and image [21] features within the shared CLIP (Contrastive Language-Image Pre-training) [16] embedding space.

Table 1
The table compares Image Style Transfer (IST) methods. 'Single' in columns one and three indicates a single text condition, while 'double' in column two refers to double text conditions for foreground (fg) and background (bg) objects.

	Text-Based IST (Single)	**Text-Based IST (Double)**	**Multimodal IST (Single)**
CS [10]	√	X	X
SemCS [6]	√	√	X
LDAST [2]	√	X	X
MMIST [19]	√	X	√
ObjMST (Ours)	√	√	√

Another issue in MMIST [19] is content mismatch, where a similar style is applied to semantically different objects, resulting in unintended spill-over. For example, Fig. 1 (e-f) shows that the texture or patterns from the style features spill over into regions such as the background or other objects, producing a visually incoherent result in MMIST [19]. SemCS [6] and CLVA [2] were also observed to be suffering from content mismatch (Fig. 1).

We address the issue of misalignment in style transfer by introducing ObjMST, the first object-focused multimodal style transfer framework. The main challenge lies in performing StyleGAN inversion, guided by directional CLIP loss, to generate intermediate style representations. These representations are then used in image-based style transfer to produce the final output. Our analysis of the intermediate style representations (Table 2 and Fig. 7) demonstrates improved visual quality when both salient objects and their surrounding elements are considered. To further enhance the representations of salient objects, we mask out arbitrary content features from the multimodal style input during the StyleGAN inversion process. This ensures the preservation of essential style elements in the intermediate stages. This entire process leads to the development of the "masked" directional CLIP loss, which ensures consistent and aligned generation of intermediate style representations.

[⋆]

[*] Corresponding author
chanda.grover_phd19@ashoka.edu.in (C.G. Kamra); indra.cse@itbhu.ac.in (I.D. Mastan); debayan.gupta@ashoka.edu.in (D. Gupta)
https://chandagrover.github.io/ (C.G. Kamra)
ORCID(s):

[1]

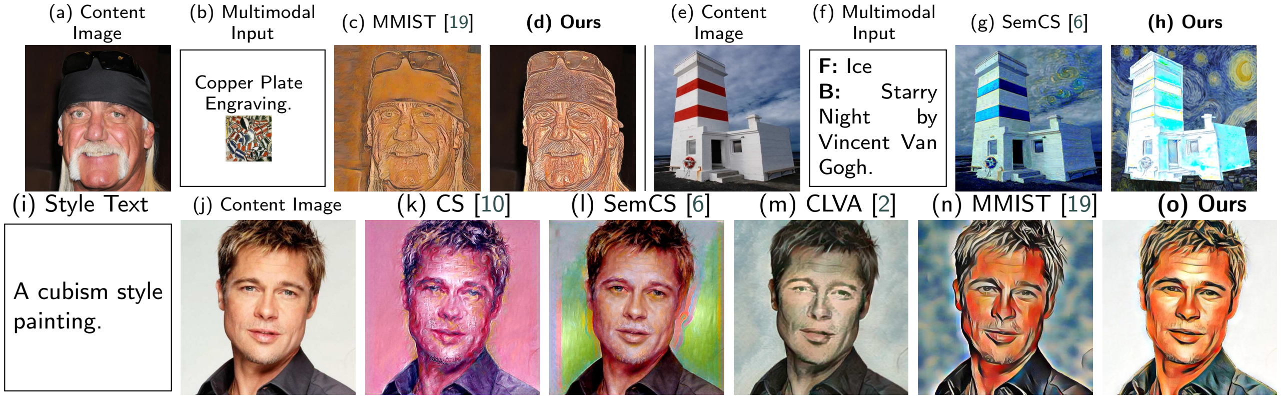

Figure 1: The comparative stylized outputs are presented as follows: (i) Top Row, Left Side: MMIST (Single); (ii) Top Row, Right Side: TIST (Double); and (iii) Bottom Row: TIST (Single). Columns (a, e, j) represent the content images, while columns (b, f, i) show the multimodal style inputs. In MMIST [19] (column c), misalignment is evident as the texture and color of the copper plate features are inconsistent compared to Ours-ObjMST (column d). In TIST (Double), SemCS [6] (column g) introduces undesired distortions, whereas ObjMST (column h) correctly applies the "Starry Night" style features to the background (sky) and ice features to the foreground. In TIST (Single), ObjMST (column o) effectively preserves the content features while accurately applying the desired style.

ObjMST resolves the content mismatch between salient objects and surrounding elements by employing salient-to-key feature matching [23]. The alignment of each salient object's content features with stable key positions of style features (S2K) facilitates semantically preserved stylization, focusing exclusively on the salient object within the image. For example, in Fig. 1-(h), ice features are applied solely to the building.

ObjMST generates distinct style representations for salient objects and surrounding elements using cross-model GAN inversion, assisted by masked directional CLIP loss for the foreground (fg) and background (bg) objects. Separating the style representation generation for fg and bg is essential to minimizing content mismatches. To ensure consistency in the final output, self-supervised image harmonization is employed [5]. Our major contributions are as follows:

- We propose the ObjMST framework, which provides separate style supervision for salient objects and surrounding elements, while addressing alignment issues in multimodal representation learning (Fig. 6 and Fig. 5).
- The ObjMST framework maps salient content features to stable style key features *(S2K)* to preserve the semantic structure of content features (Fig. 3 and Table 3).
- We validate experimental results rigorously using Contrique [13], NIMA [18] scores, and user studies (Table 3).

2. Related Work

Arbitrary Style Transfer. Arbitrary style transfer (AST) methods [1, 20, 11] have gained attention for their effectiveness in transferring arbitrary styles. Local feature distribution methods often employ attention-based feature matching [20, 11]. Adaattn [11] introduced dense correspondence via all-to-all attention [11] but it faced challenges with distorted patterns and unstable matching. The All-to-Key attention mechanism [23] addresses these issues by identifying stable style keys through distribution and progressive attention.

Multimodality-Guided Image Style Transfer. Text-guided style transfer [10, 6, 2, 19] has gained attraction by eliminating the need for a style image. However, methods such as CLIPStyler (CS) [10] and CLVA [2] encounter issues with spatial misalignment and overstylization [6]. SemCS [6] addresses these problems through controllable transfer using global foreground and background loss, while MMIST [19], the first multimodal style transfer method, also suffers from misalignment.

CLIP and StyleGAN. CLIP plays a significant role in multimodal contexts, bridging the gap between visual and textual information. StyleCLIP [15] introduced global CLIP loss to minimize cosine distance between the generated image and target text, while StyleGAN-Nada [3] employed local directional loss for better diversity. Images are either encoded or inverted into latent space for editing via latent vector manipulation [14]. Recent models use CLIP [16] embeddings to guide text-based editing [3, 10, 15].

3. Methodology

We illustrate Object-focused Multimodal Style-transfer framework (ObjMST) in Fig. 2. It comprises of two steps: generating style representations and performing object-focused style transfer. Both steps require the segmentation mask of the content image as described in the below subsections. We consider the Segment Anything method [9] to obtain the segmentation mask of the content image.

3.1. Generation of Style Representations

We generate two distinct style representations for the foreground (fg) and background (bg), referred to as the salient and surrounding style representations.

Salient-Style Representations. We propose style-specific "*masked*" directional CLIP loss to synthesize consistent salient-style representations S_{fg} as shown in Fig. 2. First, we pass foreground-input style text-image pairs $(T_{S_{fg}}, I_{S_{fg}})$ to cross-modal GAN inversion. Next, the latent vector w_{fg} is

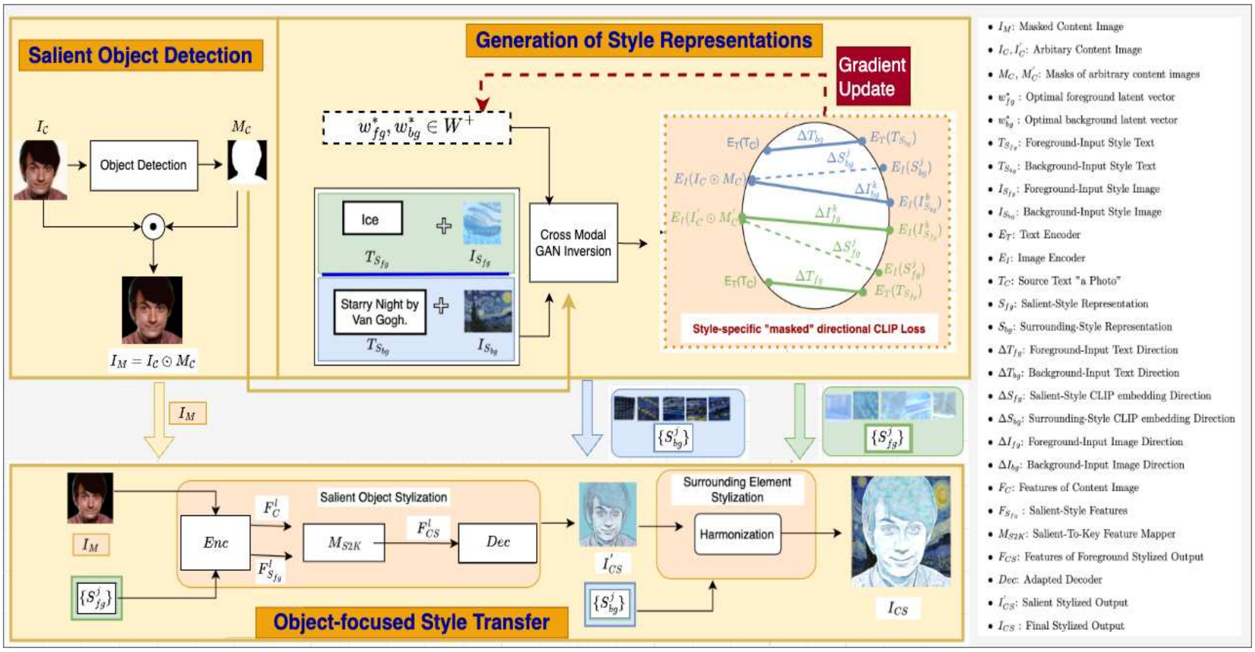
Salient Object Detection
I_C
Object Detection
M_C
$I_M = I_C \odot M_C$
Generation of Style Representations
Gradient Update
$w^*_{fg}, w^*_{bg} \in W^+$
Ice
$T_{S_{fg}}$
$I_{S_{fg}}$
Starry Night by Van Gogh.
$T_{S_{bg}}$
$I_{S_{bg}}$
Cross Modal GAN Inversion
$E_T(T_C)$
ΔT_{bg}
$E_T(T_{S_{bg}})$
ΔS^j_{bg}
$E_I(S^j_{bg})$
$E_I(I_C \odot M_C)$
ΔI^k_{bg}
$E_I(I^k_{S_{bg}})$
$E_I(I'_C \odot M'_C)$
ΔI^k_{fg}
$E_I(I^k_{S_{fg}})$
ΔS^j_{fg}
$E_I(S^j_{fg})$
$E_T(T_C)$
ΔT_{fg}
$E_T(T_{S_{fg}})$
Style-specific "masked" directional CLIP Loss
I_M
$\{S^j_{bg}\}$
$\{S^j_{fg}\}$
I_M
Salient Object Stylization
Enc
F^l_C
$F^l_{S_{fg}}$
M_{S2K}
F^l_{CS}
Dec
$\{S^j_{fg}\}$
I'_{CS}
Surrounding Element Stylization
Harmonization
$\{S^j_{bg}\}$
I_{CS}
Object-focused Style Transfer
• I_M: Masked Content Image
• I_C, I'_C: Arbitary Content Image
• M_C, M'_C: Masks of arbitrary content images
• w^*_{fg} : Optimal foreground latent vector
• w^*_{bg} : Optimal background latent vector
• $T_{S_{fg}}$: Foreground-Input Style Text
• $T_{S_{bg}}$: Background-Input Style Text
• $I_{S_{fg}}$: Foreground-Input Style Image
• $I_{S_{bg}}$: Background-Input Style Image
• E_T: Text Encoder
• E_I: Image Encoder
• T_C: Source Text "a Photo"
• S_{fg}: Salient-Style Representation
• S_{bg}: Surrounding-Style Representation
• ΔT_{fg}: Foreground-Input Text Direction
• ΔT_{bg}: Background-Input Text Direction
• ΔS_{fg}: Salient-Style CLIP embedding Direction
• ΔS_{bg}: Surrounding-Style CLIP embedding Direction
• ΔI_{fg}: Foreground-Input Image Direction
• ΔI_{bg}: Background-Input Image Direction
• F_C: Features of Content Image
• $F_{S_{fg}}$: Salient-Style Features
• M_{S2K}: Salient-To-Key Feature Mapper
• F_{CS}: Features of Foreground Stylized Output
• Dec: Adapted Decoder
• I'_{CS}: Salient Stylized Output
• I_{CS} : Final Stylized Output

Figure 2: The figure illustrates the proposed ObjMST framework. Given the segmentation mask (M_C) of the content image (I_C), we compute the masked content image (I_M). In Step 1, we compute the optimal foreground and background latent vector w^*_{fg} and w^*_{bg} to obtain the salient and surrounding style representations S_{fg} and S_{bg}. This is achieved by passing multimodal input of foreground-input style text-image pair ($T_{S_{fg}}$,$I_{S_{fg}}$) and background-input style text-image pair ($T_{S_{bg}}$,$I_{S_{bg}}$) to cross-modal GAN Inversion, which is trained using the proposed masked directional Style CLIP Loss (L_{fg}). In Step 2, the foreground stylized output (I'_{CS}) is generated by mapping salient content features (F^l_C) to stable style key features ($F^l_{S_{fg}}$) through (M_{S2K}) mapper. Finally, surrounding-style (S_{bg}) representation is applied to the background through image harmonization to generate stylized output I_{CS}.

passed to StyleGAN3 [8] generator G with random initialization to obtain $S_{fg} = G(w_{fg})$. Subsequently, the foreground-input style image ($I_{S_{fg}}$) and salient-style representations (S_{fg}) undergo cropping and augmentation to produce patch-level representations $I^k_{S_{fg}}$ and S^j_{fg}, respectively, where j and k denote the indices of the number of patches (N_{crop}).

The directional CLIP loss used in MMIST [19] generates imprecise style representations due to the misalignment between input image and text style features in the shared CLIP embedding space. To address this issue, we introduce a masked directional CLIP loss, which excludes irrelevant content features from CLIP space.

The first part of the proposed loss function involves computing the average cosine similarity between CLIP embedding directions of the patched image-image pair, specifically the foreground-input style image, ($I^k_{S_{fg}}$) and salient-style representations, (S^j_{fg}). The computation of foreground-input CLIP embedding direction (ΔI^k_{fg}) explicitly masks out the arbitrary content features ($I_C \odot M_C$) from I^k_{fg} as described in the equation below:

$$\Delta I^k_{fg} = E_I(I^k_{S_{fg}}) - E_I(I_C \odot M_C) \tag{1}$$

Here, E_I denotes the image encoder. Similarly, salient-style CLIP embedding direction (ΔS^j_{fg}) removes arbitrary content features from the salient-style representations as shown below:

$$\Delta S^j_{fg} = E_I(S^j_{fg}) - E_I(I_C \odot M_C) \tag{2}$$

The second part of the proposed loss function involves computing average cosine similarity between CLIP embedding directions of image-text pair, specifically foreground-input style text, ($T_{S_{fg}}$) and the patched salient-style representation (S^j_{fg}). The CLIP embedding direction of foreground-input style text ΔT_{fg} is defined as the difference between foreground-input style text embeddings $T_{S_{fg}}$ and source text embeddings (T_C) i.e $\Delta T_{fg} = E_T(T_{S_{fg}}) - E_T(T_C)$. Here, T_C represents the style text for any natural image and is set to "a photo" [10], while E_T denotes text encoder. The masked directional CLIP loss is defined as the averaged cosine similarity for each pair of CLIP embedding directions (image-text and image-image) as follows:

$$\begin{aligned} L_{fg} = \frac{1}{N_{\text{crop}}} \sum_{j=1}^{N_{\text{crop}}} \left(1 - \frac{\Delta S^j_{fg} \cdot \Delta T_{fg}}{\|\Delta S^j_{fg}\| \|\Delta T_{fg}\|} \right) + \\ \lambda * \frac{1}{N^2_{\text{crop}}} \sum_{j=1}^{N_{\text{crop}}} \sum_{k=1}^{N_{\text{crop}}} \left(1 - \frac{\Delta S^j_{fg} \cdot \Delta I^k_{fg}}{\|\Delta S^j_{fg}\| \|\Delta I^k_{fg}\|} \right) \end{aligned} \tag{3}$$

Here, λ is the tunable parameter. We minimize the L_{fg} to find the optimal foreground latent vector (w^*_{fg}), where $w^*_{fg} = argmin_w L_{fg}$. Finally, we pass the foreground optimal vector (w^*_{fg}) to the generator G to obtain the salient-style representations, i.e., $S_{fg} = G(w^*_{fg})$.

Surrounding-Style Representations. Similarly to salient-style representations, we determine the optimal background latent vector (w^*_{bg}) to generate surrounding-style representations S_{bg}. To achieve this, the background-input style text-image pair ($T_{S_{bg}}, I_{S_{bg}}$) is used as input. As illustrated in Fig. 2, we subtract the Hadamard product of another randomly chosen content image I'_C and its segmentation mask M'_C from the patched background-input and surrounding-style representations ($I^k_{S_{bg}}$ and S^j_{bg}) to compute their corresponding CLIP embedding directions (ΔI^k_{bg} and ΔS^j_{bg}). To ensure that these representations are not biased by features of a specific arbitrary content image, another arbitrary content image I'_C and its segmentation mask M'_C are chosen. Thus, the background-input CLIP embedding direction (ΔI^k_{bg}) is formulated as follows:

$$\Delta I^k_{bg} = E_I(I^k_{S_{bg}}) - E_I(I'_C \odot M'_C) \tag{4}$$

Similarly, surrounding-style CLIP embedding direction (ΔS^j_{bg}) is defined below:

$$\Delta S^j_{bg} = E_I(S^j_{bg}) - E_I(I'_C \odot M'_C) \tag{5}$$

The background (w^*_{bg}) latent vector for surrounding-style representations is optimized by minimizing L_{bg}, i.e. $w^*_{bg} = argmin_w L_{bg}$, where L_{bg} is defined as follows:

$$\begin{aligned} L_{bg} = \frac{1}{N_{\text{crop}}} \sum_{j=1}^{N_{\text{crop}}} \left(1 - \frac{\Delta S^j_{bg} \cdot \Delta T_{bg}}{\|\Delta S^j_{bg}\| \|\Delta T_{bg}\|} \right) + \\ \lambda * \frac{1}{N^2_{\text{crop}}} \sum_{j=1}^{N_{\text{crop}}} \sum_{k=1}^{N_{\text{crop}}} \left(1 - \frac{\Delta S^j_{bg} \cdot \Delta I^k_{bg}}{\|\Delta S^j_{bg}\| \|\Delta I^k_{bg}\|} \right) \end{aligned} \tag{6}$$

Finally, the surrounding-style representation S_{bg} is obtained by passing the optimal background vector to Stylegan3 [8] generator G, i.e., $S_{bg} = G(w^*_{bg})$. Incorporating the masked component in the computation of CLIP embedding direction ensures meaningful alignment between the generated style representations and multimodal inputs.

3.2. Object Stylization

The bottom part of Fig. 2 illustrates the object-focused style transfer step. The goal is to use salient-style representations S_{fg} to stylize the salient object of the content image I_C and use surrounding-style representations S_{bg} to stylize the background part of the content image I_C .

Salient Object Stylization. We extract content features (F_C) from the masked content image ($I_M = I_C \odot M_C$) and style features ($F_{S_{fg}}$) from salient-style representations S_{fg}, (obtained in Sec. 3.1). To obtain the multi-scale content F^l_C and style features $F^l_{S_{fg}}$ at different layers l such as $Relu\{3\_1, 4\_1, 5\_1\}$, we pass I_M and patched salient-style representations S^j_{fg} through a pretrained VGG19 [17] encoder (Enc) with fixed parameters. Here, $F \in R^{C_l * H_l * W_l}$, where C_l, H_l and W_l are the number of channels, height and width of image at different layers l. The feature extraction of F^l_C and $F^l_{S_{fg}}$ is described below:

$$F^l_C = Enc(I_C \odot M_C), \quad F^l_{S_{fg}} = Enc(S^j_{fg}) \tag{7}$$

Motivated by attention-based arbitrary style transfer methods [11, 23], we employ all-to-key [23] mapping on salient objects features to perform Salient-To-Key (S2K) feature mapping. This introduces Salient-To-Key feature mapper M_{S2K} to generate transformed multi-scale features of foreground stylized object F^l_{CS} as: $F^l_{CS} = M_{S2K}(F^l_C, F^l_{S_{fg}})$. M_{S2K} module maps salient-style features ($F^l_{S_{fg}}$) with features of content image (F^l_C) while preserving content integrity through distributive and progressive attention [23]. Finally, the salient stylized output I'_{CS} is reconstructed from multi-scale transferred feature of stylized output$\{F^l_{CS}\}$ by passing them through a decoder (Dec). We adopted the decoder (Dec) from AdaAttn [11], and is described as $I'_{CS} = Dec(\{F^l_{CS}\})$.

Surrounding Element Stylization. To stylize the background of content image, we realistically blend the surrounding-style representations $\{S^j_{bg}\}$ with the foreground stylized output (I'_{CS}). We utilized the self-supervised Harmonization framework (SSH [5]), which is trained on arbitrary unedited content images. In this way, we obtain the final stylized output I_{CS}, where distinct style features applied separately to foreground and background.

4. Experimental Results

We evaluate our method across three tasks: 1) single-condition stylization through text; 2) single condition stylization through multimodal input on salient objects; 3) and double-condition stylization using text input for both salient and surrounding elements [1]

4.1. Text-Based IST (Single)

Qualitative Analysis. As shown in Fig. 3, CS [10] distorts the original content and compromises the structure of stylized output, while SemCS [6] improves some outputs but still over-stylizes. CLVA [2] poorly matches styles to text, often producing the same color. MMIST [19] applies styles across the entire image, neglecting salient content. In contrast, our method accurately stylizes only the foreground, preserving content without affecting the background.

Fig. 4 illustrates content mismatch in the MMIST [19], CLVA [2], SemCS [6], and CS [6] methods, where style features undesirably affect the ground and sky in both the examples. In contrast, ObjMST confines style application to objects such as car and building, ensuring better semantic coherence.

Style Representations Quantitative Evaluation. We evaluate generated style representations S, consisting of both

[1] Extended results are provided in the supplementary material.

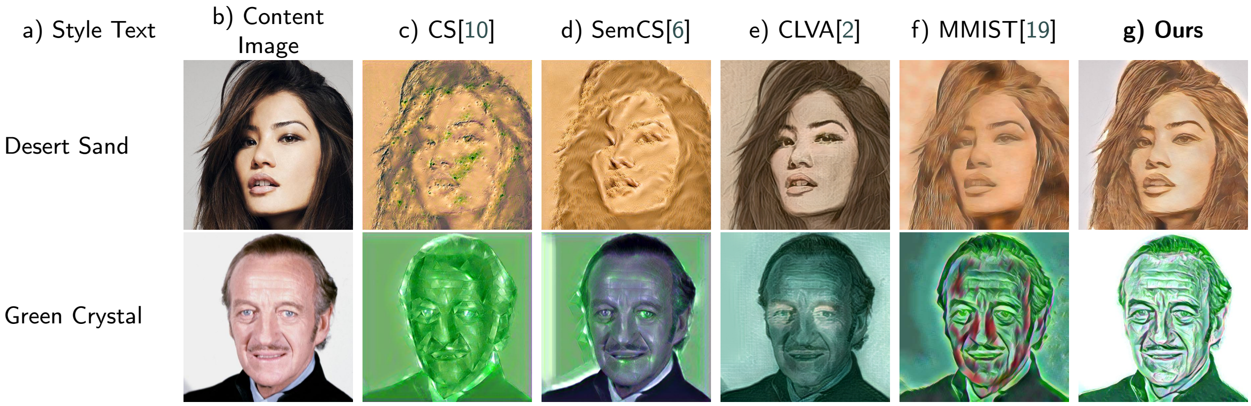

Figure 3: Text-based IST (Single). ObjMST (g) better preserves the facial structure and harmoniously integrates text and visual style cues, particularly in terms of texture and color consistency, compared to the baseline methods (c-f).

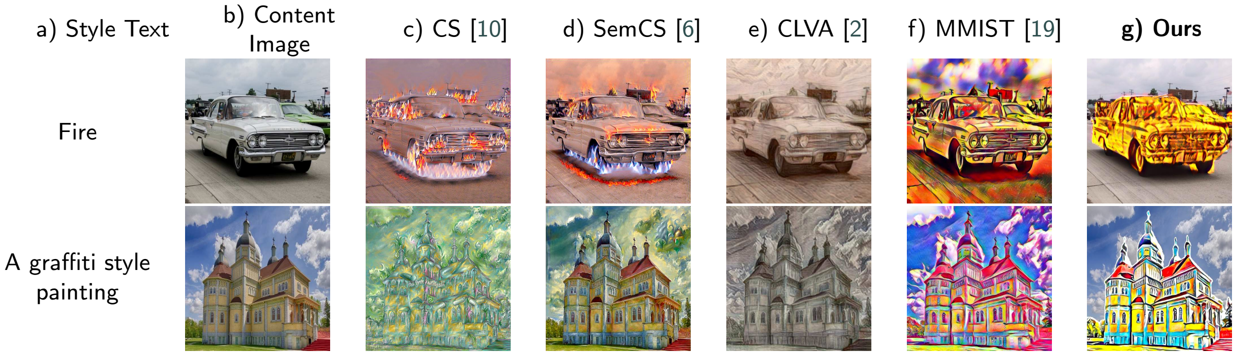

Figure 4: Content Mismatch. This figure illustrates content mismatch issues in style transfer methods, such as fire appearing on the car (first row) and graffiti style features on the building (second row). It can be observed that ObjMST (Ours) effectively minimizes these mismatches, producing more coherent stylized outputs.

Table 2
Style Representations Evaluation. T_S, I_S, and S represent the input style text, image, and generated style representations. Each of these has corresponding foreground (fg) and background bg equivalents, i.e. $T_S{=}T_{S_{fg}}$and $T_{S_{bg}}$, $I_S{=}I_{S_{fg}}$and $I_{S_{bg}}$ and $S{=}S_{fg}$and S_{bg}. Here, <> refers to cosine similarity. The average Clipscore [4] between generated style representations and image-text pair is higher, while the LPIPS [22] is lower with ObjMST.

Method	**Clipscore↑** [4]			**LPIPS↓** [22]
	<T_S, S>	<I_S,S>	Average	
MMIST [19]	0.76	0.48	0.46	0.31
Ours	0.88	0.56	**0.51**	**0.26**

salient and surrounding (S_{fg} and S_{bg}) style representations using Clipscore [4] and LPIPS [22] score. A higher Clipscore [4], indicates better alignment of salient and surrounding style representations (S_{fg} and S_{bg}) with respect to multimodal input style text-image pairs ($\{T_{S_{fg}}, I_{S_{fg}}\}$ and $\{T_{S_{bg}}, I_{S_{bg}}\}$, respectively). This improved alignment occurs because the proposed loss function excludes irrelevant content features from CLIP embedding directions during the generation process. Consequently, the style representations are closely aligned with the style text (column 2) and style image (column 3) of Table 2. Similarly, a lower LPIPS [22] score indicates better perceptual quality of generated style representations. We use foreground-input ($I_{S_{fg}}$) and background-input ($I_{S_{bg}}$) style image as reference image for computing LPIPS score on salient and surrounding style representations.

Stylized Outputs Quantitative Evaluation. We evaluated stylized outputs using both Full-Reference (FR) based - clipscore [4], LPIPS [22], Contrique-FR [13] and Non-Reference(NR) based - Nima [18] and Contrique (NR) [13] metrics. For FR metrics, the input style image serves as the reference image for contrique-FR and LPIPS, while the input style text is considered as the reference text for clipscore [4] across all baseline methods. ObjMST outperforms baseline methods (See Table 3); demonstrating superior visual quality, measured by NIMA [18]); enhanced perceptual quality, evaluated by LPIPS [22], Contrique [13]) in both FR and NR evaluations; and better alignment of style text with the stylized outputs, measured by clipscore [4]. Additionally, we conducted a user study to support this evaluation.

User Study. We applied 20 multimodal styles to 30 distinct content images, generating 600 stylized images per method. One hundred participants evaluated the images in a user study, voting for those that demonstrated the best style consistency and content preservation, particularly in the foreground of the content image. Table 3 presents the percentage of votes, highlighting the superior performance of the proposed ObjMST.

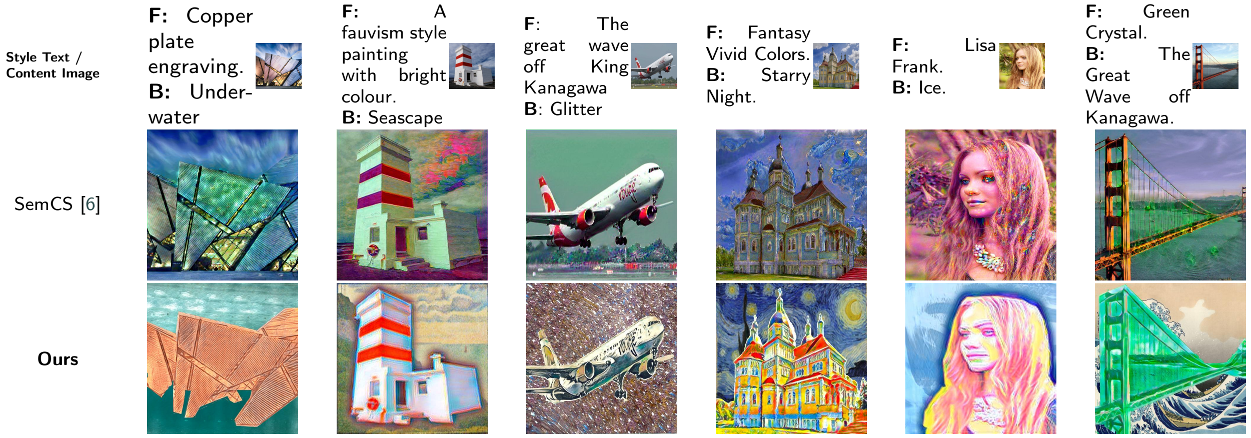
Style Text / Content Image
F: Copper plate engraving. B: Under-water
F: A fauvism style painting with bright colour. B: Seascape
F: The great wave off King Kanagawa B: Glitter
F: Fantasy Vivid Colors. B: Starry Night.
F: Lisa Frank. B: Ice.
F: Green Crystal. B: The Great Wave off Kanagawa.
SemCS [6]
Ours

Figure 5: Text-based IST (double). Distinct text style features are applied to the foreground (**F**) and the background (**B**). In SemCS [17], the absence of a clear boundary between salient objects and surrounding elements results in subpar quality of stylized outputs. In contrast, our method produces outputs with a distinct separation between foreground and background style features. Furthermore, our method provides a superior representation of style features compared to SemCS [6].

Table 3
Stylized Outputs Evaluation, TIST (Single). The table shows that quantitative scores (Nima [18], Contrique [13] (FR, NR)), Clipscore [4] and LPIPS [22] obtained with our method (ObjMST) outperform those of the baseline methods.

Scores	CS [10]	SemCS [6]	LDAST [2]	MMIST [19]	Ours
Nima↑ [18]	5.05	4.81	4.90	4.75	**5.14**
Contrique (FR)↑ [13]	0.28	0.24	0.30	0.29	**0.32**
Contrique (NR)↑ [13]	33.36	29.88	43.36	47.56	**48.14**
Clipscore↑ [4]	0.38	0.40	0.41	0.42	**0.46**
LPIPS↓ [22]	0.45	0.37	0.35	0.40	**0.32**
User Study↑	18.98	14.92	14.15	23.89	**28.06**

4.2. Text-Based IST (Double)

Fig. 5 demonstrates that our method consistently applies distinct style representations to the foreground and background based on the provided style text, avoiding the entanglement of salient and surrounding elements, unlike SemCS [6]. For instance, in column 1, our approach applies copper plate style features to the salient object and "underwater" features to the surrounding elements. In contrast, SemCS [6] fails to maintain a clear boundary, incorrectly stylizing foreground and background.

4.3. Multimodal IST (Single)

As shown in Fig. 6, MMIST (third row) distributes the style over the entire image, affecting both the foreground and background elements. In contrast, ObjMST (fourth row) selectively applies the style only to the object, leaving the surrounding areas unaffected. Unlike MMIST [19], which applies style uniformly using the All-to-All attention mechanism, ObjMST employs Salient-to-Key (S2K) attention mechanism to apply style only to salient objects(only Puppy’s face and body).

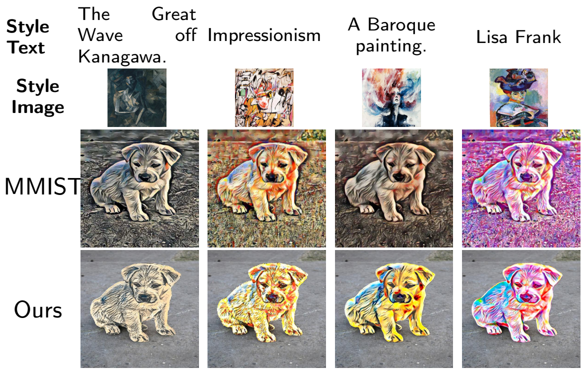

Figure 6: Multimodal IST (single). MMIST (third row) applies style to the entire image, while ObjMST (fourth row) applies it only to the object.

4.4. Ablation Studies

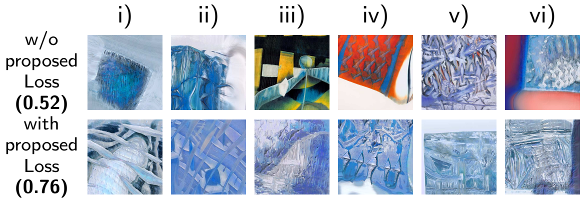

Figure 7: Ablation Study I Style representations (columns i-vi) generated with Masked Directional CLIP loss consistently align with the text "Ice." Misalignment occurs (columns iii) and iv) without this loss, resulting in a lower CLIP score (0.52).

Ablation Study I. This study examines multimodal style representations with (bottom row) and without (top row) proposed loss as shown in columns (i-vi) of Fig. 7. MMIST[19] fails to capture "Ice" features (columns iii and iv), whereas our method (bottom row) consistently aligns style features with "Ice." This is because the proposed loss eliminates the noisy features. A higher Clipscore [4] with the proposed loss function validates the consistency and alignment of the style

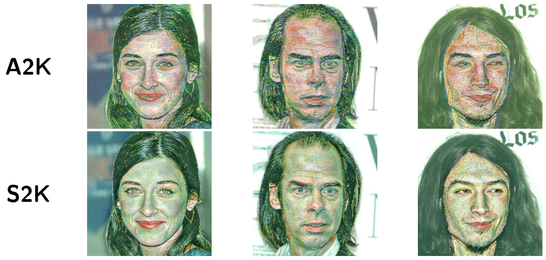

Figure 8: Ablation Study II. This study examines the effects of Salient-to-Key (S2K) mapping and All-to-All Key (A2K) mapping using the style input "A money style painting.

representations with multimodal input.

Ablation Study II. Fig. 8 demonstrates that All-to-All Key (A2K) (top row) blurs the facial features on stylization, while Salient-to-Key (S2K) applies the style features without blurring the facial features and expressions. This difference arises due to the mapping from content to stylized outputs.

5. Conclusion

In this work, we introduced ObjMST, an object-focused multimodal style transfer framework that addresses misalignment and content mismatch issues using a masked directional CLIP loss, Salient-To-Key (S2K) attention mechanism for stylizing salient objects, and image harmonization seamlessly blending the stylized objects with their surroundings. Our results demonstrate the framework's effectiveness in multimodal IST and text-based IST under both single and double-text conditions. As future work, we propose exploring multimodal IST for images with multiple objects, each stylized with distinct and unique representations.

References

[1] Batziou, E., Ioannidis, K., Patras, I., Vrochidis, S., Kompatsiaris, I., 2023. Artistic neural style transfer using cyclegan and FABEMD by adaptive information selection. Pattern Recognit. Lett. 165, 55–62. URL: https://doi.org/10.1016/j.patrec.2022.11.026, doi:10.1016/J.PATREC.2022.11.026.

[2] Fu, T.J., Wang, X.E., Wang, W.Y., 2022. Language-driven artistic style transfer, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI, Springer. pp. 717–734.

[3] Gal, R., Patashnik, O., Maron, H., Bermano, A.H., Chechik, G., Cohen-Or, D., 2022. Stylegan-nada: Clip-guided domain adaptation of image generators. ACM Transactions on Graphics (TOG) 41, 1–13.

[4] Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y., 2021. Clipscore: A reference-free evaluation metric for image captioning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 7514–7528.

[5] Jiang, Y., Zhang, H., Zhang, J., Wang, Y., Lin, Z., Sunkavalli, K., Chen, S., Amirghodsi, S., Kong, S., Wang, Z., 2021. Ssh: A self-supervised framework for image harmonization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4832–4841.

[6] Kamra, C.G., Mastan, I.D., Gupta, D., 2023. Sem-cs: Semantic clipstyler for text-based image style transfer, in: 2023 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 395–399.

[7] Kang, S., Kim, H., Seo, J., 2010. A reliable multidomain model for speech act classification. Pattern Recognit. Lett. 31, 71–74. URL: https://doi.org/10.1016/j.patrec.2009.08.013, doi:10.1016/J.PATREC.2009.08.013.

[8] Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T., 2021. Alias-free generative adversarial networks. Advances in neural information processing systems 34, 852–863.

[9] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R., 2023. Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026.

[10] Kwon, G., Ye, J.C., 2022. Clipstyler: Image style transfer with a single text condition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18062–18071.

[11] Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., Sun, Z., Li, Q., Ding, E., 2021. Adaattn: Revisit attention mechanism in arbitrary neural style transfer, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 6649–6658.

[12] Liu, Z., Cheng, Q., Song, C., Cheng, J., 2023. Cross-scale cascade transformer for multimodal human action recognition. Pattern Recognition Letters 168, 17–23. URL: https://www.sciencedirect.com/science/article/pii/S0167865523000533, doi:https://doi.org/10.1016/j.patrec.2023.02.024.

[13] Madhusudana, P.C., Birkbeck, N., Wang, Y., Adsumilli, B., Bovik, A.C., 2022. Image quality assessment using contrastive learning. IEEE Transactions on Image Processing 31, 4149–4161.

[14] Mao, Q., Wang, C., Wang, M., Wang, S., Chen, R., Jin, L., Ma, S., 2024. Scalable face image coding via stylegan prior: Toward compression for human-machine collaborative vision. IEEE Trans. Image Process. 33, 408–422. URL: https://doi.org/10.1109/TIP.2023.3343912, doi:10.1109/TIP.2023.3343912.

[15] Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D., 2021. Styleclip: Text-driven manipulation of stylegan imagery, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 2085–2094.

[16] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR. pp. 8748–8763.

[17] Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

[18] Talebi, H., Milanfar, P., 2018. Nima: Neural image assessment. IEEE transactions on image processing .

[19] Wang, H., Wu, P., Rosa, K.D., Wang, C., Shrivastava, A., 2024. Multimodality-guided image style transfer using cross-modal gan inversion, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 4976–4985.

[20] Yao, Y., Ren, J., Xie, X., Liu, W., Liu, Y.J., Wang, J., 2019. Attention-aware multi-stroke style transfer, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1467–1475.

[21] Yu, Y., Zhang, D., Wang, S., Ji, Z., Zhang, Z., 2022. Local spatial alignment network for few-shot learning. Neurocomput. 497, 182–190.

[22] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 586–595.

[23] Zhu, M., He, X., Wang, N., Wang, X., Gao, X., 2023. All-to-key attention for arbitrary style transfer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 23109–23119.