HAM: A Training-Free Style Transfer Approach via Heterogeneous Attention Modulation for Diffusion Models

Yeqi He[1,2*] Liang Li[2†] Zhiwen Yang[1,2*] Xichun Sheng[3] Zhidong Zhao[1] Chenggang Yan[1]
[1] Hangzhou Dianzi University [2] Institute of Computing Technology, Chinese Academy of Sciences [3] Macao Polytechnic University
{yeqihe,zhiwen.yang,zhaozd,cgyan}@hdu.edu.cn liang.li@ict.ac.cn p2314922@mpu.edu.mo

Abstract

*Diffusion models have demonstrated remarkable performance in image generation, particularly within the domain of style transfer. Prevailing style transfer approaches typically leverage pre-trained diffusion models' robust feature extraction capabilities alongside external modular control pathways to explicitly impose style guidance signals. However, these methods often fail to capture complex style reference or retain the identity of user-provided content images, thus falling into the trap of style-content balance. Thus, we propose a training-free style transfer approach via **h**eterogeneous **a**ttention **m**odulation (**HAM**) to protect identity information during image/text-guided style reference transfer, thereby addressing the style-content trade-off challenge. Specifically, we first introduces style noise initialization to initialize latent noise for diffusion. Then, during the diffusion process, it innovatively employs HAM for different attention mechanisms, including Global Attention Regulation (GAR) and Local Attention Transplantation (LAT), which better preserving the details of the content image while capturing complex style references. Our approach is validated through a series of qualitative and quantitative experiments, achieving state-of-the-art performance on multiple quantitative metrics.*

(a) Image-guided style transfer results. Compare with StyleID [3].

(b) Text-guided style transfer results. Compare with DiffArtist [14].

Figure 1. Comparative results of style transfer methods: Content image (top-left), style reference (bottom-left), baseline method (top-right), and our HAM (bottom-right). Red boxes denote significant identity retention disparities.

1. Introduction

The development of generative diffusion models has propelled advances in text-to-image generation [1, 7, 25, 27–29], image editing [15, 20, 35], and related fields. Among these, the generative diffusion models have also been applied to style transfer [13, 43], specifically migrating the style references of a given content image to a designated style preset while preserving its identity information. Given the powerful text-to-image generation capabilities of generative diffusion models, they have brought a efficient yet challenging new paradigm to style transfer.

Diffusion-based style transfer approaches typically utilize the inherent generative capabilities of pre-trained diffusion models [7, 25, 29] to achieve style migration for given content images. Some approachs, involves explicit style-content feature decoupling [9, 17, 37, 39], leveraging interpretability to process style features while retaining content features, and incorporates style attributes into pre-trained diffusion models via LoRA [12] or ControlNet [41] for fine-tuning-based style control. While effective for style transfer, these methods are computationally intensive and lack

*This work is done during the intern in VIPL group, ICT, CAS.
†Corresponding author

robustness, as their performance on diverse style references is highly sensitive to the extent of fine-tuning.

To fully utilize the powerful generative capabilities of diffusion models, training-free style transfer methods have been proposed. Representative works like StyleID [3] and DiffArtist [14] achieve stylization by injecting the keys and values from style features, extracted via diffusion model inversion, into the self-attention layers during generation. As theory indicates [10, 30], self-attention features, including queries, keys, and values, collectively encode various semantic and spatial relationships. Consequently, such methods that solely rely on self-attention manipulation results in insufficient style or distorted content, as Fig. 1 demonstrates, resulting in an imbalance between style and content.

In this work, we propose a training-free style transfer approach via **h**eterogeneous **a**ttention **m**odulation for diffusion models (**HAM**), which significantly improves the style-content balance capability of style transfer. Our method utilizes a style/content teacher model obtained from style references and content images, and then uses HAM to combine and share the knowledge from the teacher model to the student generator, thereby achieving style transfer.

Building upon framework process, we propose a **s**tyle-**i**nfused **n**oise **i**nitialization (**SINI**) at timestep T, where the initial latent noise is derived by fusing the inverted noise from the reference style and content images through adaptive instance normalization. To better preserve identity information, the fused initial latent noise is then modulated by the inverted content initial latent noise. Subsequently, in the process of generating and diffusing stylized images, we introduce our HAM, comprising the **g**lobal **a**ttention **r**egulation (**GAR**) and **l**ocal **a**ttention **t**ransplantation (**LAT**), respectively. GAR is a mechanism designed to preserve content and introduce style by exerting macroscopic control over attention injection, thereby maintaining the original spatial and style semantic structure. Specifically, it fuses content and style teacher's attention projections into the specific projections, ensuring their statistical distribution aligns with the corresponding attention projections in the student generator. These specific projections are then reconciled with the ones to further stabilize content/style information. After that, to implement precise style/content control in the absence of text prompts, we introduce LAT through feature operations in cross-modal cross-attention. To better preserve identity information, we inject and weight the query from the content teacher into the query from student generator, while the original key/value pairs are replaced directly with those from the style teacher to ensure effective style guidance. This heterogeneous attention modulation effectively decouples style guidance and content preservation, enabling high-fidelity stylization without compromising structural integrity.

Our GAR and LAT adapt to base model architectures: operating within self-/cross-attention for SD2.1 [29], and joint-/dual-attention for SD3.5 [7], respectively.

The main contributions are summarized as follows:

- We propose a training-free stylized image generation method, HAM, which can achieve high-quality stylized image generation without the need for gradient optimization of style images.
- In our proposed HAM, GAR effectively macroscopically introduces features from the style/content teacher into the student generator, while LAT precisely controls the guidance between style and content. The combined effect improves the quality of the generated stylized images.
- We demonstrate HAM's universal compatibility across DDIM-based (SD2.1) and DiT-based (SD3.5) architectures, achieving state-of-the-art performance on multiple metrics through comprehensive evaluations.

2. Related Work

2.1. Text-Driven Image Generation

With the advancement of deep learning [4, 5, 40], Text-driven image generation [1, 27, 28] has enabled the synthesis of highly realistic and semantically coherent images. Advances in text encoder architectures, exemplified by the SD series [7, 25, 29], have significantly improved text-to-image synthesis through structural refinements and systematic optimization. These technical developments yield quantifiable gains in output visual quality, particularly in enhanced texture detail and resolution fidelity, while reinforcing model robustness in maintaining precise semantic alignment with complex, compositional textual prompts. The progress in generative foundations [24, 33] has also propelled developments in related areas, including: (1) text-guided image editing [21, 22, 30], (2) semantic-aware style transfer [13, 43], and (3) emerging multimodal generative applications beyond static imagery [18, 34, 44–47].

2.2. Image Style Transfer

Style transfer aims to apply a reference style's visual characteristics to a content image while preserving its structural and semantic core, hinging on disentangling content and style representations. Current approaches fall into tuning-based and training-free categories. Tuning-based methods adapt models through parameter updates (e.g., coupling style-specific content or adding lightweight adapters), exemplified by ControlNet [41] training conditional copies, B-LoRA [9] innovating weight optimization, and CSGO [39] using curated adapters. Training-free methods manipulate diffusion mechanisms during inference without parameter changes, altering attention maps/activations to redirect synthesis—e.g., P2P [10] injecting reconstructed cross-attention maps and StyleID [3] fusing features into self-attention layers for zero-shot stylization.

3. Method

In this section, we present our proposed method comprising three core modules: global attention regulation, local attention transplantation, and style-infused noise initialization. Compatible with both DDIM-based SD2.1 [29] and DiT-based SD3.5 [7] architectures, our method is detailed in the following subsections, with technical descriptions primarily based on the SD2.1 framework.

3.1. Preliminaries

Before delving into the specifics of our proposed methodology, we initially provide a comprehensive background on the fundamental techniques that underpin our method.

3.1.1. Latent Diffusion Models

Latent Diffusion Models (LDMs) [29] represent a prominent image generation framework that maps images into the latent space of the Variational Auto-Encoder (VAE) [16] and subsequently leverages the powerful generative capacity of diffusion models to synthesize high-quality images while optimizing computational efficiency. The most representative class of such methods comprises the Stable Diffusion (SD) [7, 25, 29], which utilizes text prompts as input conditions to denoise latent noise for generating high-quality, high-fidelity images. The overall denoising process is formally described by Eq. 1.

$$\mathcal{L}_\theta = \mathbb{E}_{z,t,c,\epsilon\in\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right], \tag{1}$$

where t denotes the current diffusion time step, z_t represents the latent noise vector corresponding to time step t, c signifies the conditioning text prompt, ϵ is the Gaussian-distributed noise sampled from $\mathcal{N}(0,1)$, and $\epsilon_\theta(z_t, t, c)$ denotes the noise component predicted by the model.

3.1.2. Attention Mechanism

To enhance the quality of synthesized images and effectively integrate external conditioning information, Stable Diffusion incorporates multiple groups of self-attention and cross-attention blocks [36], which are typically arranged in complementary pairs. As previously established, our proposed methodology focuses on modulating this dual-attention mechanism, specifically by concurrently targeting both self-attention and cross-attention blocks. The precise mathematical formulations governing these two attention operations are formally expressed in Eq. 2.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2}$$

where Q, K, and V denote the query, key, and value matrices, respectively, with d_k representing the dimensionality of the key vectors. In self-attention, these matrices are all derived from the same input feature map, enabling intra-modality feature integration. Conversely, in cross-attention, Q is typically projected from the visual feature space, while K and V originate from the conditioning context (e.g., text embeddings), facilitating inter-modality information fusion.

3.2. Global Attention Regulation

As substantiated by prior research [10, 30], self-attention projections within Latent Diffusion Model (LDM)-based generative frameworks [7, 29] inherently encapsulate both spatial positional relationships and semantic representations pertinent to content-related information. Consequently, the preservation of content image identity information during style transfer operations emerges as a critical prerequisite for maintaining structural fidelity. To address this fundamental requirement, we propose a novel feature fusion module that strategically leverages complementary information from both content teacher models and style teacher models, as shown in Fig. 2. This module enables global modulation of the self-attention layers within the main branch's stylized generation model, thereby ensuring dual objectives: 1) consistent retention of content identity characteristics throughout the stylization process, and 2) effective incorporation of stylistic attributes derived from either image or textual style references. Specifically, the initial phase of our global attention regulation module focuses on synergistic integration of multi-source feature representations. Given the imperative for training-free feature manipulation, we employ adaptive instance normalization [13], a pioneering work in the style transfer domain, to achieve decoupled feature recombination. This operation fuses content-specific attention projections $(Q^c_{self}, K^c_{self}, V^c_{self})$ from content teacher models with style-specific projections $(Q^s_{self}, K^s_{self}, V^s_{self})$ from style teacher models, generating optimized composite projections $(Q^{cs}_{self}, K^{cs}_{self}, V^{cs}_{self})$. The mathematical formalization of this fusion process, which aligns feature distribution statistics while preserving discriminative attributes, is detailed in Eq. 3.

$$\begin{aligned}
Q^{cs}_{self} &= \sigma\left(Q^s_{self}\right)\cdot\frac{Q^c_{self} - \mu\left(Q^c_{self}\right)}{\sigma\left(Q^c_{self}\right)} + \mu\left(Q^s_{self}\right),\\
K^{cs}_{self} &= \sigma\left(K^s_{self}\right)\cdot\frac{K^c_{self} - \mu\left(K^c_{self}\right)}{\sigma\left(K^c_{self}\right)} + \mu\left(K^s_{self}\right),\\
V^{cs}_{self} &= \sigma\left(V^s_{self}\right)\cdot\frac{V^c_{self} - \mu\left(V^c_{self}\right)}{\sigma\left(V^c_{self}\right)} + \mu\left(V^s_{self}\right),
\end{aligned} \tag{3}$$

where $Q^c_{self}, K^c_{self}, V^c_{self}$ are the content-specific attention projections in the content teacher model, $Q^s_{self}, K^s_{self}, V^s_{self}$ are the style-specific attention projections in the style teacher model, and $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and variance.

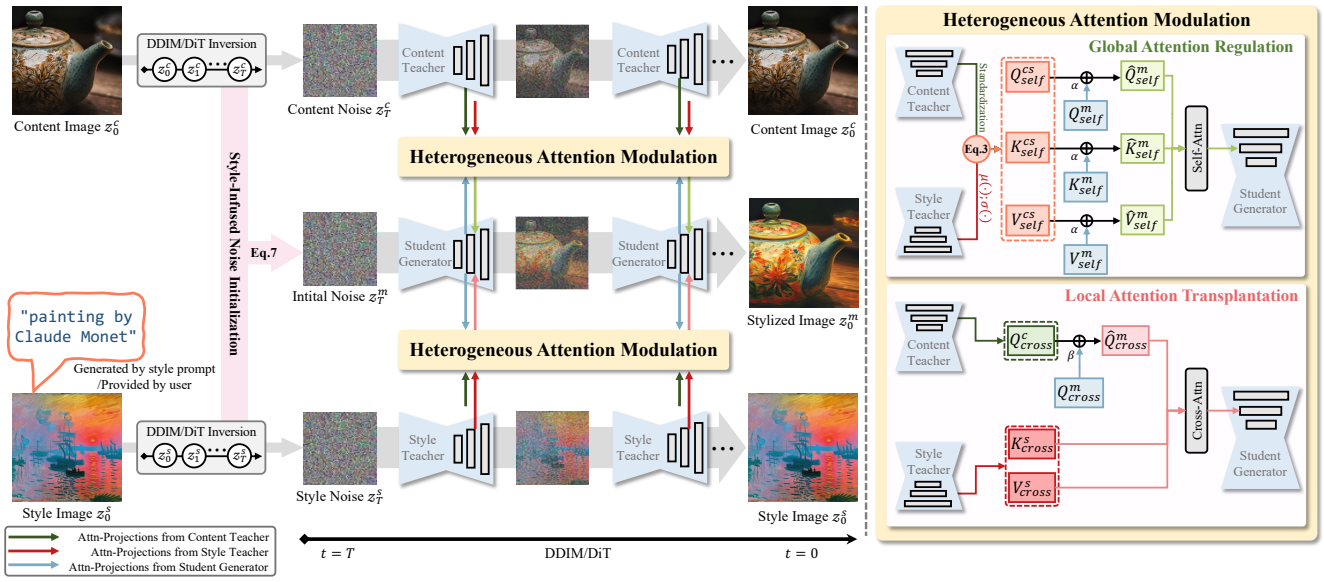

Figure 2. The overall pipeline of our method. Our proposed method consists of three main modules: global attention modulation, local attention transfer, and style injection noise initialization, which act on the self-attention, cross-attention, and noise initialization stages respectively. Through the joint modulation of the three modules, the final stylized image can retain more content identity information and capture and transfer complex style references.

Building upon this foundational alignment, our methodology ensures that the statistical properties of the optimized composite projections $(Q^{cs}_{self}, K^{cs}_{self}, V^{cs}_{self})$ exhibit distributional congruence with the intrinsic self-attention projections $(Q^{m}_{self}, K^{m}_{self}, V^{m}_{self})$ of the main branch. This critical correspondence establishes the theoretical basis for effective global modulation of the self-attention within the stylized generation model, enabling coordinated feature transformation throughout the denoising trajectory. To implement this modulation, we employ a weighted fusion strategy that integrates the optimized composite projections with the main branch's native self-attention representations using a predefined hyperparameter. This controlled combination achieves dual objectives: 1) persistent conservation of content identity information throughout stylization, and 2) regulated incorporation of stylistic attributes from style references (style images or textual descriptions). The mathematical formulation in Eq. 4 utilizes a fixed blending coefficient α to explicitly balance the trade-off between content preservation and style infusion, ensuring deterministic transformations independent of optimization dynamics.

$$\begin{aligned} \hat{Q}^{m}_{self} &= \alpha \cdot Q^{m}_{self} + (1-\alpha) \cdot Q^{cs}_{self}, \\ \hat{K}^{m}_{self} &= \alpha \cdot K^{m}_{self} + (1-\alpha) \cdot K^{cs}_{self}, \\ \hat{V}^{m}_{self} &= \alpha \cdot V^{m}_{self} + (1-\alpha) \cdot V^{cs}_{self}, \end{aligned} \tag{4}$$

where α is a hyperparameter used to control the weight of the fused attention projections.

Subsequently, the resulting modulated self-attention projections $(\hat{Q}^{m}_{self}, \hat{K}^{m}_{self}, \hat{V}^{m}_{self})$ are integrated into the main branch's stylized generation model's self-attention, ensuring that throughout the stylized generation process, the self-attention projections maintain content image identity information while simultaneously incorporating style references from style images/text. The final self-attention expression is shown in Eq. 5.

$$\text{Attention}(\hat{Q}^{m}_{self}, \hat{K}^{m}_{self}, \hat{V}^{m}_{self}), \tag{5}$$

3.3. Local Attention Transplantation

Existing methods for attention injection [3, 8] predominantly operate within the self-attention blocks of Stable Diffusion. However, since self-attention projections inherently encode substantial spatial-semantic structures, replacing key and value matrices in these blocks inevitably compromises content identity preservation. Furthermore, modifications to self-attention necessitate additional distribution alignment mechanisms (e.g., attention temperature scaling [3]) to mitigate discrepancies between query representations from different models and their corresponding key/value pairs.

To circumvent these limitations, we propose a novel paradigm shift: leveraging underutilized cross-attention channels for style transplantation. As illustrated in Fig. 2, our local attention transplantation module strategically employs feature representations extracted from both content and style teacher models. While prior methods [3, 10] manipulate attention maps by substituting key/value projec-

Figure 3. Qualitative comparison with existing text-driven and image-driven SOTA methods. For fair evaluation, all methods use fixed random seeds: text-driven methods apply prompts directly, while image-driven methods generate style references via SD2.1 using identical prompts. Our HAM method better preserves content identity while maintaining style transfer semantics.

tions in either cross-attention or self-attention blocks, our method also targets the cross-attention like them. Concretely, we transplant style-specific key and value projections $(K^s_{cross}, V^s_{cross})$ derived from style teacher models into the main stylization branch, replacing their native counterparts $(K^m_{cross}, V^m_{cross})$ to achieve localized style injection.

Simultaneously, to prevent content identity degradation during diffusion and counteract potential style intrusion from transplanted projections, we implement a content protection mechanism for query representations. This is achieved through weighted fusion between content teacher model's query projections Q^c_{cross} and the main branch's native query projections Q^m_{cross}, ensuring persistent conservation of structural identity. The complete operational formalization is provided in Eq. 6.

$$\begin{aligned} \hat{Q}^m_{cross} &= \beta \cdot Q^m_{cross} + (1-\beta) \cdot Q^c_{cross}, \\ &\text{Attention}(\hat{Q}^m_{cross}, K^s_{cross}, V^s_{cross}), \end{aligned} \tag{6}$$

where β is a hyperparameter that controls the query projection injection weight of the content teacher model.

3.4. Style-Infused Noise Initialization

As comprehensively delineated in Fig. 2, our framework achieves robust style transfer control throughout the diffusion process via the synergistic operation of global attention regulation and local attention transplantation modules. The configuration of initial noise consequently emerges as the pivotal remaining determinant of stylization efficacy, given its fundamental role in establishing the generative trajectory's starting state. While direct transplantation of the content teacher model's initial noise z^c_0 to the main stylization branch represents a conceptually straightforward approach, our quantitative evaluations and qualitative assessments (observed under $\gamma = 1$ settings in Fig. 6 and Tab. 5) demonstrate its inability to achieve meaningful style transfer, primarily attributable to insufficient style integration. Similarly, adaptive instance normalization-based fusion of content noise z^c_0 and style noise z^s_0 yields composite initial noise z^m_0, yet empirical analysis reveals pronounced content identity degradation (observed under $\gamma = 0$ settings in Fig. 6 and Tab. 5). This phenomenon indicates an inherent optimization conflict between style intensity and content fidelity in conventional fusion paradigms.

To resolve this fundamental limitation, we innovate style-infused noise initialization that incorporates a dedicated content residual noise component atop the baseline AdaIN-fused stylized noise. This dual-component architecture explicitly balances stylization intensity against content preservation, enabling precise calibration of their relative contributions throughout the denoising cascade. The complete mathematical implementation is formalized in Eq. 7.

Method	ArtFID↓	FID↓	LPIPS↓	LPIPS-Gray↓	DINO↑	CLIP-I↑	CLIP-T↑	DC↑	CC↑
DDIM(ICLR'21)	31.149	17.939	0.645	0.554	0.278	0.493	0.192	1.524	1.780
ControlNet(ICCV'23)	24.751	13.472	0.710	0.557	0.513	0.583	0.210	1.831	1.916
StyTR^2(CVPR'22)	17.460	10.433	0.527	0.416	0.433	0.487	0.206	1.729	1.794
InstructPix2Pix(CVPR'22)	28.319	17.657	0.518	0.415	0.575	0.620	0.211	1.908	1.963
InstantStyle(arxiv'24)	27.244	15.249	0.677	0.556	0.474	0.604	0.198	1.765	1.921
CSGO(arxiv'24)	27.116	15.207	0.673	0.527	0.482	0.581	0.197	1.775	1.893
StyleID(CVPR'24)	<u>15.161</u>	**8.273**	0.635	0.516	0.544	0.619	0.213	1.873	1.964
STAM(CVPR'25)	16.941	9.269	0.650	0.532	0.531	0.608	<u>0.221</u>	1.869	1.963
AttDistillation(CVPR'25)	16.170	<u>8.926</u>	0.629	0.514	0.541	0.615	0.219	1.878	1.969
DiffArtist(MM'25)	16.174	9.641	<u>0.520</u>	<u>0.413</u>	<u>0.629</u>	<u>0.626</u>	0.220	<u>1.987</u>	<u>1.984</u>
HAM(Ours)	**15.151**	9.244	**0.479**	**0.362**	**0.728**	**0.682**	**0.223**	**2.113**	**2.057**

Table 1. Quantitative comparison with existing text-driven and image-driven SOTA methods. The best results are highlighted in bold, and the second best results are underlined. The color coding indicates relative performance.

$$z_T^m = \gamma \cdot \underbrace{\left[z_T^c - \left(\sigma(z_T^s) \cdot \frac{z_T^c - \mu(z_T^c)}{\sigma(z_T^c)} + \mu(z_T^s) \right) \right]}_{\text{Content Residual Noise}} + \underbrace{\left[\sigma(z_T^s) \cdot \frac{z_T^c - \mu(z_T^c)}{\sigma(z_T^c)} + \mu(z_T^s) \right]}_{\text{Stylized Initial Noise}}, \tag{7}$$

where γ is a hyper-parameter that controls the weight of the content residual noise, z_T^c is the initial noise of the content teacher model, z_T^s is the initial noise of the style teacher model, z_T^m is the stylized initial noise for the main branch, and $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and variance.

4. Experiments

4.1. Experiment Setup

Implementation Details We employ diffusion models based on the SD2.1 and SD3.5 architectures. The denoising process utilizes 50 steps for SD2.1. Images are resized to 512×512 pixels, with SD2.1 hyperparameters configured as $\alpha = 0.75$, $\beta = 0.25$, $\gamma = 0.5$. Experiments execute on a single NVIDIA RTX3090, where SD2.1 inversion requires 4s (50 steps) and stylized image generation completes in 16s (50 steps). The results of our method HAM on SD3.5 are discussed in the supplementary material.

Datasets We conduct experiments on the MS-COCO [19] and the WikiArt [32]. Specifically, 1,000 images from MS-COCO were randomly selected as the test content images. For WikiArt, a collection of images from multiple artists was chosen as the style references to represent the style distribution for FID computation. The final test dataset comprises 1,000 content images, 1,000 corresponding generated stylized images, and the corresponding artists' works from WikiArt as style image references.

Comparison Methods We compare with existing text-driven and image-driven methods: DDIM [31], ControlNet [41], StyTR^2 [6], InstructPix2Pix [2], InstantStyle [37], CSGO [39], StyleID [3], STAM [8], AttDistillation [48] and DiffArtist [14]. For fair text/image-guided comparison, we fix random seeds and use identical prompts: (1) Text-guided: direct prompt input (2) Image-guided: prompts generate style references via SD2.1

Evaluation Metrics We evaluate both traditional metrics FID [11], LPIPS [42], ArtFID [38] and metrics based on DINO [23] and CLIP [26] (DINO, CLIP-I, CLIP-T). For CLIP-T, the input text prompts are style-specific prompts. Our evaluation framework comprises: **(1) Style strength metrics**: **FID** and **CLIP-T** measuring stylization degree; **(2) Content preservation metrics**: **LPIPS**, **DINO** and **CLIP-I** assessing content consistency; **(3) Comprehensive style transfer metric**: **ArtFID** for overall performance. To better comprehensively evaluate stylized images, we introduce two novel composite metrics following ArtFID's computation paradigm: **DC** $= (\text{DINO} + 1) \cdot (\text{CLIP-T} + 1)$ and **CC** $= (\text{CLIP-I} + 1) \cdot (\text{CLIP-T} + 1)$.

4.2. Performance Evaluation

Qualitative Evaluation As shown in Fig. 3, under our fair evaluation setting, our HAM and existing SOTA methods are compared across diverse text-driven and image-driven style transfer tasks. Due to space constraints, qualitative results omit several methods (e.g., DDIM, ControlNet) that perform poorly in quantitative assessments. Visual results indicate that StyTR^2 and InstructPix2Pix struggle to capture complex style patterns, leading to noticeable style leakage and content distortion. InstantStyle and CSGO either inadequately represent style details or suffer from severe style-content leakage, resulting in loss of identity. Although StyleID, STAM, AttDistillation, DiffArtist produce reasonable stylizations, they exhibit a consistent

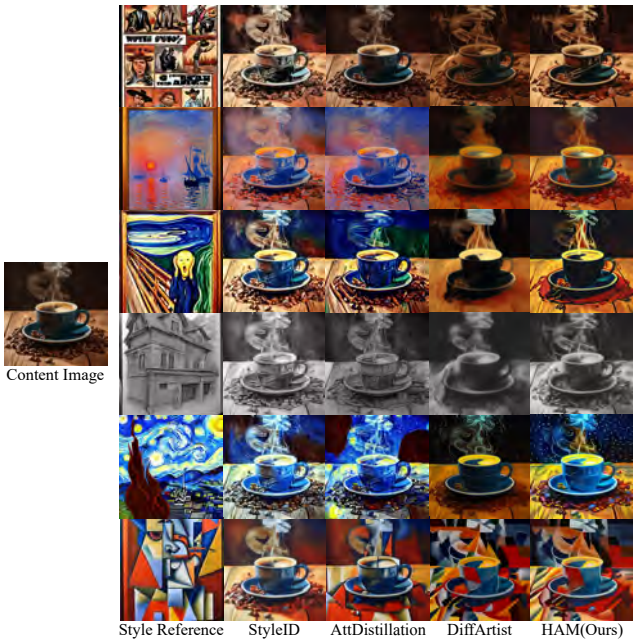

Figure 4. Qualitative results of our method HAM and the SOTA method are presented under different style references for the same content image. It can be observed that our method HAM has significant advantages in both style transfer and identity preservation.

style-content trade-off: either style is well-captured at the cost of content structure, or content is preserved with insufficient style expression. In contrast, HAM accurately captures stylistic attributes under various style-references while maintaining high content fidelity, thereby improving the overall quality of the generated stylized images.

Additionally, as illustrated in Fig. 4, another qualitative experiment is conducted using a single content image under varied style references, with fixed parameters across all methods and no specific adjustments for any styles. The results reveal that existing SOTA methods not only struggle with balancing style and content but also demonstrate limited adaptability when applied to diverse styles using the same content image. In contrast, our HAM shows stronger robustness in transferring complex stylistic information from multiple references to the same content image, while more effectively preserving structural details compared to current SOTA methods. These findings affirm the robustness of our HAM in handling style transfer tasks.

Quantitative Evaluation As shown in Tab. 1, our method HAM achieves optimal (CLIP-T) and near-optimal (FID) performance on two critical style-strength metrics, demonstrating its efficacy in distilling and transferring style reference information. Crucially, the top-ranked CLIP-T performance explicitly confirms HAM's exceptional alignment with textual style semantics across diverse prompts. For

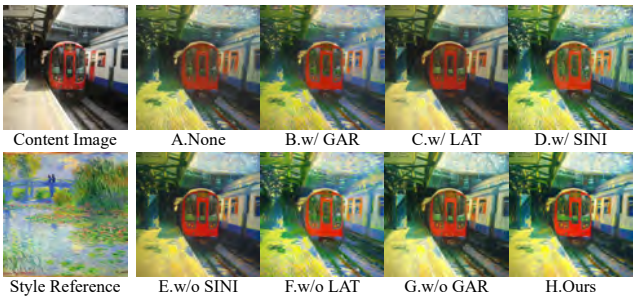

Figure 5. Qualitative ablation study of different modules in our method. The indexes are consistent with those in the quantitative experiments.

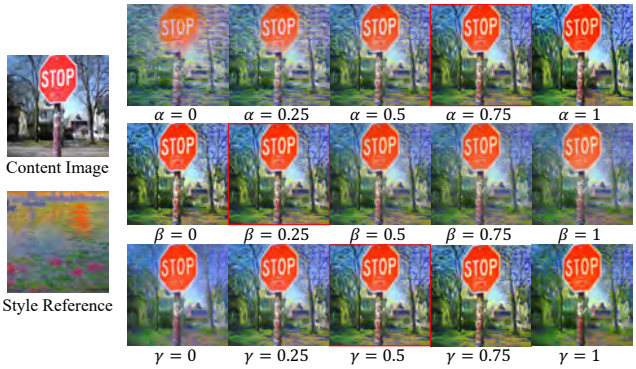

Figure 6. Qualitative ablation study of hyper-parameters in our method. The red boxes represent the hyperparameters we selected.

traditional content preservation metrics (LPIPS and LPIPS-Gray), HAM significantly outperforms all baselines by notable margins, underscoring its absolute advantage in retaining both structural integrity and fine-grained details of content images before and after color removal. This quantitatively reinforces qualitative observations of high content identity preservation. Furthermore, the other content metrics (DINO and CLIP-I) validate HAM's robustness in preserving inter-image structural coherence, global visual consistency, and visual semantic consistency. This is consistent with our qualitative conclusion regarding generation performance. Finally, regarding overall quality metrics (ArtFID, DC, CC), HAM maintains optimal balance between holistic content preservation and precise style semantics, markedly surpassing all competing methods in synthesizing perceptually harmonious outputs.

4.3. Ablation Study

Ablation on different modules As illustrated in Fig. 5 and Tab. 2, ablation studies evaluate three core modules of the HAM method. Both qualitative and quantitative results demonstrate the contributions of our modules. Global attention regulation module improves the CLIP-T score while slightly increasing DINO and CLIP-I metrics. Qualitative

No.	GAR	LAT	SINI	DINO	CLIP-I	CLIP-T	DC	CC
A	✗	✗	✗	0.609	0.626	0.220	1.963	1.984
B	✓	✗	✗	0.618	0.626	0.231	1.993	2.002
C	✗	✓	✗	0.712	0.696	0.193	2.042	2.023
D	✗	✗	✓	0.588	0.624	0.225	1.945	1.989
E	✓	✓	✗	**0.746**	**0.696**	0.202	2.099	2.040
F	✓	✗	✓	0.599	0.627	**0.235**	1.976	2.010
G	✗	✓	✓	0.695	0.679	0.200	2.034	2.015
H	✓	✓	✓	0.728	0.682	0.223	**2.113**	**2.057**

Table 2. Quantitative ablation study of different modules in our method. The indexes are the same as those in the qualitative experiments.

α	DINO↑	CLIP-I↑	CLIP-T↑	DC↑	CC↑
1	0.695	0.679	0.200	2.034	2.015
0.75	**0.728**	**0.682**	0.223	**2.113**	**2.057**
0.5	0.690	0.667	0.225	2.070	2.042
0.25	0.602	0.634	0.226	1.964	2.003
0	0.497	0.602	**0.229**	1.840	1.969

Table 3. Quantitative ablation study of hyper-parameter α in Global Attention Regulation (GAR).

results further illustrate its role in enhancing the stylization intensity and content preservation, confirming its consistency with our design objective. This is also consistent with the performance of quantitative indicators. The local attention transplantation module significantly improves DINO and CLIP-I content preservation metrics but adversely affects CLIP-T. This corresponds to qualitatively observed stronger identity retention and reduced stylization, indicating that integrating the style teacher's key/value requires protecting the main branch's query with the content teacher's query features (further evidenced in the next paragraph for the hyperparameter β). For style-infused noise initialization, a modest CLIP-T improvement is observed with no measurable impact on DINO or CLIP-I. Visually, this module enriches stylistic elements and color diversity in the output. Similarly, the results of quantitative and qualitative experiments on SINI are consistent. Collectively, the modules establish HAM's balance on DC/CC metrics.

Ablation on different hyper-parameters As shown in Fig. 6, ablation experiments assess HAM's key hyperparameters (α, β, γ) qualitatively. For α in the global attention regulation module, as shown in Tab. 3,content preservation metrics (DINO and CLIP-I) achieve optimal performance at $\alpha = 0.75$, degrading at other values, consistent with qualitative observations notably regarding text below STOP signs. Conversely, CLIP-T increases as α decreases, and qualitatively, the generated image tends to be stylized, consistent with qualitative assessments and design intent. DC and CC confirm $\alpha = 0.75$ yields best overall performance. In the local attention transfer module, as shown in Tab. 4,for parameter β, decreasing β can improve the content score (DINO, CLIP-I) while decreasing CLIP-T. Qualitatively, this manifests in the generated image as text and other main subjects gradually becoming blurred, increasing the stylization of the image, consistent with the quantitative results. β weights the query injection, governing content protection versus style transfer, mutually constrained. Based on DC and CC, $\beta = 0.25$ provides optimal balance, validated quantitatively and qualitatively. For γ in the style-infused noise initialization module, as shown in Tab. 5, CLIP-T peaks at $\gamma = 0.5$, while content metrics decline as γ decreases, consistent with enhanced style information at content identity expense. DC and CC indicate $\gamma = 0.5$ offers best stylization outcome, validated quantitatively and qualitatively.

β	DINO↑	CLIP-I↑	CLIP-T↑	DC↑	CC↑
1	0.599	0.627	**0.235**	1.976	2.010
0.75	0.646	0.630	0.227	2.019	1.999
0.5	0.689	0.650	0.226	2.071	2.023
0.25	0.728	0.682	0.223	**2.113**	**2.057**
0	**0.739**	**0.704**	0.201	2.088	2.046

Table 4. Quantitative ablation study of hyper-parameter β in Local Attention Transplantation (LAT).

γ	DINO↑	CLIP-I↑	CLIP-T↑	DC↑	CC↑
1	**0.746**	**0.696**	0.202	2.099	2.040
0.75	0.733	0.689	0.212	2.101	2.048
0.5	0.728	0.682	**0.223**	**2.113**	**2.057**
0.25	0.714	0.678	0.217	2.086	2.042
0	0.708	0.674	0.212	2.070	2.029

Table 5. Quantitative ablation study of hyper-parameter γ in Style-Infused Noise Initialization (SINI).

5. Conclusion and Limitations

We propose HAM, a training-free style transfer framework that addresses the core challenge of content-style balance via heterogeneous attention modulation. Our method begins with SINI, followed by two heterogeneous modulation mechanisms: GAR and LAT. Together, these components work synergistically to achieve superior content-style equilibrium. Extensive experiments demonstrate that HAM outperforms the state-of-the-art methods in both fidelity and stylistic quality. Although our method advances content-style balancing, transferring highly abstract or surrealistic artistic styles remains an open challenge for future work.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (62322211, 62336008), the "Pioneer" and "Leading Goose" R&D Program of Zhejiang Province(2024C01023).

References

[1] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 1, 2

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 6

[3] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8795–8805, 2024. 1, 2, 4, 6

[4] Yiming Cui, Liang Li, Jiehua Zhang, Chenggang Yan, Hongkui Wang, Shuai Wang, Heng Jin, and Li Wu. Stochastic context consistency reasoning for domain adaptive object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1331–1340, 2024. 2

[5] Yiming Cui, Liang Li, Haibing Yin, Yuhan Gao, Yaoqi Sun, and Chenggang Yan. Debiased teacher for day-to-night domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2577–2587, 2025. 2

[6] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11326–11336, 2022. 6

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1, 2, 3

[8] Masud An Nur Islam Fahim, Nazmus Saqib, and Jani Boutellier. Stam: Zero-shot style transfer using diffusion model via attention modulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6333–6343, 2025. 4, 6

[9] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. In *European Conference on Computer Vision*, pages 181–198. Springer, 2024. 1, 2

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023. 2, 3, 4

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1

[13] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 3

[14] Ruixiang Jiang and Chang Wen Chen. Diffartist: Towards structure and appearance controllable image stylization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, page 9598–9607, New York, NY, USA, 2025. Association for Computing Machinery. 1, 2, 6

[15] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6007–6017, 2023. 1

[16] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 3

[17] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation, 2023. 1

[18] Liang Li, Gaoxiang Cong, Yuankai Qi, Zheng-Jun Zha, Qi Wu, Quan Z. Sheng, Qingming Huang, and Ming-Hsuan Yang. Dubbing movies via hierarchical phoneme modeling and acoustic diffusion denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):10361–10377, 2025. 2

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[20] Chang Liu, Xiangtai Li, and Henghui Ding. Referring image editing: Object-level image editing via referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13128–13138, 2024. 1

[21] Henglei Lv, Jiayu Xiao, and Liang Li. Pick-and-draw: Training-free semantic guidance for text-to-image personalization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10535–10543, 2024. 2

[22] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475, 2024. 2

[23] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 6

[24] Yuxin Peng, Zishuo Wang, Geng Li, Xiangtian Zheng, Sibo

Yin, and Hulingxiao He. A survey on fine-grained multimodal large language models. *Authorea Preprints*, 2025. 2

[25] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2, 3

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 6

[27] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 1, 2

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1, 2, 3

[30] Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv preprint arXiv:2406.14555*, 2024. 2, 3

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 6

[32] Wei Ren Tan, Chee Seng Chan, Hernan Aguirre, and Kiyoshi Tanaka. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28(1):394–409, 2019. 6

[33] Jingyi Tang, Li Liang, Beichen Zhang, and Qingming Huang. Lmda: Llm-guided marginal distribution alignment for open-set active learning. *Chinese Journal of Electronics*, 2026. 2

[34] Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, and Qingming Huang. Smart: Syntax-calibrated multi-aspect relation transformer for change captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4926–4943, 2024. 2

[35] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali 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 (CVPR)*, pages 1921–1930, 2023. 1

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[37] Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 1, 6

[38] Matthias Wright and Björn Ommer. Artfid: Quantitative evaluation of neural style transfer. In *DAGM German Conference on Pattern Recognition*, pages 560–576. Springer, 2022. 6

[39] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*, 2024. 1, 2, 6

[40] Beichen Zhang, Liang Li, Shuhui Wang, Shaofei Cai, Zheng-Jun Zha, Qi Tian, and Qingming Huang. Inductive state-relabeling adversarial active learning with heuristic clique rescaling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9780–9796, 2024. 2

[41] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 1, 2, 6

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[43] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10146–10156, 2023. 1, 2

[44] Zhedong Zhang, Liang Li, Gaoxiang Cong, Haibing Yin, Yuhan Gao, Chenggang Yan, Anton van den Hengel, and Yuankai Qi. From speaker to dubber: Movie dubbing with prosody and duration consistency learning. In *Proceedings of the 32nd ACM International Conference on Multimedia, 2024*, pages 7523–7532, 2024. 2

[45] Zhedong Zhang, Liang Li, Chenggang Yan, Chunshan Liu, Anton Van Den Hengel, and Yuankai Qi. Prosody-enhanced acoustic pre-training and acoustic-disentangled prosody adapting for movie dubbing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 172–182, 2025.

[46] Zhiqian Zhao, Liang Li, Jiehua Zhang, Yaoqi Sun, Xichun Sheng, Haibing Yin, and Shaowei Jiang. Heterogeneous prompt-guided entity inferring and distilling for scene-text aware cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10537–10545, 2025.

[47] Zhiqian Zhao, Liang Li, Lei Shen, Xichun Sheng, Yaoqi Sun, Fang Kang, and Chenggang Yan. Temporal calibrating and distilling for scene-text aware text-video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13323–13331, 2026. 2

[48] Yang Zhou, Xu Gao, Zichong Chen, and Hui Huang. Attention distillation: A unified approach to visual characteristics transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18270–18280, 2025. 6