ArtAdapter: Text-to-Image Style Transfer using Multi-Level Style Encoder and Explicit Adaptation

Dar-Yen Chen Hamish Tennent Ching-Wen Hsu
PicCollage, Taiwan
{daryen.chen, hamish.tennent, gina.hsu}@cardinalblue.com
https://cardinalblue.github.io/artadapter.github.io/

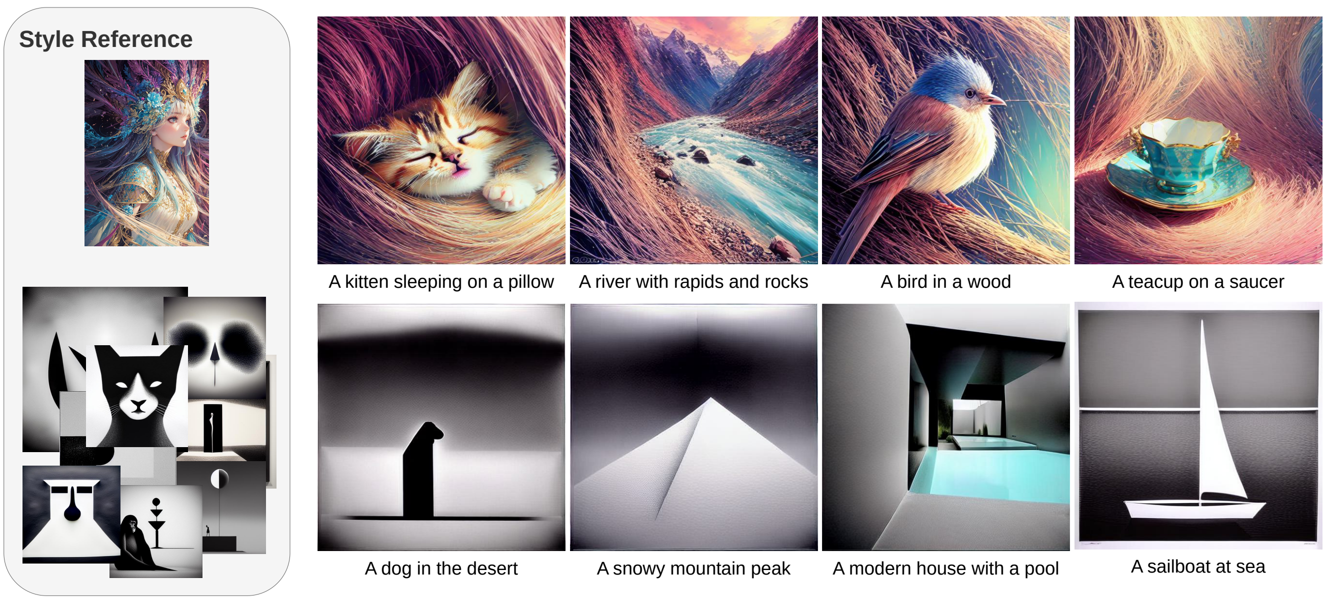

Figure 1. Our framework is capable of capturing faithful style representation, from low-level delicate texture to high-level minimalism composition, in either single or multiple style references, closely adhering to the textual prompts.

Abstract

This work introduces ArtAdapter, a transformative text-to-image (T2I) style transfer framework that transcends traditional limitations of color, brushstrokes, and object shape, capturing high-level style elements such as composition and distinctive artistic expression. The integration of a multi-level style encoder with our proposed explicit adaptation mechanism enables ArtAdapter to achieve unprecedented fidelity in style transfer, ensuring close alignment with textual descriptions. Additionally, the incorporation of an Auxiliary Content Adapter (ACA) effectively separates content from style, alleviating the borrowing of content from style references. Moreover, our novel fast finetuning approach could further enhance zero-shot style representation while mitigating the risk of overfitting. Comprehensive evaluations confirm that ArtAdapter surpasses current state-of-the-art methods.

1. Introduction

Bridging the realms of artificial intelligence and artistic creativity, text-to-image (T2I) style transfer [13, 24, 25] stands out as a captivating domain that masterfully transforms textual descriptions into visually rich and stylistic representations. The core challenge lies in not just generating text-resonant images but infusing them with artistic depth and nuance, spanning from delicate brushstrokes to bold compositional elements—thereby capturing the essence of artistic vision. Conventional arbitrary style transfer (AST) methods [4, 7–9, 15, 21, 28, 54] typically struggle beyond low-level features such as medium and colors, failing to grasp the more sophisticated realms of artistic expression. Diffusion approaches [25, 55], including Textual Inversion [14], DreamBooth [39], and Low-Rank Adaptation (LoRA) [2, 20], have shown potential in style representation. Yet, these methods are hindered by laborious finetuning processes and a tendency towards overfitting, leading to outputs overly influenced by the style references' content at the expense of the textual context. Especially when dealing with the unique nature of artworks, these methods face difficulties in style transfer from a single reference.

Addressing these challenges, we present ArtAdapter, an innovative T2I style transfer framework utilizing style embeddings derived from a multi-level style encoder. These style embeddings, representing various layers of artistic attributes, then intricately interplay with textual context within the text encoder. The style information is refined further through the novel Explicit Adaptation mechanism, situated within the cross-attention layers of the diffusion model. Here, the Explicit Adaptation focuses explicitly on the style pathway, while leaving the text pathway frozen. This ensures that the artistic elements, from foundational textures to the distinctive expression of the artworks, are faithfully and precisely represented in the generated images without compromising the textual generalizability. Moreover, our approach incorporates the Auxiliary Content Adapter (ACA) during the training phase. The ACA plays a vital role by offering weak content guidance, aiding in the separation of content structure from style references. This ensures that the final images are not overwhelmed by the content of the style references, maintaining a clear representation of style elements and the narrative intent of the text prompts. Furthermore, we introduce a fast finetuning method, which further refines the model's ability to capture nuanced style details. This finetuning is effective for both single- and multi-image style references, where it achieves detailed style representation with minimal steps, greatly reducing the time and computational resources typically required. This innovation addresses the challenge of overfitting and speeds up the adaptation process. A notable feature of ArtAdapter is its adeptness at style mixing. Leveraging the multi-level style encoder, ArtAdapter can blend styles from various references, extracting and combining distinct stylistic elements at different levels. This enables the production of images that blend a diverse array of artistic influences, thereby offering remarkable flexibility and creativity to style transfer.

The principal contributions of our work are summarized as follows: (i) A groundbreaking T2I style transfer model, ArtAdapter, that harnesses a multi-level style encoder and the Explicit Adaptation mechanism to capture diverse levels of style representation, and ensures a subtle balance between style and textual semantics; (ii) The Auxiliary Content Adapter (ACA) separates content structure from style references, addressing the issue of content dominance from style references. (iii) A fast finetuning approach that enables rapid and effective style adaptation; (iv) ArtAdapter allows for style mixing across different hierarchical levels, enriching the creative potential of T2I style transfer.

2. Related Work

2.1. Text-to-Image Synthesis

Text-to-image (T2I) synthesis has evolved remarkably, with early methods primarily leveraging Generative Adversarial Networks (GANs) [16]. A paradigm shift was marked by the introduction of discrete Variational Autoencoders (VAEs) [12, 36] and autoregressive transformers [46], leading to a more stable training process and high-quality image generation as demonstrated by Esser *et al*. [12] and the DALL-E [34]. Subsequent advancements with diffusion models [19] further refined the generation with a gradual denoising process, offering enhanced control and fidelity in outputs, as showcased by DALL-E2 [35], Imagen [41], and Stable Diffusion [37]. Recent developments [50, 51, 56] have extended these models' capabilities in various context, with innovations like ControlNet [53] and T2I adapters [31] for spatial semantics, enriching the conditioning options well beyond textual inputs.

2.2. T2I Personalization

In T2I personalization, the central goal is to tailor pretrained models to a specific concept using a collection of reference images. Innovations like Textual Inversion [14] and DreamBooth [39] pioneered this direction, learning text embeddings and optimizing diffusion backbone respectively. The quest for efficiency has led to the development of Low-Rank Adaptation (LoRA) [2, 20], which reduces parameters by decomposing residual weights into two low-rank estimated matrices. SVDiff [17] and Lightweight DreamBooth [40] further refine this process using Singular Value Decomposition (SVD) and orthogonal incomplete basis within LoRA weight-space respectively. Perfusion [45] incorporates Rank-One Model Editing (ROME) [30], exemplifies targeted model edits aligned with conceptual directions. Meanwhile, HyperDreamBooth [40] enables rapid adaptation to new concepts through hypernetwork-initialized rank-1 residuals. Beyond finetuning, various approaches [43, 49] deploy task-specific components for zero-shot adaptation to novel concepts, including Taming Encoder [23] and IP-Adapter [51] for content semantics. Interestingly, T2I style transfer aligns closely with style personalization, treating artistic style as a unique conceptual entity. Our work contributes to this evolving landscape by developing a novel zero-shot T2I style transfer framework whose performance can be further improved through fast finetuning.

2.3. Style Transfer

Traditionally, Style Transfer [4, 8, 28] has depended on extracting content from target images and matching style via second-order statistics [15, 21]. However, the biases [7, 9] inherent in statistics cripple the representations. IEST [7] and CAST [54], address this issue by introducing contrastive learning. The advent of vision transformers [11] has facilitated the capture of long-range features, as seen in StyleFormer *et al*. [48] and StyTr^2 [9]. ArtFusion [6] presents an integration of diffusion models [19] into style transfer, bypassing reliance on statistical matching. Following the booming of T2I models, InST [55] exemplifies the growing interest in T2I style transfer, with semantics provided by textual prompt, learning artistic features and guiding the generating with the CLIP image encoder [33]. T2I-Style-Adapter [31] aligns style representation by harnessing cross-attention layers to process CLIP [33] style image embeddings. Meanwhile, FreeDoM [52] employs time-dependent energy guidance, where the similarity between style references and generated samples is measured using the distance of the Gram matrix. Nonetheless, T2I style transfer confronts challenges such as content from style references overshadowing the textual context [47] and obvious artefacts. Addressing this, our framework introduces an auxiliary component to foster content-style disentanglement, mitigating the influence of content semantics from style references.

3. Approach

Preliminary on T2I Diffusion Models. A Text-to-Image (T2I) diffusion model comprises two foundation components: the diffusion backbone ϵ_θ and the text encoder c_θ. The diffusion backbone meticulously manages a process of adding and removing noise. The inference process begins with a distribution of random noise, which is gradually denoised step by step according to the textual prompt to form a coherent image. The textual prompt, denoted as y, first undergoes tokenization and index-based lookup, linking it to text embeddings F_{txt}, a sequence of vectors. These embeddings are further refined by the text encoder, which contextualizes the information, producing enriched text embeddings E_{txt} that encapsulate the textual description's meaning, intent, and subtleties. Typically, T2I diffusion models employ cross-attention layers to access and utilize the semantic information contained in E_{txt}. The T2I diffusion models' objective is defined by the equation [10]:

$$\mathcal{L} = \mathbb{E}_{x,y,\epsilon,t}\Big[||\epsilon_\theta(x_t, t, c_\theta(F_{txt})) - \epsilon||_2^2\Big], \tag{1}$$

where x is the image, and ϵ represents the noise component in the corrupted x_t at timestep t. This equation aims to estimate the noise involved in the diffusion process, which is crucial to the gradual denoising process, aligning samples with the text description.

Overall Framework. In Figure 2, we present the architecture of our ArtAdapter, utilizing a pretrained diffusion model as its backbone. Our method instructs the diffusion model in articulating diverse styles through learnable style embeddings. For encoding the style reference x_{sty} into

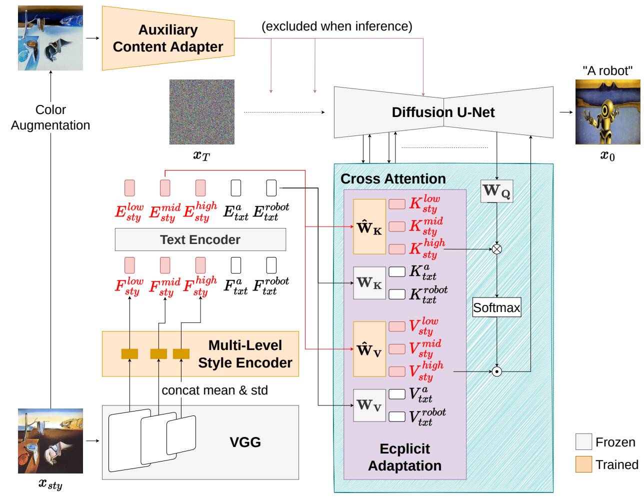

Figure 2. **Architecture of ArtAdapter.** Style embeddings, extracted through a cascade of a pretrained VGG [27] followed by the multi-level style encoder, interact with text embeddings in the text encoder. In the cross-attention layers, the Explicit Adaptation exclusively optimizes the style-related projection to align outputs with style references. The Auxiliary Content Adapter provides weak content guidance during training, helping disentangle the content structure in the style reference. Our approach faithfully captures the style features without content semantics.

style embeddings F_{sty}, ArtAdapter employs a multi-level style encoder to process activations of pretrained VGG network [27]. Subsequently, these style embeddings are concatenated with text embeddings F_{txt}, and processed through the text encoder to obtain the style encodings E_{sty}. To allow the diffusion backbone to adapt to these style concepts, we introduce the Explicit Adaptation mechanism within the cross-attention layers. This mechanism exclusively adapts to style encodings while preserving the integrity of text encoding. This ensures the robust generalization of the pretrained diffusion model while enabling precise adaptation to complex style nuances. Furthermore, our framework incorporates the Auxiliary Content Adapter (ACA) - a key component during training, that is excluded when inference. ACA provides rough content structure information to the UNet backbone [38], helping eliminate the influence of the content semantics in the style reference. Conclusively, our fast finetuning method is designed to capture more nuanced style characteristics, applicable to both individual and collective style references.

3.1. Multi-Level Style Encoder

In the realm of style transfer, second-order statistics of VGG [27] activations, specifically the mean and standard deviation, have been crucial for their effectiveness in achieving style similarity [15, 22]. These statistics lack overt content information, enabling more unadulterated style features. A core of our ArtAdapter lies in the extraction of multi-level style features, amplifying the expressiveness and interpretability of the style transfer. Specifically, we select VGG activations - `relu3_3`, `relu4_3`, and `relu5_3` - to represent low-, mid-, and high-level features, respectively. Once we concatenate these activations' mean and standard deviation [6, 21], low-, mid-, and high-leve style embeddings $F_{sty}^{low}, F_{sty}^{mid}, F_{sty}^{high}$ are extracted using distinct MLPs. These style embeddings interact with the textual context in the text encoder and infuse the nuanced style information into the diffusion model.

Intuitively, this mechanism is akin to captioning the style reference with pseudo-words, resulting in stylized prompts in the form of "$[style_{low}]$ $[style_{mid}]$ $[style_{high}]$ a robot", where $[style_{low/mid/high}]$ serve as learnable descriptors for styles at each level. Notably, our implementation generates 3 embeddings per level, culminating in a comprehensive style embedding F_{sty} of length 9. These multi-level style embeddings provide rich style context, significantly enhancing the style fidelity.

3.2. Explicit Adaptation

Significant advancements in adapting pretrained diffusion models to new contexts, notably with efficient methods like LoRA [20], have demonstrated effective adaptation with fewer parameters. To enhance the integration of style within the diffusion backbone, we introduce Explicit Adaptation, a novel adaptation mechanism different from pursuing minimal parameters. Within the cross-attention layers, the Explicit Adaptation mechanism is meticulously applied to the key and value projections of the style encodings E_{sty}, while leaving the pathways of the text embeddings E_{txt} frozen. The output h of K- or V-projection in a cross-attention layer can be represented as:

$$h = W\{E_{sty}, E_{txt}\} + \alpha \Delta W\{E_{sty}, 0\} \tag{2}$$

where W represents the original weights, which are frozen to maintain existing knowledge, α is a learnable scale, and ΔW is the residual weight focusing on the style encodings.

The term "Explicit" signifies a direct and intentional approach, not only in the operational dynamics of the adaptation but also in the model's learning objectives. This deliberate demarcation ensures the adaptation process is sharply concentrated on the incorporation of style nuances hidden in the style encodings. This innovation enables our ArtAdapter to refine style representations without compromising the established robust linguistic knowledge base. The merits of Explicit Adaptation, particularly in its role in the precise acquisition of fine-grained style features, will be further dissected and validated in Section 4.4.

3.3. Auxiliary Content Adapter

One of the major challenges in T2I style transfer is the unintentional overpowering of the style reference's content within the generated images, eclipsing the textual semantics. Even conditioning on VGG [27] statistics features without content structure, due to certain highly recurrent patterns in the dataset, like human faces [6], the model might become prone to overfitting specific statistical features of content, leading it to transfer content patterns from the style reference. To address this challenge, we introduce the Auxiliary Content Adapter (ACA) into our framework, a variant of the T2I adapter [31]. It takes the image x as input, providing the diffusion backbone with essential content cues. This ensures that the style components in ArtAdapter, including the multi-level style adapter and the Explicit Adaptation, concentrate solely on capturing pure style features, avoiding any unwanted diversion towards mimicking content semantics from the style reference.

To calibrate the influence of ACA, its features are infused within the deepest input block of the UNet backbone [38], and positioned to impact only the initial denoising stages (20% in this work), where the model delineates only the rudimentary content outlines. Moreover, we subject the ACA's input to a series of random color augmentation, including color jitter, greyscale, inversion, and polarization, to prevent the model from learning the style's color palette using ACA — reserving color learning for the style encoder. These approaches confine the ACA's function to offering only the rough content structure, devoid of intricate details, thus paving the way for a more nuanced style representation.

It's imperative to underscore that ACA's role is confined to the training phase. In an ideal training process, the ACA and textual prompts, provide the content structure and semantics, respectively. During inference, ACA and style references' captions are excluded, allowing our ArtAdapter to draw exclusively from the learned style features. This ensures that the generated images present only the desired style characteristics without content from style references.

3.4. Fast Finetuning and Multiple Style References

Building on the Explicit Adaptation mechanism, we introduce a fast finetuning method to enhance the precision and subtlety of zero-shot style representation. The crux of this finetuning lies in the following equation:

$$h = W\{E_{sty}, E_{txt}\} + \alpha \Delta W\{E_{sty}, 0\} + \{\Delta h_{sty}, 0\} \tag{3}$$

Δh_{sty} is the residual vector that directly adjusts the style-related activations, refining the model's ability to mirror the characteristics of the style references. During finetuning, all components besides the vector residuals remain frozen, ensuring an efficient adaptation process focused on the enhancement of style representation. The principle of explicit adaptation mechanism ensures that the textual context pathways remain intact, offering a safeguard against

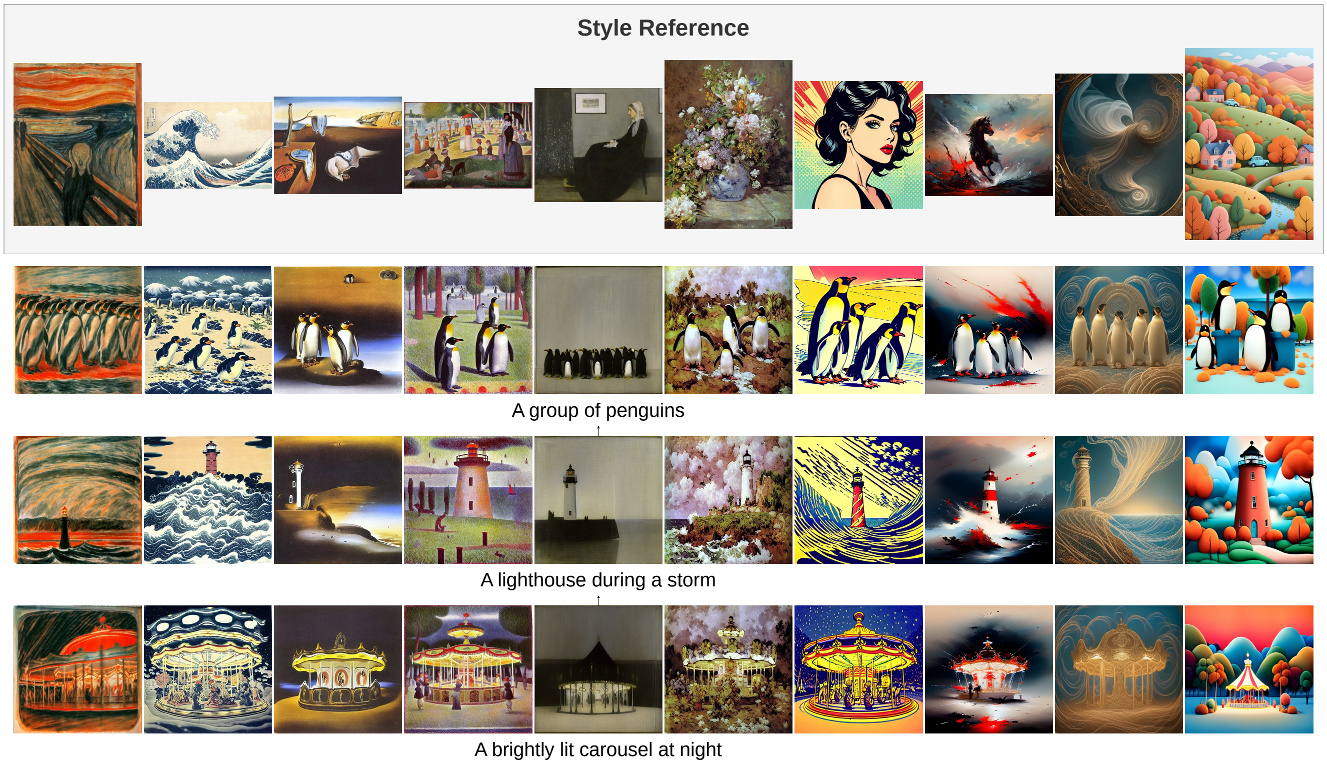

Figure 3. **Qualitative results.** This collection of images exhibits the ArtAdapter's capability to present faithful style representation across diverse artworks without compromising on semantics, showcasing the versatility and deep understanding of artistic and textual contexts.

overfitting—a common challenge in single-image personalization. This finetuning is characterized by its efficiency. With only minimal extra parameters from the rank-1 residual vector, and a focus on style operations, the process is expedited to dozens of steps, typically completed within minutes. Notably, this approach can be easily extended to multi-reference style transfer. By averaging the style embeddings from all references, the model can optimize the vector residuals to reflect the aggregate style.

3.5. Style Mixing

Our innovative style mixing approach leverages the multi-level style encoder to expand the T2I style transfer's horizons. With a trio of style reference images, we extract their respective style embeddings at low, mid, and high levels. These are subsequently constructed into compositional style embeddings by concatenation. Such a fusion of embeddings enables ArtAdapter to synthesize a new, mixed style that simultaneously exhibits features from the individual styles at corresponding hierarchical levels. Furthermore, this method of style mixing can easily extend to styles that have undergone finetuning, by applying the residual vectors to the corresponding entries.

4. Experiments

Data. For training, we harness the LAION AESTHETICS [42] and WikiArt [32] datasets. The WikiArt dataset has been captioned using BLIP-2 [26]. For testing, we adopt a test dataset comprising 35 prompts, including some derived from Wang *et al*. [47]. We collect 10 styles for both single- and multi-reference style transfer separately, culminating in a total of 700 image pairs for the test dataset.

Evaluation Metrics. We employ the CLIP [33] metrics for objective assessment, including similarity for text and style. The aesthetics score [1] functions as an indicator, gauging the aesthetic appeal of the generated images. Additionally, we undertake a user study involving 212 participants utilizing a 7-point scale, with 5 tuples per user, to assess subjective text and style accuracy, alongside the overall quality.

Implementation Details. This work leverages the SD V1.5 [37] as the backbone. We employ an AdamW optimizer [29] with a batch size of 16, a learning rate of 1e-4 for the multi-level style encoder and the ACA, and a reduced learning rate of 1e-7 for Explicit Adaptation residual weights. Specifically, we apply LoRA [2, 20] to reduce the parameters. During the finetuning, we optimize the vector residual using a learning rate of 0.02 over 25 iterations, taking $\sim$1 minute. All experiments are undertaken on a single Nvidia

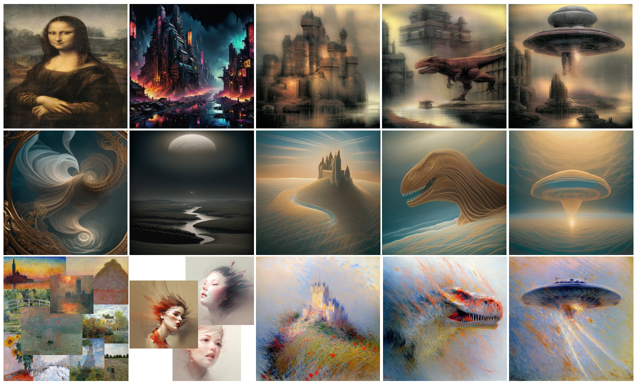

Figure 4. **Illustration of Style Mixing.** The seamless integrations of the two styles reflect the distinct contributions of style features from different hierarchical levels to the final images and demonstrate the remarkable flexibility of ArtAdapter.

Type	Approach	Text ↑	Style ↑	Aesth ↑
single ref	CAST	0.297	0.583	5.410
	StyTr^2	0.299	0.577	5.308
	T2I-Style-Adapter	0.181	0.822	5.705
	InST	0.235	0.660	5.255
	FreeDoM	0.262	0.634	5.059
	Ours (zero-shot)	0.269	0.656	5.532
	Ours	**0.255**	**0.707**	**5.601**
multi ref	TI	0.265	0.678	5.712
	LoRA	0.254	0.678	5.701
	Ours	**0.258**	**0.680**	**5.610**

Table 1. **Quantitative Comparison.** ArtAdapter demonstrates superior balances between text and style similarity, as well as competitive aesthetics score, even with zero-shot.

Model	Text ↑	Style ↑	Quality ↑
TI	4.49	4.35	4.38
LoRA	3.91	4.08	4.18
Ours	**4.74**	**4.76**	**4.43**

Table 2. **Perceptual Comparison.** ArtAdapter exhibits preeminence in all subjective metrics over TI [14] and LoRA [2, 20], highlighting its effectiveness in T2I multi-reference style transfer based on user preferences.

GTX 4090 GPU. For inference, we set the sampling [44] steps to 50, with the classifier-free guidance scale [18] being fixed at 9. Unless otherwise specified, all reported results were obtained with finetuning.

4.1. Qualitative Evaluation

The ArtAdapter's prowess is showcased in Figure 1, where it captures nuanced styles from various either single or multiple style references, while faithfully reflecting the textual prompts. This is particularly evident in the top row of Figure 1, where a clear facial structure is depicted in the style reference, a challenging scenario often encountered in T2I style transfer. Remarkably, our model adeptly reproduces the ornate texture and aesthetic of the style reference without erroneously incorporating the facial structure into the final results. The bottom row of Figure 1 illustrates ArtAdapter's capability to go beyond mere low-level style representation. Here, we observe not only the replication of black-and-white tones and geometric patterns but, more crucially, the high-level abstract composition that resonates with the style reference collection. This demonstrates the multi-level style encoder and Explicit Adaptation mechanism's collaborative efficacy in rendering styles that are both authentic and expressive.

Figure 3 provides further insights into ArtAdapter's capacity, showcasing side-by-side comparisons of various styles within the same textual context. From the precise transfer of textures and tones to the emulation of geometric shapes and the overall composition, ArtAdapter showcases a deep understanding of the distinctive artistic concepts inherent in the style references. ArtAdapter effectively circumvents borrowing content from style references, creatively reinterpreting artistic elements to yield visually compelling art that remains faithful to the textual descriptions. This attests to the efficacy of the explicit mechanisms in our approach, which align the subject matter with the style, while preventing overfitting that commonly happens in single-image personalization finetuning.

In summary, our method facilitates T2I style transfer, advancing toward a more genuine and precise representation of styles. It encapsulates the essence of the artistic intent, offering a substantial leap.

4.2. Style Mixing

Figure 4 demonstrates ArtAdapter's proficiency in seamlessly mixing two styles within one image. Achieving this fusion involves applying distinct styles to different hierarchical levels: one affects low- and mid-level features, while the other shapes high-level attributes. This mixing reveals the significant roles of different levels in shaping the final image. The low and mid-levels primarily dictate tone and texture, as evidenced by the haziness of the "Mona Lisa" and Monet's brushstroke texture. Conversely, high-level features dictate the overall compositional structure and artistic expression, as exemplified by the simplistic style in the second row. This multi-level approach underscores our model's flexibility and interpretability in style transfer, facilitating a creative combination of artistic elements.

4.3. Comparison with State-of-the-art Methods

Single Style Reference. In Figure 5, our ArtAdapter is compared with conventional AST models, such as CAST

Figure 5. **Qualitative comparison on single style reference.** Our results showcase ArtAdapter's superior style alignment over other approaches [5, 9, 54, 55]. Note that the SD [37] column works as content target images for conventional AST models [9, 54].

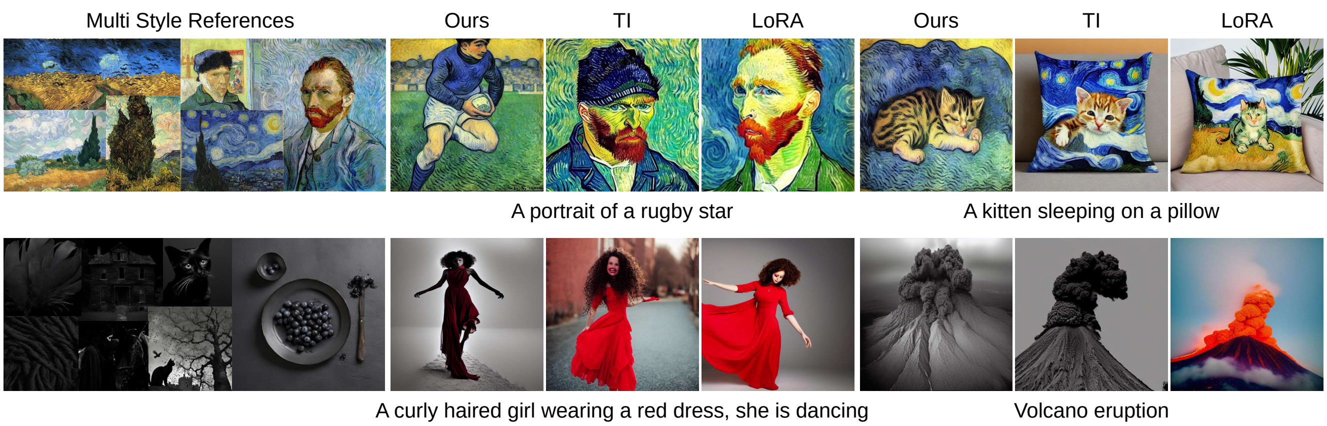

Figure 6. **Qualitative comparison on multiple style references.** Compared to TI [14] and LoRA [2, 20], our ArtAdapter provides greater consistency and coherence, while preventing content borrowed from the style references to the results.

[54] and StyTr^2 [9], the diffusion-based T2I-Style-Adapter [31], InST [55], FreeDoM [52] and the advanced DALL-E3 [5] in single-reference T2I style transfer. Utilizing SD [37] as for T2I generation, stylized by AST models, we have the T2I style transfer results of conventional AST approaches [9, 54]. These models typically struggle to transcend basic color transfer, struggling with the conveyance of high-level style attributes, particularly abstract geometric configurations, and holistic style elements. As diffusion-based approaches, T2I-Style-Adapter [31], while adept at representing the style, often discards the textual context, resulting in a loss of controllability. InST [55] tends to produce results with unnatural style representation and unintended content borrowings from style references; FreeDoM [52], which relies on energy guidance, often introduces noticeable artefacts into the stylized images. DALL-E3 [5], leveraging ChatGPT [3] to generate style descriptions, struggles in the actual reproduction of style features, revealing a bottleneck in prompt engineering.

Distinctly, our zero-shot ArtAdapter captures the original style's brushwork, texture, and overall aesthetic with exceptional fidelity, as evidenced by the quantitative results in Table 1. This fidelity is further enhanced by the finetuning, leading to an alignment closely mirroring the original artworks. The quantitative analysis, as shown in Table 1, highlights our style similarity score of 0.707 and an aesthetics score of 5.601, surpassing the benchmarks, except for T2I-Style-Adapter. While T2I-Style-Adapter edges out in style and aesthetics, the disproportionately low text similarity reveals a significant shortcoming in its T2I style transfer capability. In contrast, our ArtAdapter excels, maintaining high style and aesthetics scores with only a minimal trade-

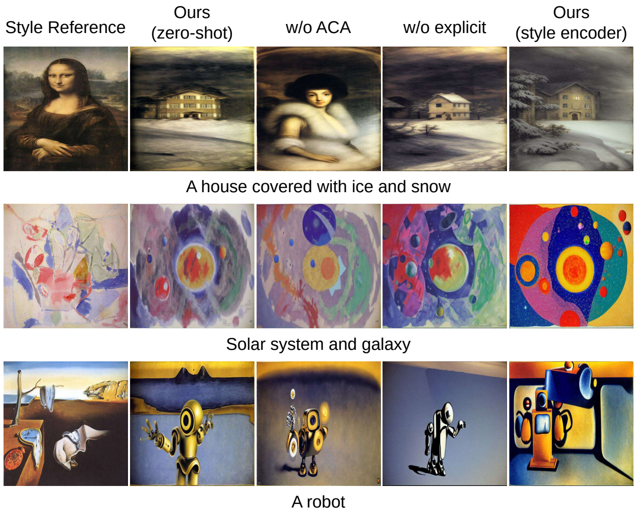

Figure 7. **Ablation research.** These comparisons highlight the importance of each mechanism in our framework.

off in text fidelity, which is 0.255.

Multiple Style Reference. In assessing multi-reference T2I style transfer, our ArtAdapter is juxtaposed with Textual Inversion (TI) [14] and Low-Rank Adaptation (LoRA) [20], detailed in Figure 6 and Table 1. Despite the successes of TI [14] and LoRA [20] in capturing subtle style nuances, they exhibit overfitting, as evidenced by their dominance of style reference content, such as the face of Van Gogh, leading to discrepancies with the textual prompts. Instances such as the top-right corner of the sample manifest the incoherence in their style transfer, rooted in localized stylization problems. Moreover, varying degrees of style assimilation within a single style reference in the bottom row reveals a lack of consistency. These highlight the instability of TI [14] and LoRA [20] in style transfer. Our ArtAdapter excels in rendering outputs that accurately embody style features from the collection and eliminate content borrowing, with consistent representation in diverse textual contexts. It nicely preserves the inherently strong generality of SD [37]. Quantitative analysis demonstrates parity in performance among the methods, highlighting our approach's efficiency with a fast finetuning process of $\sim$1 minute.

User Study. To evaluate the efficacy of our multi-reference T2I style transfer approach, a comprehensive user study, detailed in Table 2, is conducted. The study unveils a marked discrepancy in text and style precision, with ArtAdapter achieving 4.74 and 4.76 scores, respectively, thus underscoring its enhanced adherence to textual prompts and style references, surpassing TI [14] and LoRA [20] benchmarks. Moreover, the elevated quality score of 4.43 underscores the visual appeal of our generated images. Collectively, these results indicate that our approach provides an improved T2I style transfer experience from a user perspective.

4.4. Ablation Study

The ablative analysis illustrated in Figure 7 dissects our ArtAdapter, emphasizing the distinct contributions and impact of each core mechanism on style transfer performance.

Auxiliary Content Adapter. The first row underscores the crucial role of the Auxiliary Content Adapter (ACA). In its absence, the model inappropriately adopts content semantics. The outputs will be dominated by content features in the style reference—such as the face in "Mona Lisa". The ACA's function to separate content and style learning is confirmed; it empowers the model to preserve style features while filtering out content during inference, guaranteeing that the generated images adhere to the textual prompts.

Explicit Adaptation. In the second and third rows, we scrutinize the effectiveness of the Explicit Adaptation mechanism. Without this mechanism, the model shows a marked decrease in its ability to accurately represent style across different levels. This is evidenced by distorted colors, imprecise geometric formations, and a lack of depth in capturing the compositional elements and artistic intent, such as those found in Salvador Dali's surrealist works. The explicit adaptation mechanism is thereby shown to be indispensable for the model's ability to internalize style contexts, generating outputs with high style fidelity.

Exclusive Style Encoder. Sole reliance on the multi-level style encoder, without incorporating adaptation in the backbone, leads to significant degradation in style representation, especially obvious in the second and third rows. The style encoder can only capture features such as imprecise color and rough texture patterns, failing to grasp the more profound stylistic expressions. This clearly highlights the critical role of the diffusion backbone in adapting to various style contexts, affirming that the integration of all components is critical for optimal style transfer performance.

5. Conclusion

In this work, our innovative T2I style transfer framework, ArtAdapter, has effectively demonstrated its ability to synthesize images that faithfully align with given textual prompts and style references. By employing a multi-level style encoder for nuanced style capture, the Explicit Adaptation for effective style integration, and the Auxiliary Content Adapter (ACA) for content-style separating, our ArtAdapter sets a new standard in achieving both style and text fidelity. The implementation of our fast finetuning strategy significantly enhances the efficiency of style alignment, establishing our model as a benchmark in the realm of T2I style transfer. However, ArtAdapter does have limitations, particularly in style mixing. We observe that high-level style embeddings often inadvertently incorporate elements from lower levels, causing interference in the style mixing process. We aim to refine the disentanglement of hierarchical style features to improve the authenticity and precision of style mixing in future works.

References

[1] Clip+mlp aesthetic score predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022.

[2] Low-rank adaptation for fast text-to-image diffusion fine-tuning. https://github.com/cloneofsimo/lora, 2022.

[3] Chatgpt. https://openai.com/blog/chatgpt, 2023.

[4] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. 2023.

[6] Dar-Yen Chen. Artfusion: Controllable arbitrary style transfer using dual conditional latent diffusion models, 2023.

[7] Haibo Chen, Lei Zhao, Zhizhong Wang, Zhang Hui Ming, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Artistic style transfer with internal-external learning and contrastive learning. In *Advances in Neural Information Processing Systems*, 2021.

[8] Haibo Chen, lei zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Artistic style transfer with internal-external learning and contrastive learning. In *Advances in Neural Information Processing Systems*, 2021.

[9] 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 (CVPR)*, 2022.

[10] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[13] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-Driven Artistic Style Transfer. In *European Conference on Computer Vision (ECCV)*, 2022.

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023.

[15] Golnaz Ghiasi, Honglak Lee, Manjunath Kudlur, Vincent Dumoulin, and Jonathon Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. In *BMVC*, 2017.

[16] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Proc. NeurIPS*, 2014.

[17] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020.

[20] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[22] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[23] Xuhui Jia, Yang Zhao, Kelvin C. K. Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models, 2023.

[24] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[25] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. In *The Eleventh International Conference on Learning Representations*, 2023.

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[27] Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015.

[28] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6649–6658, 2021.

[29] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

[30] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. 2022.

[31] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[32] Fred Phillips and Brandy Mackintosh. Wiki art gallery, inc.: A case for critical thinking. *Issues in Accounting Education*, 2011.

[33] 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 *ICML*, 2021.

[34] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

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

[36] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*, 2019.

[37] 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*, 2022.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015.

[39] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation. In *CVPR*, 2022.

[40] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models, 2023.

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. 2022.

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[43] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning, 2023.

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[45] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[47] Zhouxia Wang, Xintao Wang, Liangbin Xie, Zhongang Qi, Ying Shan, Wenping Wang, and Ping Luo. Styleadapter: A single-pass lora-free model for stylized image generation, 2023.

[48] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. Styleformer: Real-time arbitrary style transfer via parametric style composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[49] Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv*, 2023.

[50] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. *arXiv preprint arXiv:2305.16223*, 2023.

[51] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*, 2023.

[52] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[53] 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 (ICCV)*, 2023.

[54] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH*, 2022.

[55] 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)*, 2023.

[56] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. 2023.