StyleStudio: Text-Driven Style Transfer with Selective Control of Style Elements

Mingkun Lei, Xue Song, Beier Zhu, Hao Wang, Chi Zhang

AGI Lab, Westlake University, Fudan University, Nanyang Technological University, The Hong Kong University of Science and Technology (Guangzhou)

Outline

Background: Text-Driven Style Transfer is Transforming Image Synthesis

- Text-driven style transfer is a critical task in image synthesis, blending the style of a reference image with content described by a text prompt.
- This field has significant applications in digital art, advertising, and game design, enabling creative workflows.
- Recent advancements in text-to-image generative models, such as Stable Diffusion, have improved style transformations while preserving content fidelity.

The Problem: Challenges in Text-Driven Style Transfer

- Defining “style” is inherently ambiguous, encompassing elements like color palettes, textures, lighting, and brush strokes.
- Existing models often overfit to reference styles, reducing flexibility and adaptability.
- Maintaining alignment with textual prompts and avoiding artifacts like layout instability remain unresolved issues.

Our Core Contributions

- Proposed a cross-modal Adaptive Instance Normalization (AdaIN) mechanism to integrate style and text features, improving alignment.
- Developed Style-based Classifier-Free Guidance (SCFG) to selectively control stylistic elements, filtering out irrelevant influences.
- Incorporated a Teacher Model to stabilize spatial layouts during early generation stages, mitigating artifacts.
- Demonstrated significant improvements in style transfer quality and text alignment, compatible with existing frameworks without fine-tuning.

Methodology: Overview of Our Approach

- Our method introduces three key components to address the challenges in text-driven style transfer:
- Cross-Modal Adaptive Instance Normalization (AdaIN): Ensures balanced fusion of style and text features.
- Teacher Model: Stabilizes spatial layouts during early generation stages.
- Style-Based Classifier-Free Guidance (SCFG): Enables selective control over stylistic elements.

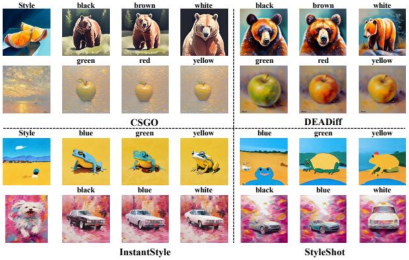

Figure: Overfitting in text-to-image models: style dominates text prompts.

Cross-Modal Adaptive Instance Normalization (AdaIN)

- Normalizes text features based on style features, ensuring balanced fusion and minimizing conflicts between text and style inputs.
- Replaces the traditional weighted sum approach, enabling effective feature integration without additional training.
- Improves alignment between textual prompts and reference styles, reducing generation conflicts.

Teacher Model for Layout Stabilization

- Shares spatial attention maps during early denoising steps to stabilize layout structures.
- Mitigates artifacts such as checkerboard patterns by selectively replacing self-attention maps in the stylized image with those from the original diffusion model.
- Ensures consistent layout arrangements, improving the overall quality of generated images.

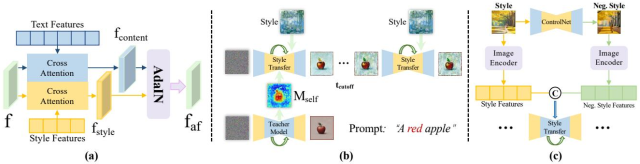

Figure: Cross-Modal AdaIN with Teacher Model and Style-Based CFG.

Style-Based Classifier-Free Guidance (SCFG)

- Inspired by classifier-free guidance, SCFG uses a negative style image to disentangle and emphasize desired style elements.
- Filters out irrelevant or conflicting features, ensuring precise control over stylistic components in complex scenarios.
- Improves the ability to selectively apply style elements, avoiding unintended influences.

Figure: Checkerboard artifact in CSGO method vs. SDXL results with same noise.

Experimental Setup

- Datasets: Evaluated on diverse datasets to test style transfer quality and text alignment.
- Metrics: Used text alignment accuracy, style fidelity, and user preference as evaluation metrics.
- Baselines: Compared against state-of-the-art methods, including IP-Adapter, InstantStyle, and StyleAlign.

Key Results: Quantitative Comparison

- Our method achieves the highest text alignment accuracy, outperforming state-of-the-art methods.
- Improves text alignment by 8.7
- Demonstrates significant improvements in style fidelity and user preference.

Table: Table 1: Quantitative comparison with state-of-the-art methods

Metric	SDXL-based Methods			SD15-based Methods			Ours	
	IP-Adapter	InstantStyle	CSGO	StyleAlign	StyleCrafter	StyleShot	DEADiff	Ours
Text Alignment ↑	0.221	0.229	0.216	0.180	0.189	0.202	0.229	**0.235**
Infer Time (s)	6	6	9	48	4	3	2	17
User-study Text (%)	7.48	6.46	7.99	5.78	3.06	2.55	1.87	**62.92**
User-study Style (%)	6.63	8.67	6.97	7.82	8.67	5.10	5.27	**50.85**

Ablation Study: Impact of Key Components

- Cross-Modal AdaIN improves text alignment accuracy by 5.5%.
- Teacher Model contributes a 3.2% improvement in text alignment.
- Combining both components achieves an 8.7% improvement, demonstrating their complementary effects.

Table: Table 2: Ablation study evaluating the impact of our proposed methods

Cross-Modal AdaIN	Teacher Model	Text Alignment ↑
		0.216
	✓	0.223 (+3.2%)
✓		0.228 (+5.5%)
✓	✓	**0.235 (+8.7%)**

Conclusion and Future Work

- Our method addresses critical limitations in text-driven style transfer, improving alignment, control, and stability.
- Demonstrated significant improvements in style fidelity and text alignment, compatible with existing frameworks.
- Future work includes improving efficiency and exploring strategies to further mitigate style overfitting.

Questions & Discussion

- Thank you for your attention!
- Questions and feedback are welcome.