Semantix: An Energy Guided Sampler for Semantic Style Transfer

Huiang He[1*], Minghui Hu[2*], Chuanxia Zheng[3], Chaoyue Wang[4], Tat-Jen Cham[2]

[1] South China University of Technology, [2] & Nanyang Technological University

[3] University of Oxford, [4] The University of Sydney

Contribution

☑ **Consistency.** **Semantix** can transfer style through semantic correspondence with semantic alignment and visual consistency.

☑ **Generic.** **Semantix** can be applied across both images and videos as it is an energy-guided sampler. It is not restricted by the fundation models.

☑ **Training-free.** Benefiting from energy guidance, **Semantix** can steer style transfer without the need of model training or finetuning.

Background

1. DIFT[1]: Diffusion models can capture rich semantic information and establish precise **semantic correspondence** between the context and reference images.

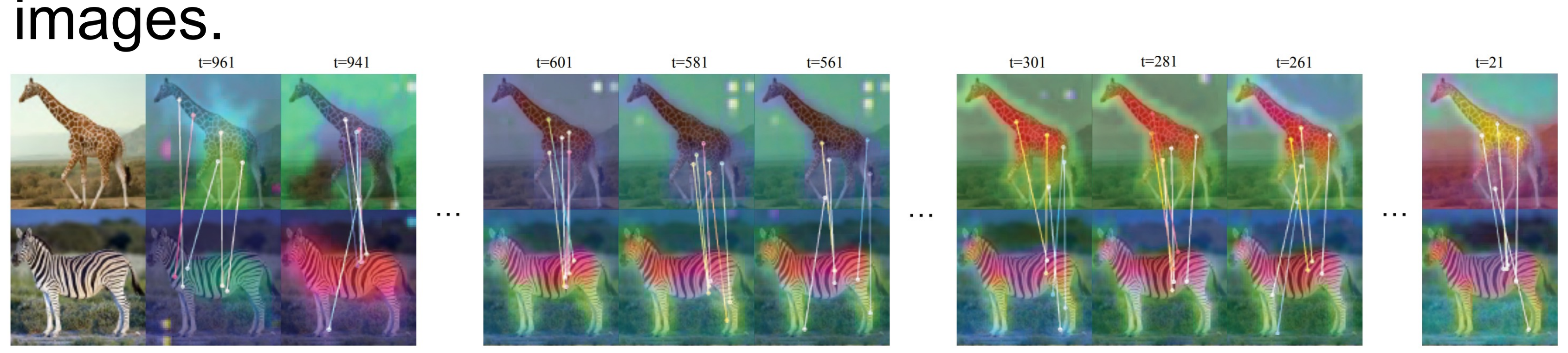

Fig. 1 Visualizing feature maps. We extracted features from the second block of the diffusion model decoder and visualized the top three PCA components and **feature mapping** at each timestep.

2. Energy Guidance[2,3]: The energy function can provide additional directional information to guide the sampling process along with classifier-free guidance.

[1] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems. 2023.*

[2] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *International Conference on Learning Representations. 2024.*

[3] Zhao, Min, et al. "Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations." *Advances in Neural Information Processing Systems* 35 (2022): 3609-3623.

Method

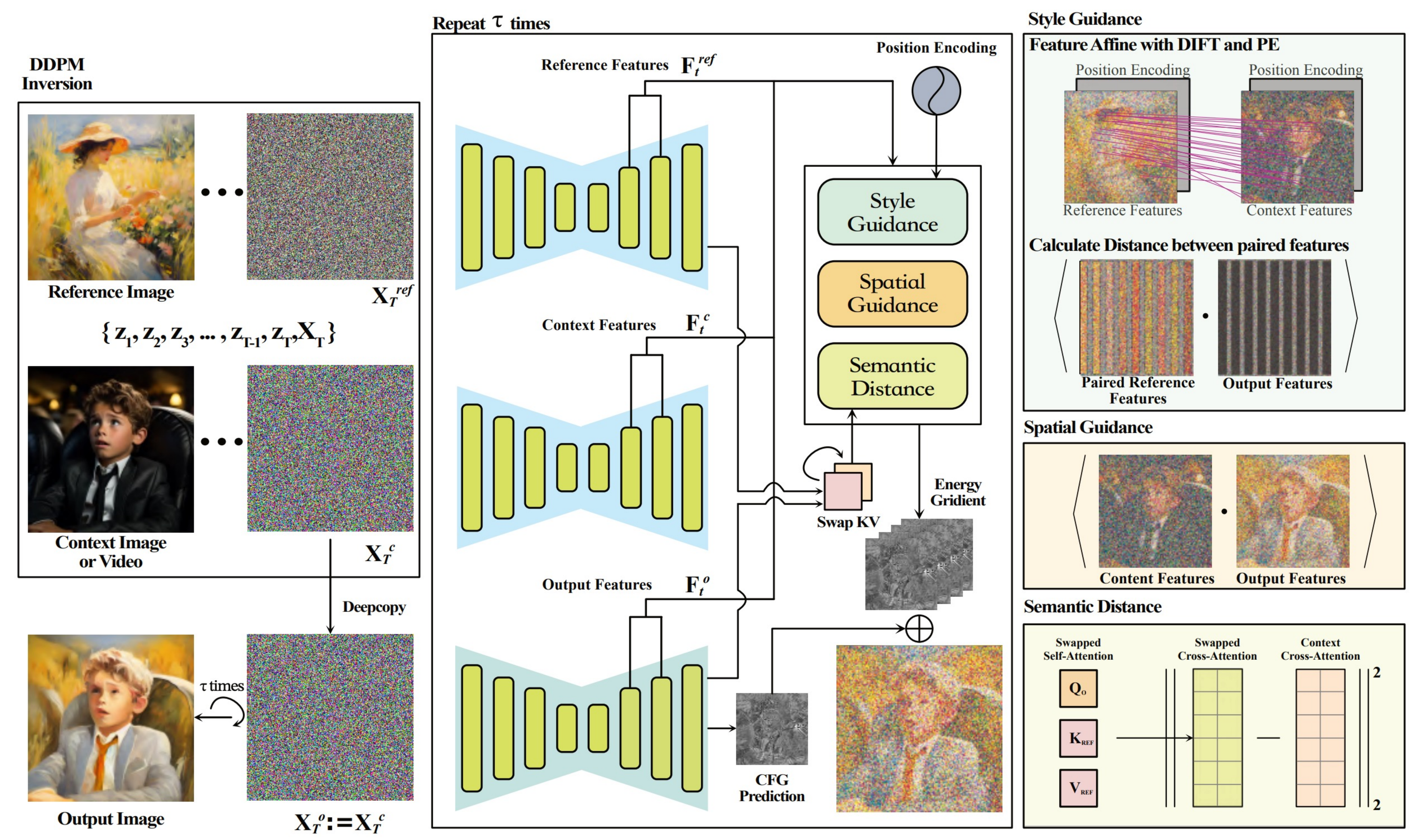

Fig. 2 We present **Semantix**, a **training-free** energy-guided sampler for *Semantic Style Transfer* that achieves style and appearance transfer across image and video through semantic alignment.

• Design of Energy Function

We regulate the sampling process (Eq.1) via an **energy function** to achieve semantic style transfer, the energy function consists of three components (Eq.2)

$$\hat{\epsilon}_t = (1+\omega)\epsilon_\theta(\boldsymbol{x}_t; t, \mathcal{C}) - \omega\epsilon_\theta(\boldsymbol{x}_t; t, \phi) + \nabla_{\boldsymbol{x}_t}\mathcal{F}(\boldsymbol{x}_t; t, \mathcal{C}), \quad (1)$$

$$\mathcal{F}(\boldsymbol{x}_t; t, \mathcal{C}) = \gamma_{ref}\mathcal{F}_{ref} + \gamma_c\mathcal{F}_c + \gamma_{reg}\mathcal{F}_{reg}, \quad (2)$$

1. **Style Feature Guidance**: to align the style features with the reference image.

$$D_{ij} = \|\mathrm{v}^c_{p_i} - \mathrm{v}^{ref}_{p_j}\|^2_2, \quad \forall \mathrm{v}^c_{p_i} \in F^c_t, \quad \forall \mathrm{v}^{ref}_{p_j} \in F^{ref}_t,$$

$$p^*_j = \arg\min_{p_j} D_{ij}.$$

$$\bar{F}^c_{t_{\{i\}}} \leftarrow F^c_t + \lambda_{pe}\cdot\boldsymbol{pe}_{\{i\}},$$

$$\bar{F}^{ref}_{t_{\{i\}}} \leftarrow F^{ref}_t + \lambda_{pe}\cdot\boldsymbol{pe}_{\{i\}}.$$

$$\mathcal{F}_{ref} \propto \boldsymbol{d}\left(F^{out}_t, F^{ref*}_t\right),$$

2. **Spatial Feature Guidance**: to maintain spatial coherence with context.

$$\mathcal{F}_c \propto \boldsymbol{d}\left(F^{out}_t, F^c_t\right).$$

3. **Semantic Distance**: to regularise the whole energy function.

$$\mathcal{F}_{reg} = \|\text{Cross-Attn}^{out}_{swap} - \text{sg}(\text{Cross-Attn}^c)\|^2_2,$$

Results and Comparison

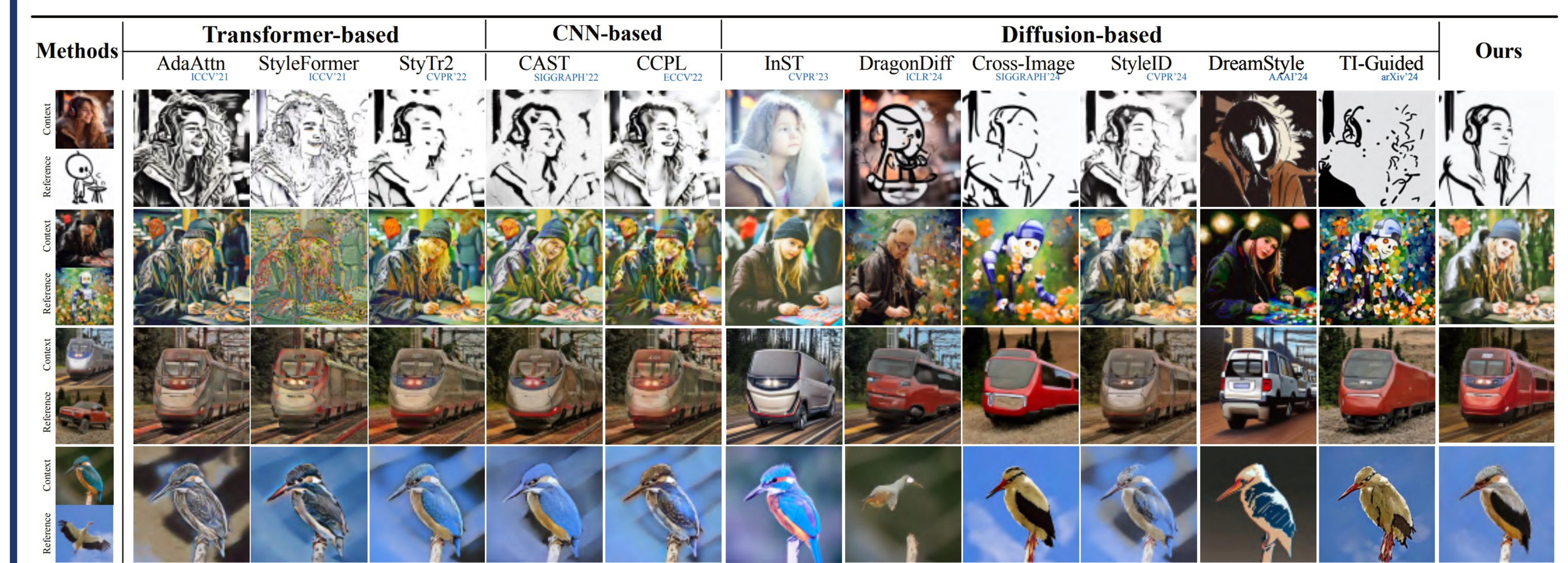

Fig. 3 Comparison Results for Image Transfer

Fig. 4 Qualitative Results for Video Transfer.

Tab. 1 Quantitative Results for Image Transfer.

Metrics	AdaAttn	StyleFormer	StyTR2	CAST	CCPL	InST	Cross-Image	StyleID	DreamStyler	TI-Guided	Ours
LPIPS ↓	0.581	0.560	0.476	0.465	0.523	0.548	0.703	0.514	0.580	0.649	**0.461**
CFSD ↓	0.189	0.156	0.155	0.133	0.133	0.408	0.232	0.160	0.789	0.183	**0.117**
SSIM ↑	0.403	0.331	0.561	0.514	0.536	0.383	0.454	0.527	0.334	0.453	**0.589**
Gram Metrics $_{\times 10^2}$ ↓	7.929	2.822	5.403	6.594	4.861	4.917	5.850	2.878	6.990	4.811	**2.524**
PickScore ↑	16.87	18.85	16.76	16.72	16.75	16.80	17.45	19.68	16.80	18.39	**19.95**
HPS ↑	16.81	18.20	16.81	16.77	16.79	16.87	16.59	18.70	16.87	17.56	**18.78**

The **best** results are highlighted in **bold font**, and the second-best are underlined.
We compare our method with recent state-of-the-art methods in terms of structure preservation, style similarity and image aesthetics.
* To measure structure preservation capability, we calculate the LPIPS, CFSD and SSIM.
* For style similarity, we compute Gram Metrics as style loss.
* We utilize PickScore and HPS as aesthetic evaluation metrics.

Tab. 2 Quantitative Results for Video Transfer.

Metric	MCCNet	UNIST	Cross-Image	CCPL	Ours
Semantic Consistency ↑	0.714	0.861	0.936	0.942	**0.944**
Object Consistency ↑	0.723	0.777	0.939	0.943	**0.955**
Motion Alignment ↑	-5.251	-4.178	-3.878	**-1.792**	-1.894
Visual Quality ↑	52.11	43.97	47.33	48.92	**55.86**
Motion Quality ↑	53.35	**55.07**	53.14	53.25	53.99
Temporal Consistency ↑	59.14	45.43	55.85	59.64	**60.05**

The **best** results are highlighted in **bold font**, and the second-best results are underlined.

Project Page