Specialist Diffusion: Plug-and-Play Sample-Efficient Fine-Tuning of Text-to-Image Diffusion Models to Learn Any Unseen Style

Haoming Lu, Hazarapet Tunanyan, Kai Wang, Shant Navasardyan, Zhangyang Wang, Humphrey Shi

Picsart
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Specialist Diffusion

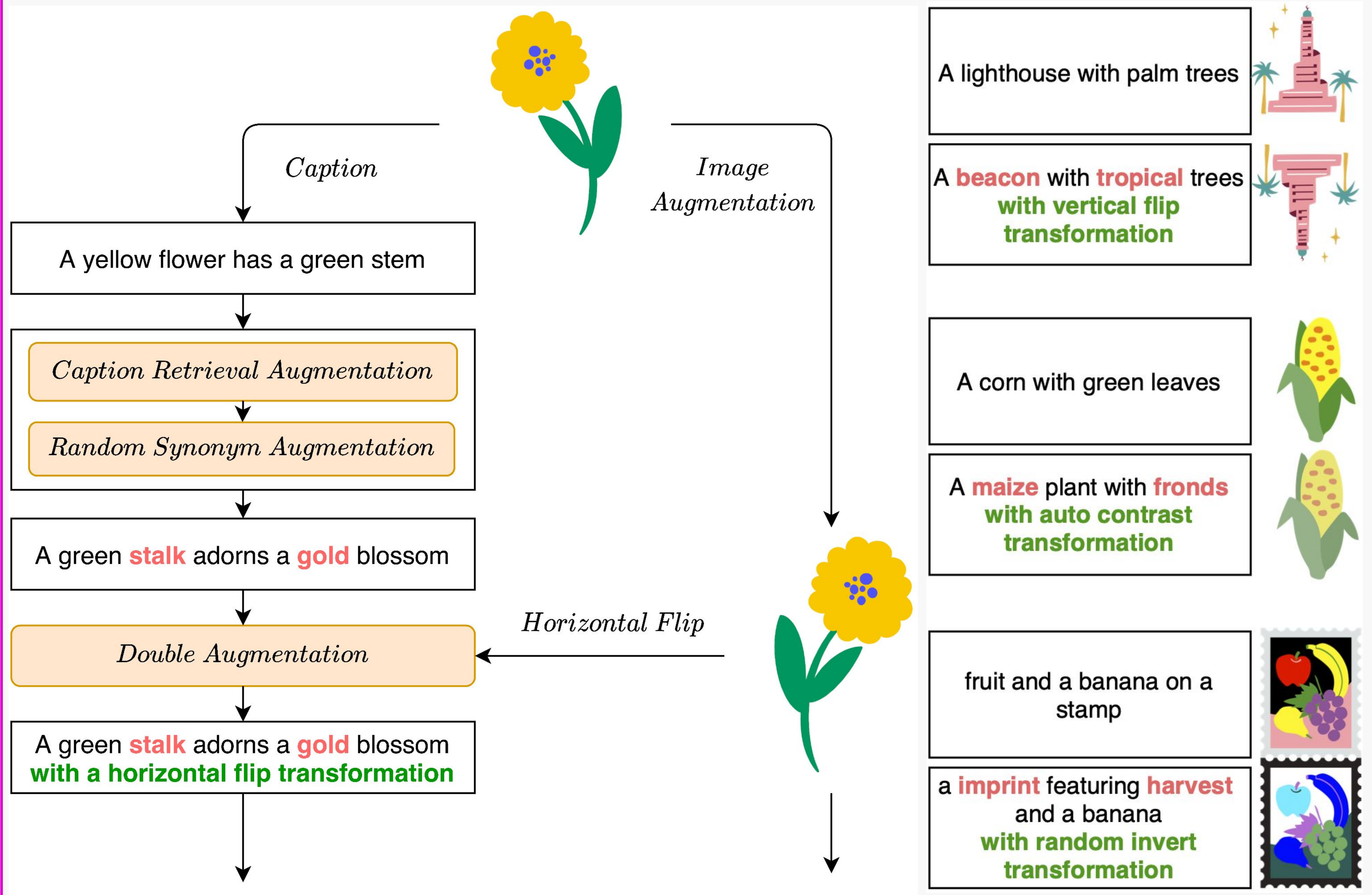
- ➔ Any unseen style can be specialized
- ➔ Less than 10 examples needed
- ➔ Less than 30 m. on a single RTX A6000 GPU

Contributions

- ➔ Advanced Augmentation
- ➔ Content-Loss to avoid forgetting
- ➔ Sparse Timesteps Updating

Advanced Augmentation

Overview / Examples

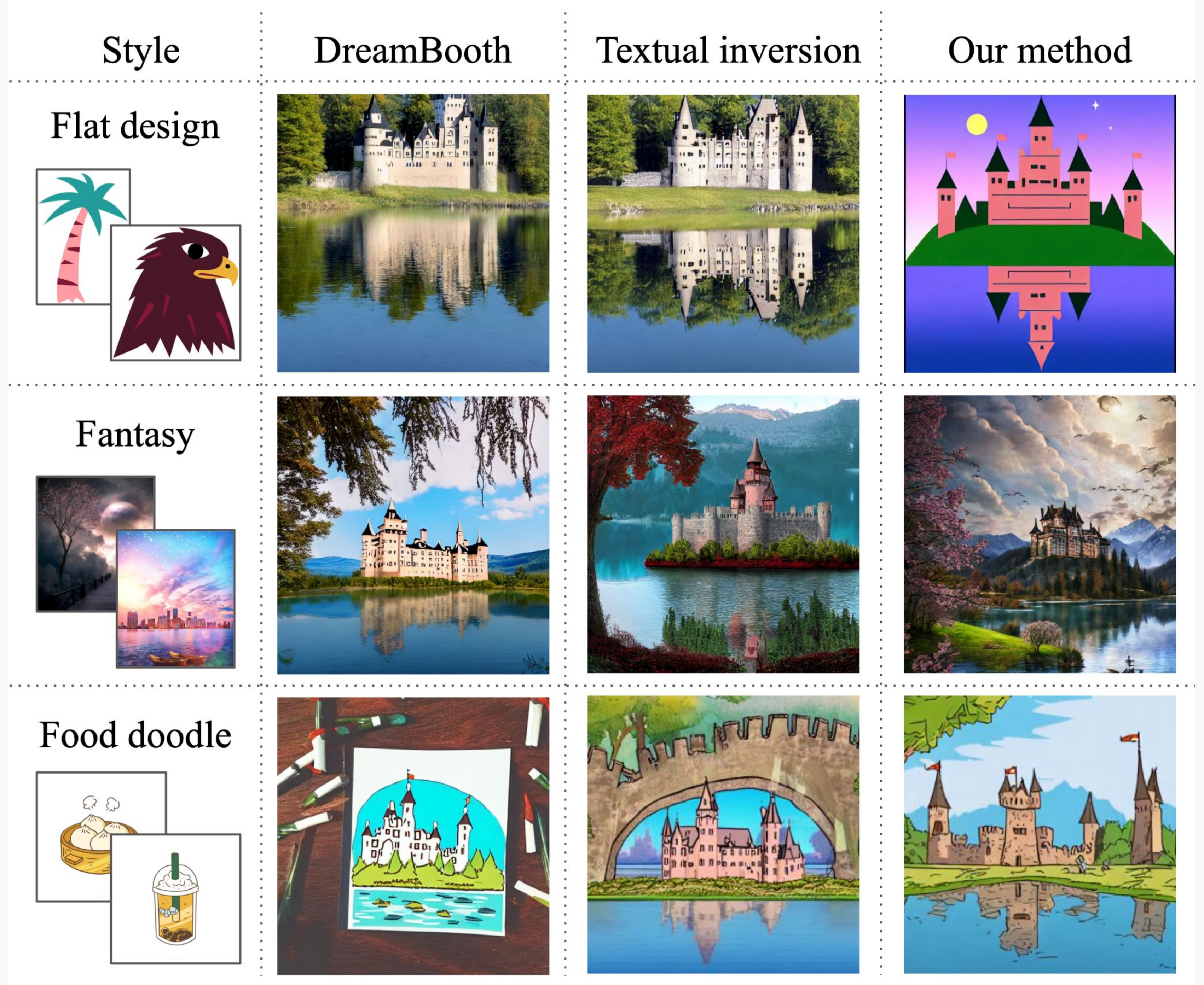


Content Loss

$$\arg\min_{w \in w^+} D_{CLIP}(G(w), t) + R$$

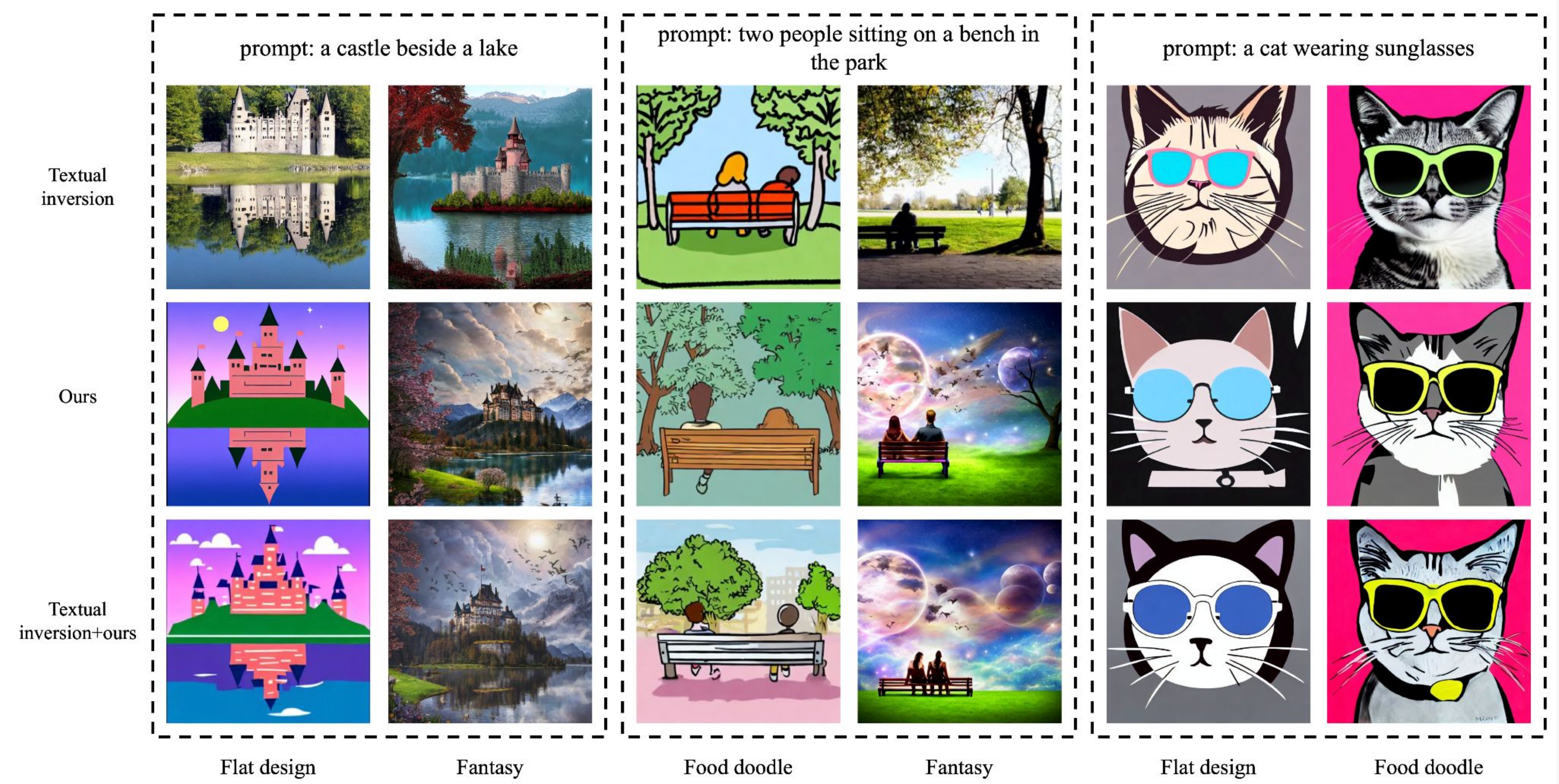
D_{CLIP}	Cosine distance between CLIP embeddings
$G(w)$	Generative model with w parameters
$t,\ R$	Text and a regularization term

Qualitative Comparison



Prompt: "a castle beside a lake"

Plug-and-Play



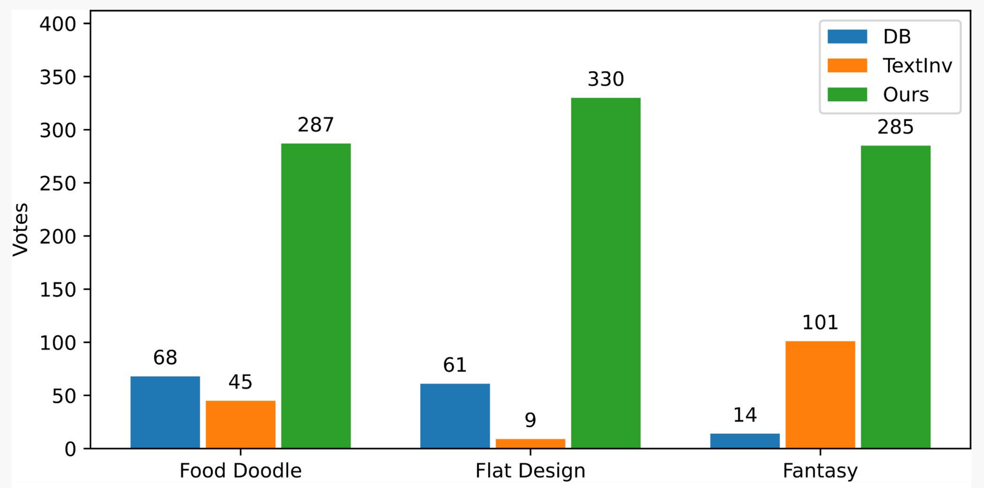
Quantitative Comparison

Style\Method	DB	TextInv	Ours	Ours+Inv
Fantasy	437.254	452.150	399.351	**352.650**
Flat design	445.089	460.252	421.410	**363.362**
Food doodle	491.302	451.466	441.640	**409.607**

Table 1. FIDs on different styles × different methods

Style\Method	DB	TextInv	Ours	Ours+Inv
Fantasy	1.087	0.202	0.202	**0.147**
Flat design	0.839	0.276	0.116	**0.098**
Food doodle	0.492	0.104	0.046	**0.037**

Table 2. Average style loss (VGG-based) between generated images & corresponding training examples. Numbers scaled by 10^2.



User study with the question: "*which one aligns best with the reference image in style?*"