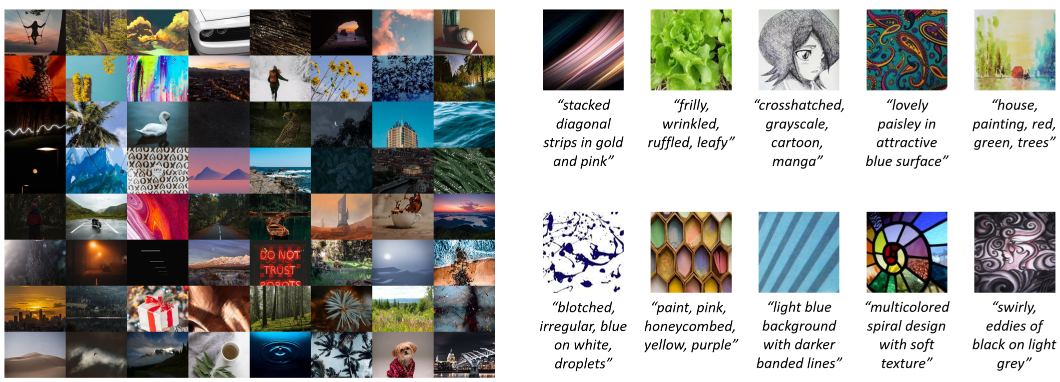
SUPPLEMENTARY MATERIAL OF LANGUAGE-DRIVEN IMAGE STYLE TRANSFER

Anonymous authors
Paper under double-blind review

A USED DATASET ON LDIST

We build a dataset to evaluate language-based image style transfer (`LDIST`). As shown in Fig. 1, we collect 14,924 wallpapers from WallpapersCraft as content images ($\mathcal{C}$), including diverse scenes like *building*, *animal*, or *island*. We apply DTD^2 (Wu et al., 2020) that provides 5,368 pairs of texture image ($\mathcal{S}$) and textual description ($\mathcal{X}$) as reference styles, such as *striped*, *smeared*, or *paisley*.



(a) Content Image from WallpapersCraft **(b) Style Image and Style Instruction from DTD²**

Figure 1: The used dataset is built upon content images and reference styles.

B RETRIEVAL-BASED BASELINE

Apart from generating the transferred result ($\hat{\mathcal{O}}$) directly by the style instruction ($\mathcal{X}$), we also investigate a two-step retrieval-based baseline. Firstly, we search the related style image ($\mathcal{S}$) via $\mathcal{X}$ and then perform the standard style transfer from both. We adopt the learned BERT encoder from DTD^2(Wu et al., 2020) for the text-image retrieval with *MAP* 13.5, *R@5* 5.2, and *R@20* 17.3. Table 1 shows that the two-step baseline performs slightly better than our CLVA on automatic metrics. However, this retrieval-based method still relies on the existing collections of style images and may limit the diversity of style patterns due to the collection size.

	Automatic Metrics (*vs.* Semi-GT)				
Method	SSIM (↑)	Percept (↓)	FAD (↓)	VLS (↑)	RS (↑)
CLVA	60.586	0.02076	**0.11318**	22.785	98.798
Retrieval	**60.745**	**0.02059**	0.11931	**22.942**	**98.942**

Table 1: Testing results of the two-step retrieval-based baseline.

C HUMAN EVALUATION

We investigate the quality of `LDIST` results from the human aspect through Amazon Mechanical Turk (AMT). Fig. 2 illustrates the screenshots of the human ranking task. MTurkers rank the correlation of the LDIST result from each method between the style instruction (*vs.* Instruction) or the style image (*vs.* Style). Each MTurker rewards \$1.0 and takes a mean of 17 minutes for 6 tasks.

(a) Human Evaluation (*vs.* Instruction) **(b) Human Evaluation (*vs.* Style)**

Figure 2: The screenshots of the human ranking task for evaluating the quality of `LDIST` results.

D PHOTOREALISTIC `LDIST`

In this paper, we focus on *artistic* style transfer, which manipulates colors and textures of a content image. Ideally, our CLVA can also support *photorealistic* `LDIST` by replacing style instructions $\mathcal{X}$ with photorealistic instructions. However, there is a practical issue where the caption is not detailed enough to represent itself *visual concept*. For example, from the ARTEMIS dataset (Achlioptas et al., 2021), descriptions are usually too abstract to provide explicit style patterns. Therefore, we leave the collection of *photorealistic* style instructions and *photorealistic* `LDIST` as a future work.

Figure 3: The description of natural image is too abstract to support *photorealistic* `LDIST`.

E ETHICS DISCUSSION

Though our work benefits creative visual applications, there may be a "fake as real" doubt for those manipulated images. To mitigate this issue, we can apply techniques from image forensics (Wang et al., 2020; Huh et al., 2018; Frank et al., 2020) to detect the authenticity of an image. Regarding guided instructions, for example, hate speech detection (Aluru et al., 2020; Huang et al., 2020; Samghabadi et al., 2020; Samanta et al., 2019) can help to filter out malicious texts and prevent from producing controversial results with ethics concerns.

REFERENCES

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. ArtEmis: Affective Language for Visual Art. In *CVPR*, 2021.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. Deep Learning Models for Multilingual Hate Speech Detection. In *ECML-PKDD*, 2020.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *ICML*, 2020.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. Multilingual Twitter Corpus and Baselines for Evaluating Demographic Bias in Hate Speech Recognition. In *LREC*, 2020.

Please visit project website for more visualization results: `https://ai-sub.github.io/ldist/`

Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. Fighting Fake News: Image Splice Detection via Learned Self-Consistency. In *ECCV*, 2018.

Bidisha Samanta, Niloy Ganguly, and Soumen Chakrabarti. Improved Sentiment Detection via Label Transfer from Monolingual to Synthetic Code-Switched Text. In *ACL*, 2019.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In *LREC*, 2020.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-Generated Images are Surprisingly Easy to Spot...for Now. In *CVPR*, 2020.

Chenyun Wu, Mikayla Timm, and Subhransu Maji. Describing Textures using Natural Language. In *ECCV*, 2020.