Language-Driven Artistic Style Transfer

Tsu-Jui Fu[†], Xin Eric Wang[‡], William Yang Wang[†]

[†]UC Santa Barbara [‡]UC Santa Cruz
{tsu-juifu, william}@cs.ucsb.edu xwang366@ucsc.edu

Abstract. Despite having promising results, style transfer, which requires preparing style images in advance, may result in lack of creativity and accessibility. Following human instruction, on the other hand, is the most natural way to perform artistic style transfer that can significantly improve controllability for visual effect applications. We introduce a new task—language-driven artistic style transfer (`LDAST`)—to manipulate the style of a content image, guided by a text. We propose contrastive language visual artist (CLVA) that learns to extract visual semantics from style instructions and accomplish `LDAST` by the patch-wise style discriminator. The discriminator considers the correlation between language and patches of style images or transferred results to jointly embed style instructions. CLVA further compares contrastive pairs of content images and style instructions to improve the mutual relativeness. The results from the same content image can preserve consistent content structures. Besides, they should present analogous style patterns from style instructions that contain similar visual semantics. The experiments show that our CLVA is effective and achieves superb transferred results on `LDAST`.

1 Introduction

Style transfer [16,33,23,41,32,25] adopts appearances and visual patterns from another reference style images to manipulate a content image. Artistic style transfer has a considerable application value for creative visual design, such as image stylization and video effect [55,70,15,21]. However, it requires preparing collections of style image in advance. It even needs to redraw new references first if there is no expected style images, which is impractical due to an additional overhead. In contrast, language is the most natural way for humans to communicate. If a system can follow textual descriptions and automatically perform style transfer, we can significantly improve accessibility and controllability.

In this paper, we introduce Language-driven Artistic Style Transfer (`LDAST`). As illustrated in Fig. 1, `LDAST` treats a content image and a text as the input, and the style transferred result is manipulated based on the style description. It should preserve the structure of the content yet simultaneously modifies the style pattern that corresponds to the instruction. `LDAST` is different from the general language-based image-editing (LBIE) [39,31,36,10] that aims at altering objects or properties of objects. The main challenge of `LDAST` is to extract visual semantics from language. Humans use not only explicit visual attributes but

Fig. 1. Language-driven Artistic Style Transfer (`LDAST`). `LDAST` performs style transfer for a content image $\mathcal{C}$, guided by the visual attribute (the lower row) or even the visual content and emotional effect (the upper row) from a style instruction $\mathcal{X}$.

also visual content or emotional effects to describe style feelings. For example, it requires connecting *"water, sketching, and painting"* or *"peaceful, feel content"* with their visual concepts and further carrying out correlated style transfer.

We present contrastive language visual artist (CLVA), including language visual artist (LVA) and contrastive reasoning (CR), to perform style transfer conditioning on guided texts. LVA preserves content structures from content images $\mathcal{C}$ and extracts visual semantics from style instructions $\mathcal{X}$. LVA learns the latent style pattern based on the distinguishment between patches of style imags or transferred results from the patch-wise style discriminator. Furthermore, CR boosts by comparing contrastive pairs where relative content images or style instructions should present similar content structures or style patterns.

To evaluate `LDAST`, we conduct experiments upon DTD^2 [61] and ArtEmis [1]. DTD^2 provides texture images with its colors or texture patterns in text. ArtEmis collects explanations of visual contents and emotional effects for artworks. We treat these annotations as style instructions for the challenging `LDAST` concerning visual attributes or human style feelings. The experiments show that our CLVA is effective for `LDAST` and achieves superb yet efficient transferred results on both automatic metrics and human evaluation. Our contributions are four-fold:

- We introduce `LDAST` that follows natural language for artistic style transfer;
- We present CLVA, which learns to extract explicit visual semantics from style instructions and provide sufficient style patterns for `LDAST`;
- We conduct the evaluation on DTD^2 and ArtEmis to consider diverse style instructions with visual attributes and emotional effects;
- Extensive experiments and qualitative examples demonstrate that our CLVA outperforms baselines regarding both effectiveness and efficiency.

2 Related Work

Artistic Style Transfer. Style transfer [16,25,57,7,18,26,52] redraws an image with a specific style. Since being a popular form of art, incorporating painting with digital design can produce attractive visual effect (VFX). In general, style transfer can be divided into two categories: *photorealistic* and *artistic*. Photorealistic style transfer [38,34,67,42] aims at applying reference styles on scenes without hurting details and satisfying contradictory objectives. By contrast, artistic style transfer [16,33,23,41,32,35,5] captures style concepts from reference and

modifies color distributions and texture patterns of content images. However, it requires preparing numerous style images in advance, which limits practicality of style transfer. To tackle this issue, `LDAST` allows following textual descriptions to perform *artistic* style transfer and improves the accessibility of VFX design.

Language-based Image Editing. The general task of `LDAST` is language-based image editing (LBIE), which also uses language to edit input images. With rule-based instructions and predefined semantic labels, they [8,30] first carry out LBIE but under limited practicality. Inspired by text-to-image generation [46,69,65], previous works [6,39,31,63,36,10,12,13] perform LBIE by conditional GAN, which modifies the properties of objects in the image. In contrast, `LDAST` aims at preserving the scene structure from the content image and performing stylization guided by the style instruction.

CLIP-guided Optimization. Recently, based on the powerful visual-linguistic connection of CLIP [53], CLIP-guided image synthesis [45,40] has shown exciting results. StyleCLIP [43] and NADA [14] tweak the latent code of a pre-trained StyleGAN [27] for image editing. Since heavily relying on a pre-trained generator, both are confined to the training domain, and the results can only present limited stylization. CLIPstyler [29] updates the style transfer network for target style patterns from the CLIP alignment. Though supporting arbitrary content images, CLIPstyler still requires hundreds of iterations and takes lots of time with considerable GPU memory, suffering from the efficiency and practicality overhead. Moreover, our experiments show that CLIP poorly captures detailed style patterns from instructions, which is intractable to perform explicit `LDAST`.

3 Language-Driven Artistic Style Transfer

3.1 Overview of CLVA

We introduce language-driven artistic style transfer (`LDAST`) to manipulate the style of a content image $\mathcal{C}$, guided by a style instruction $\mathcal{X}$, as illustrated in Fig. 1. For training, we have pairs of style images $\mathcal{S}$ with style instructions $\mathcal{X}$ to learn the mutual correlation. During testing, only $\mathcal{X}$ are provided for `LDAST` to carry out artistic style transfer purely relied on language. We present contrastive language visual artist (CLVA) in Fig. 2. Language visual artist (LVA) extracts content structures from $\mathcal{C}$ and visual patterns from $\mathcal{X}$ to perform `LDAST`. LVA adopts the patch-wise style discriminator D to connect extracted visual semantics to patches of paired style image ($\mathcal{P}_{\mathcal{S}}$ in Fig. 2). Contrastive reasoning (CR) allows comparing contrastive pairs $\mathcal{C}_1$-$\mathcal{X}_1$, $\mathcal{C}_2$-$\mathcal{X}_1$, and $\mathcal{C}_2$-$\mathcal{X}_2$ of content image and style instruction. In this way, it should present consistent content structures from the same content image $\mathcal{C}_2$ or analogous style patterns from related style images $\mathcal{S}_1$ and $\mathcal{S}_2$, despite using different style instructions.

3.2 Language Visual Artist (LVA)

To tackle `LDAST`, language visual artist (LVA) first adopts visual encoder G_{E} to extract the content feature $h^{\mathcal{C}}$ and the style feature $h^{\mathcal{S}}$ for an image. Text

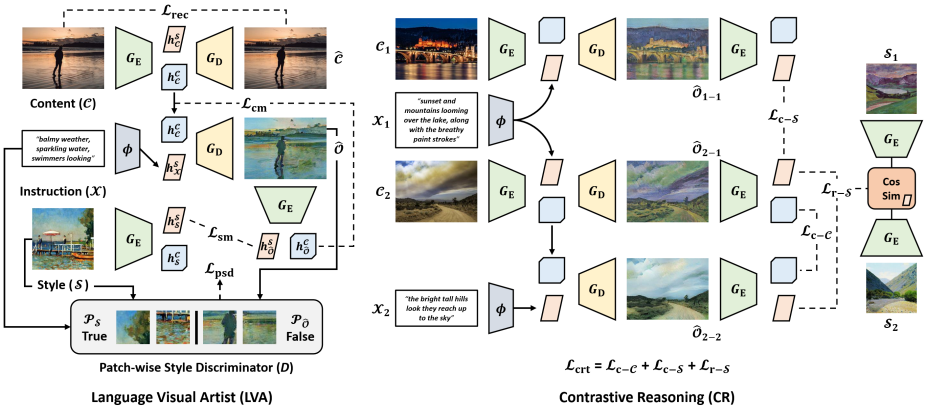

Fig. 2. Contrastive language visual artist (CLVA). Language Visual Artist (LVA) learns to jointly embed style images $\mathcal{S}$ and style instructions $\mathcal{X}$ by the patch-wise style discriminator D and perform `LDAST` for content images $\mathcal{C}$. Contrastive Reasoning (CR) compares contrasitve pairs to improve the relativeness between transferred results $\hat{\mathcal{O}}$.

encoder ϕ also extracts the style instruction feature $h^{\mathcal{S}}_{\mathcal{X}}$ from an instruction. $h^{\mathcal{C}}$ is a spatial tensor containing the content structure feature, and $h^{\mathcal{S}}$ represents the global style pattern. $\mathcal{S}^{S}_{\mathcal{X}}$ embeds into the same space of $h^{\mathcal{S}}$ to reflect the extracted visual semantic. Then, visual decoder G_{D} produces transferred results $\hat{\mathcal{O}}$ from $h^{\mathcal{C}}_{\mathcal{C}}$ and $h^{\mathcal{S}}_{\mathcal{X}}$, which performs style transfer by style instructions:

$$\begin{aligned} h^{\mathcal{C}}_{\mathcal{C}}, h^{\mathcal{S}}_{\mathcal{C}} = G_{\mathrm{E}}(\mathcal{C}), \quad h^{\mathcal{S}}_{\mathcal{X}} = \phi(\mathcal{X}), \\ \hat{\mathcal{O}} = G_{\mathrm{D}}(h^{\mathcal{C}}_{\mathcal{C}}, h^{\mathcal{S}}_{\mathcal{X}}). \end{aligned} \tag{1}$$

In particular, G_{D} applies self-attention [68,41] to fuse $h^{\mathcal{C}}$ and $h^{\mathcal{S}}$ over the global spatial dimension. There are two goals to train LVA for `LDAST`: (i) preserving *content structures* from content images; (ii) presenting *style patterns* correlated with visual semantics of style instructions.

Structure Reconstruction. To preserve content structures, we consider that visual decoder G_{D} should be able to reconstruct input content images using extracted content features $h^{\mathcal{C}}_{\mathcal{C}}$ and style features $h^{\mathcal{S}}_{\mathcal{C}}$ from visual encoder G_{E}:

$$\begin{aligned} \hat{\mathcal{C}} = G_{\mathrm{D}}(h^{\mathcal{C}}_{\mathcal{C}}, h^{\mathcal{S}}_{\mathcal{C}}), \\ \mathcal{L}_{\mathrm{rec}} = ||\hat{\mathcal{C}} - \mathcal{C}||_2, \end{aligned} \tag{2}$$

where the reconstruction loss $\mathcal{L}_{\mathrm{rec}}$ is computed as the mean L2 difference between reconstructed content images $\hat{\mathcal{C}}$ and input content images $\mathcal{C}$.

Patch-wise Style Discriminator (D). Regarding style patterns, results $\hat{\mathcal{O}}$ guided by style instructions $\mathcal{X}$ are expected to present analogously to reference style images $\mathcal{S}$. To address the connection between linguistic from $\mathcal{X}$ and visual semantics from $\mathcal{S}$, we introduce the patch-wise style discriminator D. Inspired by texture synthesis [64,17], images with analogous patch patterns should appear

perceptually similar texture patterns. D tries to recognize the correspondence between an image patch $\mathcal{P}$ and a style instruction $\mathcal{X}$:

$$\begin{aligned} \mathcal{P}_{\hat{\mathcal{O}}}, \mathcal{P}_{\mathcal{S}} &= \text{Crop}(\hat{\mathcal{O}}), \text{Crop}(\mathcal{S}), \\ \mathcal{L}_{\text{psd}} &= \log(1 - D(\mathcal{P}_{\hat{\mathcal{O}}}, \mathcal{X})), \\ \mathcal{L}_{D} &= \log(1 - D(\mathcal{P}_{\hat{\mathcal{O}}}, \mathcal{X})) + \log(D(\mathcal{P}_{\mathcal{S}}, \mathcal{X})), \end{aligned} \tag{3}$$

where `Crop` is to randomly crop an image into patches. The patch-wise style loss $\mathcal{L}_{\text{psd}}$ aims at generating transferred results that are correlated with $\mathcal{X}$. Contrarily, by the discriminator loss $\mathcal{L}_D$, D learns to distinguish that a patch $\mathcal{P}$ is from style images ($\mathcal{P}_{\mathcal{S}}$) or transferred results ($\mathcal{P}_{\hat{\mathcal{O}}}$). This adversarial loss [19,48] encourages that transferred results from style instructions are presented similarly with style images, which jointly embeds the extracted visual semantics.

Content Matching and Style Matching. To further enhance the alignment with inputs, inspired by cycle consistency [71,62,44,66], we consider the content matching loss $\mathcal{L}_{\text{cm}}$ and the style matching loss $\mathcal{L}_{\text{sm}}$ of transferred results $\hat{\mathcal{O}}$. We adopt G_{E} again to extract content features $h^{\mathcal{C}}_{\hat{\mathcal{O}}}$ and style features $h^{\mathcal{S}}_{\hat{\mathcal{O}}}$ for $\hat{\mathcal{O}}$, where $h^{\mathcal{C}}_{\hat{\mathcal{O}}}$ and $h^{\mathcal{S}}_{\hat{\mathcal{O}}}$ should correlate with $h^{\mathcal{C}}_{\mathcal{C}}$ from $\mathcal{C}$ and $h^{\mathcal{S}}_{\mathcal{S}}$ from $\mathcal{S}$:

$$\begin{aligned} (h^{\mathcal{C}}_{\hat{\mathcal{O}}}, h^{\mathcal{S}}_{\hat{\mathcal{O}}}), (h^{\mathcal{C}}_{\mathcal{S}}, h^{\mathcal{S}}_{\mathcal{S}}) &= G_{\text{E}}(\hat{\mathcal{O}}), G_{\text{E}}(S), \\ \mathcal{L}_{\text{cm}}, \mathcal{L}_{\text{sm}} &= ||h^{\mathcal{C}}_{\hat{\mathcal{O}}} - h^{\mathcal{C}}_{\mathcal{C}}||_2, ||h^{\mathcal{S}}_{\hat{\mathcal{O}}} - h^{\mathcal{S}}_{\mathcal{S}}||_2. \end{aligned} \tag{4}$$

Therefore, transferred results are required to align with content structures and style patterns from inputs, which meets the goal of `LDAST`.

3.3 Contrastive Reasoning (CR)

The content image should transfer to various styles while preserving the same structure. Related style instructions can apply analogous style patterns to arbitrary content images. As shown in Fig. 2, contrastive reasoning (CR) compares content structures or style patterns from transferred results of contrastive pairs. The contrastive pair consists of two different content images $\mathcal{C}_1$ and $\mathcal{C}_2$ with two reference styles $\{\mathcal{S}_1, \mathcal{X}_1\}$ and $\{\mathcal{S}_2, \mathcal{X}_2\}$. We follow the LVA inference to acquire cross results for pairs of content images and style instructions:

$$\begin{aligned} (h^{\mathcal{C}}_{\mathcal{C}_1}, h^{\mathcal{S}}_{\mathcal{C}_1}), (h^{\mathcal{C}}_{\mathcal{C}_2}, h^{\mathcal{S}}_{\mathcal{C}_2}) &= G_{\text{E}}(\mathcal{C}_1), G_{\text{E}}(\mathcal{C}_2), \\ h^{\mathcal{S}}_{\mathcal{X}_1}, h^{\mathcal{S}}_{\mathcal{X}_2} &= \phi(\mathcal{X}_1), \phi(\mathcal{X}_2), \\ \hat{\mathcal{O}}_{\mathcal{C}_1\text{-}\mathcal{X}_1}, \hat{\mathcal{O}}_{\mathcal{C}_1\text{-}\mathcal{X}_2} &= G_{\text{D}}(h^{\mathcal{C}}_{\mathcal{C}_1}, h^{\mathcal{S}}_{\mathcal{X}_1}), G_{\text{D}}(h^{\mathcal{C}}_{\mathcal{C}_1}, h^{\mathcal{S}}_{\mathcal{X}_2}), \\ \hat{\mathcal{O}}_{\mathcal{C}_2\text{-}\mathcal{X}_1}, \hat{\mathcal{O}}_{\mathcal{C}_2\text{-}\mathcal{X}_2} &= G_{\text{D}}(h^{\mathcal{C}}_{\mathcal{C}_2}, h^{\mathcal{S}}_{\mathcal{X}_1}), G_{\text{D}}(h^{\mathcal{C}}_{\mathcal{C}_2}, h^{\mathcal{S}}_{\mathcal{X}_2}). \end{aligned} \tag{5}$$

Consistent Matching. Transferred results should present similar content structures ($\hat{\mathcal{O}}_{\mathcal{C}_2\text{-}\mathcal{X}_1}$ and $\hat{\mathcal{O}}_{\mathcal{C}_2\text{-}\mathcal{X}_2}$) or analogous style patterns ($\hat{\mathcal{O}}_{\mathcal{C}_1\text{-}\mathcal{X}_1}$ and $\hat{\mathcal{O}}_{\mathcal{C}_2\text{-}\mathcal{X}_1}$) if

Algorithm 1 Training Process of Language Visual Artist (LVA)

G_{E}, G_{D}: Visual Encoder, Visual Decoder
ϕ: Text Encoder
D: Patch-wise Style Discriminator
while TRAIN_VLA **do**
  $\mathcal{C}$, $\{\mathcal{S}, \mathcal{X}\} \leftarrow$ Sampled content/style

  $h^{\mathcal{C}}_{\mathcal{C}}$, $h^{\mathcal{S}}_{\mathcal{C}} \leftarrow G_{\mathrm{E}}(\mathcal{C})$ $\hat{\mathcal{C}} \leftarrow G_{\mathrm{D}}(h^{\mathcal{C}}_{\mathcal{C}}, h^{\mathcal{S}}_{\mathcal{C}})$
  $\mathcal{L}_{\mathrm{rec}} \leftarrow$ Reconstruction loss ▷ Eq. 2
  $h^{\mathcal{S}}_{\mathcal{X}} \leftarrow \phi(\mathcal{X})$ $\hat{\mathcal{O}} \leftarrow G_{\mathrm{D}}(h^{\mathcal{C}}_{\mathcal{C}}, h^{\mathcal{S}}_{\mathcal{X}})$
  $\mathcal{P}_{\mathcal{S}}$, $\mathcal{P}_{\hat{\mathcal{O}}} \leftarrow \mathrm{Crop}(\mathcal{S})$, $\mathrm{Crop}(\hat{\mathcal{O}})$
  $\mathcal{L}_{\mathrm{psd}} \leftarrow$ Patch-wise style loss ▷ Eq. 3
  $(h^{\mathcal{C}}_{\hat{\mathcal{O}}}, h^{\mathcal{S}}_{\hat{\mathcal{O}}})$, $(h^{\mathcal{C}}_{\mathcal{S}}, h^{\mathcal{S}}_{\mathcal{S}}) \leftarrow G_{\mathrm{E}}(\hat{\mathcal{O}}), G_{\mathrm{E}}(\mathcal{S})$
  $\mathcal{L}_{\mathrm{cm}} \leftarrow$ Content matching loss ▷ Eq. 4
  $\mathcal{L}_{\mathrm{sm}} \leftarrow$ Style matching loss ▷ Eq. 4

  $\mathcal{L}_G \leftarrow \mathcal{L}_{\mathrm{rec}} + \mathcal{L}_{\mathrm{psd}} + \mathcal{L}_{\mathrm{cm}} + \mathcal{L}_{\mathrm{sm}}$
  $\mathcal{L}_D \leftarrow$ Discriminator loss for D ▷ Eq. 3
  Update G_{E}, G_{D}, ϕ by minimizing $\mathcal{L}_G$
  Update D by maximizing $\mathcal{L}_D$
end while

using the same content image ($\mathcal{C}_2$) or the same style instruction ($\mathcal{X}_1$):

$$
\begin{aligned}
h^{\mathcal{C}}_{\hat{\mathcal{O}}_{\mathcal{C}_i\text{-}\mathcal{X}_j}} &= G_{\mathrm{E}}(\hat{\mathcal{O}}_{\mathcal{C}_i\text{-}\mathcal{X}_j}), \\
\mathcal{L}_{\mathrm{c}-\mathcal{C}} &= ||h^{\mathcal{C}}_{\hat{\mathcal{O}}_{\mathcal{C}_1\text{-}\mathcal{X}_1}} - h^{\mathcal{C}}_{\hat{\mathcal{O}}_{\mathcal{C}_1\text{-}\mathcal{X}_2}}||_2 + ||h^{\mathcal{C}}_{\hat{\mathcal{O}}_{\mathcal{C}_2\text{-}\mathcal{X}_1}} - h^{\mathcal{C}}_{\hat{\mathcal{O}}_{\mathcal{C}_2\text{-}\mathcal{X}_2}}||_2, \\
\mathcal{L}_{\mathrm{c}-\mathcal{S}} &= ||h^{\mathcal{S}}_{\hat{\mathcal{O}}_{\mathcal{C}_1\text{-}\mathcal{X}_1}} - h^{\mathcal{S}}_{\hat{\mathcal{S}}_{2-1}}||_2 + ||h^{\mathcal{S}}_{\hat{\mathcal{O}}_{\mathcal{C}_1\text{-}\mathcal{X}_2}} - h^{\mathcal{S}}_{\hat{\mathcal{O}}_{\mathcal{C}_2\text{-}\mathcal{X}_2}}||_2,
\end{aligned}
\tag{6}
$$

where *consistent matching* of content structure $\mathcal{L}_{\mathrm{c}-\mathcal{C}}$ or style pattern $\mathcal{L}_{\mathrm{c}-\mathcal{S}}$ is aligned by content features or style features, extracted by G_{E}.

Relative Matching. Apart from consistent matching, distinct style instructions, which imply corresponding visual semantics, should still present relative patterns. For example, we can only discover *"reach up to the sky"* literally from $\mathcal{X}_2$. If comparing reference style images $\mathcal{S}_1$ and $\mathcal{S}_2$, we can perceive the sharing of a similar style pattern and link the visual concept of *"bright tall hills"* in $\mathcal{X}_2$ to *"mountains looming over the lake"* in $\mathcal{X}_1$. We define *relative matching* $\mathcal{L}_{\mathrm{r}-\mathcal{S}}$ with the cosine similarity (CosSim) between reference style images:

$$
\begin{aligned}
(h^{\mathcal{C}}_{\mathcal{S}_i}, h^{\mathcal{S}}_{\mathcal{S}_i}) &= G_{\mathrm{E}}(\mathcal{S}_i), \\
r &= \mathrm{CosSim}(h^{\mathcal{S}}_{\mathcal{S}_1}, h^{\mathcal{S}}_{\mathcal{S}_2}), \\
\mathcal{L}_{\mathrm{r}-\mathcal{S}} &= (||h^{\mathcal{S}}_{\hat{\mathcal{O}}_{\mathcal{C}_1\text{-}\mathcal{X}_1}} - h^{\mathcal{S}}_{\hat{\mathcal{O}}_{\mathcal{C}_1\text{-}\mathcal{X}_2}}||_2 + \\
&\quad ||h^{\mathcal{S}}_{\hat{\mathcal{O}}_{\mathcal{C}_2\text{-}\mathcal{X}_1}} - h^{\mathcal{S}}_{\hat{\mathcal{O}}_{\mathcal{C}_2\text{-}\mathcal{X}_2}}||_2) \cdot r.
\end{aligned}
\tag{7}
$$

When style images are related, it has to align style features to certain extent even if paired style instructions are different. Otherwise, $\mathcal{L}_{\mathrm{r}-\mathcal{S}}$ will be close to 0 and ignore this unrelated style pair. The overall contrastive reasoning loss $\mathcal{L}_{\mathrm{ctr}}$ considers both consistent matching and relative matching:

$$
\mathcal{L}_{\mathrm{ctr}} = \mathcal{L}_{\mathrm{c}-\mathcal{C}} + \mathcal{L}_{\mathrm{c}-\mathcal{S}} + \mathcal{L}_{\mathrm{r}-\mathcal{S}}. \tag{8}
$$

3.4 Learning of CLVA

For each epoch of CLVA training, we first train with the LVA process and then CR. As algo. 1, we consider reconstruction loss $\mathcal{L}_{\text{rec}}$ to preserve content structure and patch-wise style loss $\mathcal{L}_{\text{psd}}$ between style instruction and visual pattern of transferred results. Both content matching loss $\mathcal{L}_{\text{cm}}$ and style matching loss $\mathcal{L}_{\text{sm}}$ enhance the matching with the inputs. Simultaneously, we update D by maximizing discriminator loss $\mathcal{L}_D$ to distinguish between true patches $\mathcal{P}_{\mathcal{S}}$ or false patches $\mathcal{P}_{\hat{\mathcal{O}}}$, concerning style instructions. During CR, contrastive pairs of content images and style instructions are randomly sampled, and the transferred results are across produced. We further update by minimizing contrastive reasoning loss $\mathcal{L}_{\text{ctr}}$ to allow considering content consistency and mutual style relativeness. The overall optimization of CLVA is summarized as:

$$\begin{aligned} \mathcal{L}_G =& \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{psd}} + \mathcal{L}_{\text{cm}} + \mathcal{L}_{\text{sm}}, \\ \min_{G,\phi} \max_{D} \; & \mathcal{L}_G + \mathcal{L}_D + \mathcal{L}_{\text{ctr}}. \end{aligned} \tag{9}$$

4 Experiments

4.1 Experimental Setup

Dataset. To evaluate our CLVA, we consider DTD[2] [9] and ArtEmis [1] as reference style instructions. DTD[2] contains 5K texture images with its natural descriptions for visual attributes such as colors and texture patterns. ArtEmis provides 80K artworks from WikiArt[1] with annotations of visual contents and emotional effects as human style feelings. We also collect 15K wallpapers from WallpapersCraft[2], which presents diverse scenes as content images. Each content image is resized to 256x192 in our experiment. We randomly sample 100 unseen content images and 100 testing reference styles to evaluate the generalizability of `LDAST`. Note that both style images and style instructions appear for training, but only style instructions are accessible during testing.

Evaluation Metrics. To support large-scale evaluation, we treat transferred results directly from style images as semi-groundtruth (Semi-GT) [2,4,49] by the SOTA style transfer AdaAttn [35]. We apply the following metrics:

- **SSIM** [59] compares images in the luminance, contrast, and structure aspects. A higher SSIM has a higher structural similarity;
- **Percept** [26] computes from the gram matrix of visual features. A lower Percept loss shows that two images share a similar style pattern;
- **FAD** [20] is computed by the mean L2 distance of the activations from the InceptionV3 [56] feature. As a distance metric, a lower FAD represents that `LDAST` results and Semi-GT are more relevant.

[1] WikiArt: `https://www.wikiart.org`
[2] WallpapersCraft: `https://wallpaperscraft.com/`

Note that we consider SSIM and FAD to compare with Semi-GT and calculate Percept loss directly with reference style images. Apart from visual similarity, we consider the correlation between style instructions and `LDAST` results:
– **VLS** [60] calculates the cosine similarity between each other from CLIP [53].
Since each metric has different deficiencies, we also conduct a comprehensive human evaluation from aspects of content, instruction, and style matching. We randomly sample 75 `LDAST` results and adopt MTurk[3] to rank over all methods. We also hire 3 MTurkers for each task to avoid the potential ranking bias.

Baselines. We conduct baselines for LDAST from various aspects:
– **Style Transfer**: We consider previous artistic style transfer methods NST [16], WCT [33], AdaIn [23], SANet [41], and LST [32] that support arbitrary contetn images. We use the same style (instruction and image) encoding from our CLVA as style features and follow their own training process to perform `LDAST` upon them. Due to the space issue, we only show the comparison with more recent SANet and LST. Please refer to Appendix for the complete results.
– **Language-based Image Editing**: We adopt ManiGAN [31] with affine combination module (ACM) as the general language-based editing baseline, where it modifies the content image by the style instruction. We treat normal style transferred results as groundtruth for ManiGAN to learn from.
– **CLIP-based Optimization**: StyleCLIP [43], NADA [14], and CLIPstyler [29] manipulate the content image based on the CLIP alignment of the guided instruction. Since StyleCLIP and NADA are restricted by the pre-trained generator, we compare them with the training domains of car and church. Differently, CLIPstyler can carry out arbitrary content images for `LDAST`.

4.2 Quantitative Results

Instruction with Visual Attributes. Table 1 illustrates the comparison of `LDAST` with baselines on DTD[2]. As regards automatic metrics, CLVA preserves content structures (highest 36.65 SSIM) and stylizes with related visual attributes to style images (lowest 0.2033 Percept loss). Furthermore, CLVA brings out the highest overall similarity as Semi-GT (lowest 0.1493 FAD). Since CLIPstyler directly optimizes by CLIP [53], it makes the highest VLS. Through the patch-wise discriminator, our CLVA can still produce style patterns correlated to given instructions (competitive 24.00 VLS) even without the pre-trained CLIP.

The human evaluation investigates the matching between transferred results with content images (Content), style instructions (Instruction), style images (Style), and Semi-GT (Semi-GT). In particular, content and instruction matching are the two most crucial, which concern the goal of `LDAST`: *content structure preservation* and *style pattern presentation*; style image and semi-gt matching are provided for different comparing targets from a human aspect. The results are calculated by the mean ranking score (from 1 to 5, the higher is better) of each method. In general, MTurkers indicate that our CLVA has an apparent advantage in preserving content structures (highest 3.852 Content) and presenting aligned

[3] Amazon Mechanical Turk: `https://www.mturk.com`

Method	Automatic Metrics				Human Evaluation			
	SSIM↑	Percept↓	FAD↓	VLS↑	Content↑	Instruction↑	Style↑	Semi-GT↑
SANet [41]	35.50	0.2129	0.1627	23.57	2.701	2.477	2.738	2.630
LST [32]	34.84	0.2129	0.1533	23.16	2.743	2.831	2.651	2.528
ManiGAN [31]	32.70	0.2401	0.1663	23.25	2.757	2.562	2.937	2.922
CLIPstyler [29]	25.24	0.2598	0.1818	**24.62**	2.948	3.388	3.073	3.265
CLVA	**36.65**	**0.2033**	**0.1493**	24.00	**3.852**	**3.742**	**3.603**	**3.655**

Table 1. Testing results of `LDAST` using visual attribute instructions on DTD[2].

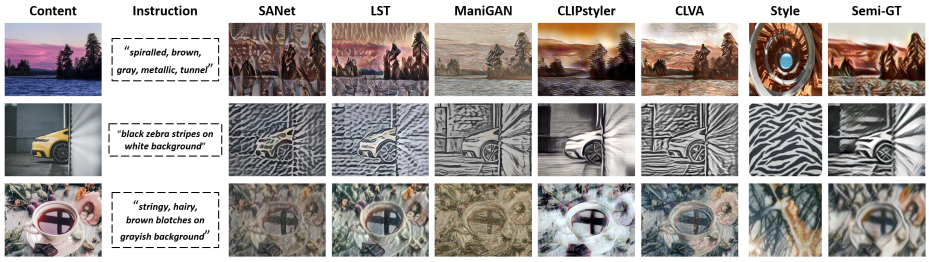

Fig. 3. Visualized comparison using visual attribute instructions on DTD[2].

style patterns (highest 3.742 Instruction). Though with the aid of CLIP, CLIPstyler is still behind CLVA (-0.4 Instruction), with an even higher gap in style image matching (-0.5 Style). Contributed by contrastive reasoning that compares the mutual relativeness between pairs of contents and instructions, CLVA can stylize with the captured visual attributes. We adopt Pearson correlation and investigate the coefficients between automatic metrics and human evaluation as 77.2 (FAD→Instruction), 84.5 (FAD→Semi-GT), 81.3 (VLS→Instruction), and 77.8 (VLS→Semi-GT). This high correlation indicates that our metric design is adequate for evaluating large-scale LDAST experiments. The even higher 88.2 correlation (Instruction→Semi-GT) between instruction and Semi-GT matching in human evaluation further supports the usage of Semi-GT.

From the aspect of visualized comparison in Fig. 3, previous SANet and LST only produce repetitive and disorder textures in their transferred results. ManiGAN modifies the style directly over pixels, suffering from blurring objects; this deficiency can also be found in Table 1 (lower SSIM and lower Content matching). CLIPstyler is sometimes misguided by CLIP, making irrelevant patterns, such as the bright white background in the third case. Contrary to baselines, CLVA extracts a more detailed style from different kinds of guidance (*"brown metallic"* in the first row and *"stringy hairy"* in the third case), leading to superior `LDAST` results that correspond to style instructions.

Instruction with Emotional Effects. Unlike visual attributes, emotional effect instructions are more challenging as connecting to visual semantics of described objects or style patterns from human feelings. For example, "*yellowish and green*" from "*sunset and mountains*" or "*scaring charcoal grey*" from "*nightmare*". We consider this human style feeling on ArtEmis [1], where the model has to express the latent visual concepts of emotional effect instructions. CLVA performs with more balance (both second-highest SSIM and second-lowest Per-

Method	Automatic Metrics				Human Evaluation			
	SSIM↑	Percept↓	FAD↓	VLS↑	Content↑	Instruction↑	Style↑	Semi-GT↑
SANet [41]	38.36	**0.0352**	0.1548	19.30	3.170	2.978	2.980	2.890
LST [32]	**42.13**	0.0386	0.1595	19.92	2.967	2.714	2.614	2.757
ManiGAN [31]	38.46	0.0500	0.1554	19.69	2.729	2.583	2.879	3.192
CLIPstyler [29]	24.17	0.0659	0.1759	**21.04**	2.777	3.140	2.998	2.952
CLVA	40.32	0.0357	**0.1418**	20.11	**3.357**	**3.586**	**3.530**	**3.208**

Table 2. Testing results of LDAST using emotional effect instructions on ArtEmis.

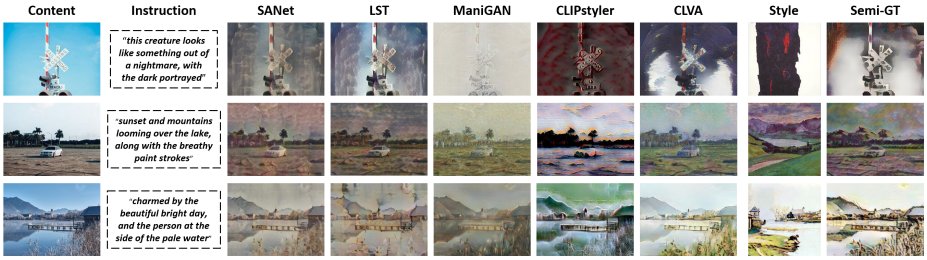

Fig. 4. Visualized comparison using emotional effect instructions on ArtEmis.

cept) from Table 2, especially the lowest 0.1418 FAD, making the most similar transferred results to Semi-GT. Though CLIPstyler [29] achieves higher VLS by optimizating over CLIP, from human aspects, CLVA can preserve more concrete contents and present more correlated style patterns (higher 3.357 content and 3.586 instruction matching). The visualized comparison in Fig. 4 illustrates that previous SANet [41] and LST [32] contain unsmooth and fragmentary patterns with blurring contents. Without a style transformation process, ManiGAN [31] modifies with only monotonous colors. CLIPstyler is failed to capture human style feelings well, suffering from weird and unpleasant results. Different from them, our CLVA learns the visual semantic during contrastive reasoning by comparing mutual relativeness between literal instructions and style images, leading to a more colorful and corresponding stylization as human emotion. More surprisingly, despite not instructed literally, CLVA perceives "*side of the water*" and reveals the latent yet correlated "*grassland*" precisely in the third row.

Specific Content Domain. To compare with StyleCLIP [43] and NADA [14] that are restricted by the pre-trained generator, we evaluate LDAST on the specific content domain. We consider the same domain images in StyleGAN2 [27] and visual attribute instructions on DTD[2]. Table 3 indicates the numerical comparison on Car and Church. Our CLVA still produces superior results and is the most admirable by human. Since StyleCLIP and NADA rely on StyleGAN, they can only preserve content (highest 3.459 Content by StyleCLIP) but with limited stylization (lower Instruction and Style). Similar observations can be found in Fig. 5, where StyleCLIP shows almost no modification for the second car. They can neither deal with the background; NADA even destroys the scene in the third row. In contrast to CLIPstyler [29] that only contains abstractive and obscure styles, CLVA presents the detailed *"read interplaced cloth"* behind the car and the color *"cream"* precisely on the surface of the church.

Method	Automatic Metrics				Human Evaluation			
	SSIM↑	Percept↓	FAD↓	VLS↑	Content↑	Instruction↑	Style↑	Semi-GT↑
ManiGAN [31]	26.45	0.2329	0.1672	23.44	2.861	2.894	2.978	2.893
StyleCLIP [43]	28.03	0.2609	0.1812	21.55	**3.459**	2.845	2.930	2.829
NADA [14]	16.98	0.2733	0.1876	23.38	2.542	2.798	2.846	2.932
CLIPstyler [29]	18.43	0.2493	0.1826	**24.16**	2.986	3.067	3.003	3.032
CLVA	**30.98**	**0.1957**	**0.1544**	23.68	3.153	**3.465**	**3.344**	**3.315**

Table 3. Testing results of LDAST on specific content domain (Car and Church).

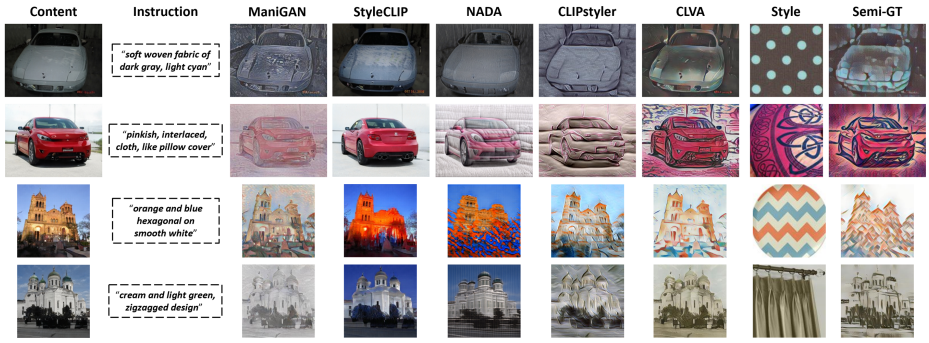

Fig. 5. Visualized comparison on specific content domain (Car and Church).

4.3 Ablation Study

We conduct an ablation study of each component effect on DTD[2] in Table 4. At row (a), with the reconstruction $\mathcal{L}_{\text{rec}}$ and the patch-wise style $\mathcal{L}_{\text{psd}}$, CLVA achieves feasible LDAST results by concrete structures and extracted style semantics. Row (b)-(d) shows the strength of content matching $\mathcal{L}_{\text{cm}}$ and style matching $\mathcal{L}_{\text{sm}}$. In particular, content matching helps the structure similarity to content images (higher 36.05 SSIM). Style matching aims at analogous visual patterns to style images, which leads to better stylization quality (lower 0.2049 Percept and higher 23.69 VLS). If considering altogether, it can benefit and strike a balance between both. Finally, contrastive reasoning $\mathcal{L}_{\text{ctr}}$ further enables CLVA to consider contrastive pairs, making a comprehensive improvement at row (e).

Why CLVA is better than CLIP-based? Despite no CLIP optimized, CLVA demonstrates superior results on LDAST with all aspects of automatic metrics and human evaluation. To investigate it, we conduct instruction-to-style retrieval based on the similarity between features of style instructions and style images. Table 5 shows that our learned CLVA performs higher Recall@k on both DTD[2] and ArtEmis, leading to a better instruction-style alignment than the used CLIP. The visualization in Fig. 6 also indicates the flaw of CLIP on detailed style patterns. For example, in the first row, CLIP only presents either *"bright color"* or *"town"* in the retrieval results. In contrast, CLVA can capture both and present more related LDAST to *"happy place to live"*. From Table 6, even CLIP has been fine-tuned ahead; our CLVA still produces preferrable LDAST results from all human aspects of content, instruction, and style matching. This observation supports that contrastive reasoning, which considers contrastive pairs of content images and style instructions, is required to benefit from mutual relativeness.

	Ablation Settings				Automatic Metrics			
	$\mathcal{L}_{\text{rec}}$+$\mathcal{L}_{\text{psd}}$	$\mathcal{L}_{\text{cm}}$	$\mathcal{L}_{\text{sm}}$	$\mathcal{L}_{\text{ctr}}$	SSIM↑	Percept↓	FAD↓	VLS↑
(a)	✓	✗	✗	✗	34.73	0.2290	0.1568	23.29
(b)	✓	✓	✗	✗	<u>36.05</u>	0.2304	0.1512	23.27
(c)	✓	✗	✓	✗	35.73	<u>0.2049</u>	<u>0.1508</u>	<u>23.69</u>
(d)	✓	✓	✓	✗	35.86	0.2100	0.1499	23.54
(e)	✓	✓	✓	✓	**36.65**	**0.2033**	**0.1493**	**24.00**

Table 4. Ablation study of CLVA using visual attribute instructions on DTD^2.

	DTD^2		ArtEmis	
Method	R@1	R@5	R@1	R@5
CLIP [53]	13.9	30.7	9.8	20.7
CLVA	**19.3**	**45.1**	**13.9**	**30.7**

Table 5. Instruction-to-style retrieval on DTD^2 and ArtEmis.

	Human Evaluation			
Method	Content↑	Instruction↑	Style↑	Semi-GT↑
CLIPstyler (ft.)	1.208	1.347	1.292	1.333
CLVA	**1.792**	**1.653**	**1.708**	**1.667**

Table 6. Human comparison between CLVA and CLIPstyle with fine-tuned CLIP on DTD^2.

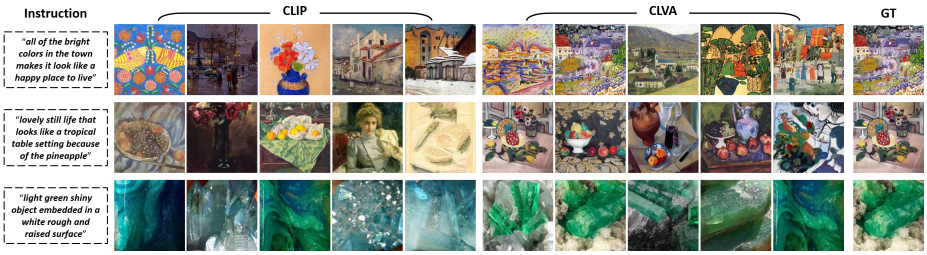

Fig. 6. Visualization examples of instruction-to-style retrieval by CLIP and CLVA.

Apart from transfer quality, CLVA also holds a higher efficiency than CLIP-based methods. Table 7 illustrates the time and GPU cost on a single TITAN X (12GB) with content image size 256x192. All CLIP-based methods take more than 30 seconds for only one pair of content images and style instructions. Instead of numerous iterations to align with CLIP, we extract style semantics and carry out `LDAST` in one shot, taking merely 0.03 seconds for one input. Without updating the model during inference, CLVA supports parallelization and can accomplish 50 pairs in half a second. Besides, as a lightweight style transfer network, CLVA requires the least GPU memory for `LDAST`. In summary, our CLVA surpasses those CLIP-based methods on both quality and efficiency because of the detailed style deficiency and the required optimizing iteration from CLIP.

Qualitative Results. As shown in Fig. 7, we investigate the linear interpolation of extracted style patterns by CLVA. Considering style features $h^{\mathcal{S}}_{\mathcal{X}_1}$ and $h^{\mathcal{S}}_{\mathcal{X}_2}$ of instructions $\mathcal{X}_1$ and $\mathcal{X}_2$, the interpolated $h^{\mathcal{S}}_{\text{p}}$ should be:

$$h^{\mathcal{S}}_{\text{p}} = (1-\alpha)h^{\mathcal{S}}_{\mathcal{X}_1} + \alpha h^{\mathcal{S}}_{\mathcal{X}_2}, \tag{10}$$

where α is the style ratio between the two. Fig. 7 presents a smooth transformation from one style instruction to another. By training on DTD^2 and ArtEmis altogether, CLVA even performs interpolated stylization by both visual attribute and emotional effect instructions in the third row. Fig. 8 illustrates diverse `LDAST`

Method	Time (sec)			GPU (MB)		
	BS=1	32	50	BS=1	32	50
ManiGAN [31]	0.079	0.533	1.148	3312	6572	8129
StyleCLIP [43]	32.38	*	*	4149	*	*
NADA [14]	63.49	*	*	6413	*	*
CLIPstyler [29]	99.98	*	*	5429	*	*
CLVA	**0.029**	**0.246**	**0.405**	**1525**	**3207**	**4441**

Table 7. Time and GPU cost when performing LDAST on TITAN X with content image size 256x192. * means this method can only run one input at a time.

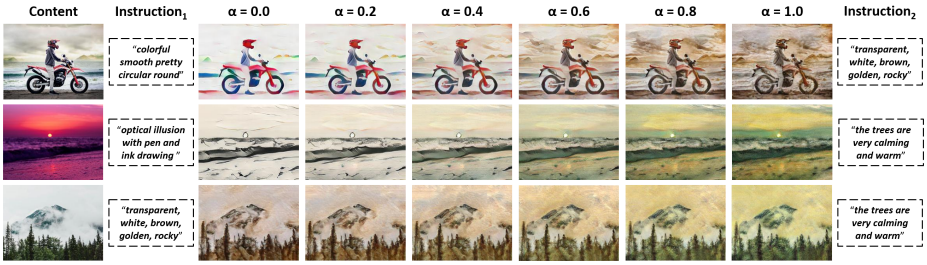

Fig. 7. Style interpolation results of LDAST over instructions.

results by our CLVA. Since CLVA supports arbitrary content images, we can also modify the style detail for high-resolution inputs in Fig. 9.

5 Conclusion

We introduce language-driven artistic style transfer (LDAST) to do stylization for a content image by a style instruction. We propose contrastive language visual artist (CLVA) that adopts the patch-wise style discriminator and contrastive reasoning to jointly learn between style images and style instructions. We demonstrate that CLVA can express various style patterns of visual attributes as well as emotional effects and perform LDAST efficiently. CLVA also outperforms baselines on both automatic metrics and human evaluation. We believe that LDAST can make visual applications like image/video effect more controllable for humans.

Acknowledgments. Research was sponsored by the U.S. Army Research Office and was accomplished under Contract Number W911NF-19-D-0001 for the Institute for Collaborative Biotechnologies. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

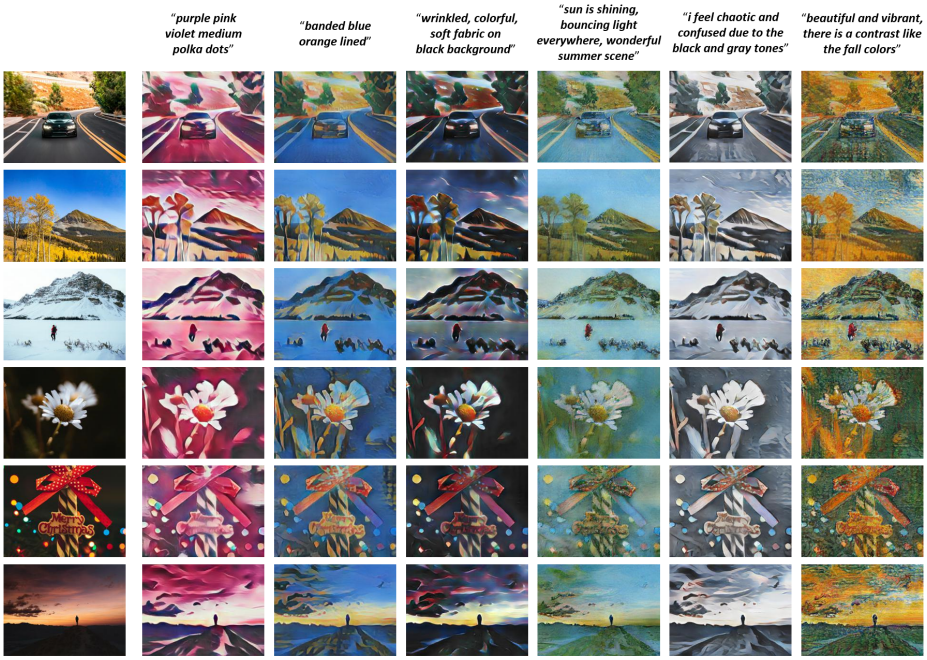

Fig. 8. CLVA results on diverse pairs of content images and style instructions.

Fig. 9. High-resolution (1920x1080) `LDAST` results by CLVA with upper right: *"the lonely world makes me feel scared and nostalgic how sky and sea merge together"*; lower left: *"the snow and lights in the shop windows looks like a winter scene"*; lower right: *"ink painting, black dotted line, whiteboard"*.

A Implementation Detail

We adopt VGG-19 [54,41] as our visual encoder G_{E} and visual decoder G_{D}. Text encoder ϕ first adopts RoBERTa [37,47] for a general linguistic and then expands its spatial dimension to jointly embed with style features. We follow the self-attention layer from SANet [41] to fuse between content and style features in G_{D}. The patch-wise style discriminator D contains a similar architecture with a dense layer to determine the correlation between instructions and image patches. Both G_{E} and G_{D} are initialized from SANet and further update during the CLVA training process. We adopt Adam [28] to optimize CLVA with learning rate 3e-4 for $\mathcal{L}_G$, 1e-4 for $\mathcal{L}_D$, and 3e-5 for $\mathcal{L}_{\mathrm{ctr}}$.

B Complete Results on all Baselines

Table 8 shows the complete `LDAST` results using visual attribute instructions on DTD^2 [9]. As with previous style transfer methods, NST [16] and WCT [33] cannot handle the target style well (higher Percept loss), resulting in the irrelevant produced patterns to style instructions (lower VLS). AdaIn [23] better performs stylization from guided texts, but the content will be much modified with a relatively lower SSIM. Similar observations can be found in Fig. 10, where the scissors and color pencils by AdaIn are mostly distorted in the second row. In contrast, CLVA presents the detailed multi-color pattern (the second case) yet preserves the concrete structure of the flower (the third case) at the same time.

We also illustrate the results of SANet [41] and LST [32] on specific content domain (Car and Church) in Table 9. By learning from pairs of style instructions and images, style transfer methods can perform better stylization than StyleGAN-restricted methods (StyleCLIP [43] and NADA [14]). Despite having similar results on automatic metrics, there is a noticeable quality gap compared to our CLVA in Fig. 11. Unlike SANet and LST, which contain repetitive and chaotic patterns, CLVA accomplishes `LDAST` with even more concrete contents (the second car and the fourth church) than Semi-GT through consistent matching during contrastive reasoning.

Method	Automatic Metrics			
	SSIM↑	Percept↓	FAD↓	VLS↑
NST [16]	19.70	0.2441	0.1920	18.97
WCT [33]	**37.88**	0.2617	0.1720	19.76
AdaIn [23]	29.43	<u>0.2081</u>	0.1748	21.65
SANet [41]	35.50	0.2129	0.1627	23.57
LST [32]	34.84	0.2137	<u>0.1533</u>	23.16
ManiGAN [31]	32.7	0.2401	0.1663	23.25
CLIPstyler [29]	25.24	0.2598	0.1818	**24.62**
CLVA	<u>36.65</u>	**0.2033**	**0.1493**	<u>24.00</u>

Table 8. Complete results using visual attribute instructions on DTD^2.

Method	Automatic Metrics			
	SSIM↑	Percept↓	FAD↓	VLS↑
SANet [41]	30.95	<u>0.1982</u>	0.1638	23.20
LST [32]	**31.16**	0.2045	<u>0.1606</u>	23.34
ManiGAN [31]	26.45	0.2329	0.1672	23.44
StyleCLIP [43]	28.03	0.2609	0.1812	21.55
NADA [14]	16.98	0.2733	0.1876	23.38
CLIPstyler [29]	18.43	0.2493	0.1826	**24.16**
CLVA	<u>30.98</u>	**0.1957**	**0.1544**	<u>23.68</u>

Table 9. Complete results on specific content domain (Car and Church).

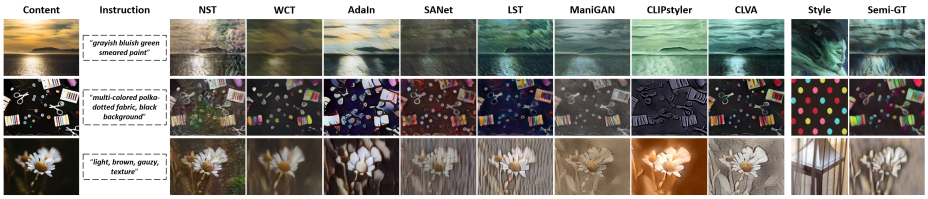

Fig. 10. Complete visualization using visual attribute instructions on DTD[2].

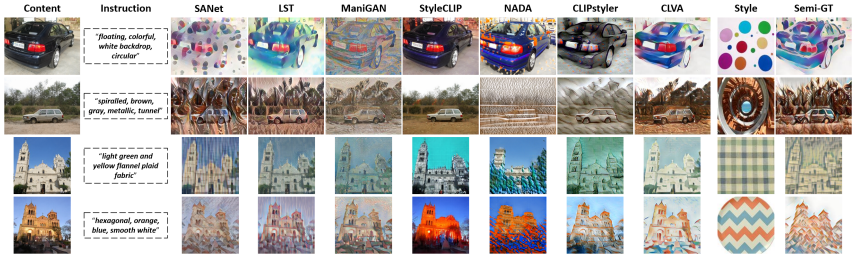

Fig. 11. Complete visualization on specific content domain (Car and Church).

C Retrieval-based Baseline

Apart from producing the transferred result by the style instruction $\mathcal{X}$, we also investigate a two-step retrieval-based baseline. We first adopt the CLIP [53] alignment to find the most similar style image $\mathcal{S}$ (with 13.9 R@1 and 30.7 R@5 on DTD[2] [9]) via $\mathcal{X}$. Then, the retrieved $\mathcal{S}$ is used to carry out standard style transfer. Table 10 shows that the two-step retrieval baseline performs slightly better on visual similarity to Semi-GT. However, by stylization from guided texts, CLVA produces more correlated style patterns to instructions (higher 24.00 VLS). In addition, this retrieval-based method still relies on an existing set of style images and limits the diversity of stylization due to the collection size.

	Automatic Metrics			
Method	SSIM↑	Percept↓	FAD↓	VLS↑
SANet [41]	35.50	0.2129	0.1627	23.57
SANet (rtrv.)	**37.74**	**0.2005**	**0.1421**	<u>23.68</u>
CLVA	<u>36.65</u>	<u>0.2033</u>	<u>0.1493</u>	**24.00**

Table 10. Testing results of the two-step retrieval-based baseline using visual attribute instructions on DTD[2].

D Human Evaluation

We investigate the quality of `LDAST` results from the human aspect through Amazon Mechanical Turk. Fig. 12 illustrates the screenshots of the human tasks. MTurkers rank the correlation of the `LDAST` result according to Content, Instruction, Style, and Semi-GT matching. Each MTurker rewards $2.0 and takes a mean of 15 minutes.

Fig. 12. The screenshots of the ranking tasks for human evaluation on `LDAST`.

E Limitation and Ethics Discussion

Though our work benefits creative visual applications, there are several remaining technical issues. At first, complicated instructions that contain excessive visual attributes or emotional effects are still difficult to address by our CLVA. CLVA may lean towards specific visual concepts, resulting in correlated but monotonous stylization. Secondly, since the learning of CLVA relies on patch-wise style discriminator D, the quality of the randomly sampled patches will crucially influence the transferred results. On the other hand, there may be a "fake as real" doubt for those manipulated content images. To mitigate this issue, we can apply techniques from image forensics [58,24,11] to detect the authenticity of an image. Regarding guided instructions, for example, hate speech detection [3,22,51,50] can help to filter out malicious texts and prevent from producing controversial results with ethics concerns.

References

1. Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M., Guibas, L.: ArtEmis: Affective Language for Visual Art. In: CVPR (2021)
2. Al-Sarraf, A., Shin, B.S., Xu, Z., Klette, R.: Ground Truth and Performance Evaluation of Lane Border Detection. In: ICCVG (2014)
3. Aluru, S.S., Mathew, B., Saha, P., Mukherjee, A.: Deep Learning Models for Multilingual Hate Speech Detection. In: ECML-PKDD (2020)
4. Borkar, A., Hayes, M., Smith, M.T.: An Efficient Method to Generate Ground Truth for Evaluating Lane Detection Systems. In: ICASSP (2010)
5. Chen, H., Zhao, L., Wang, Z., Zhang, H., Zuo, Z., Li, A., Xing, W., Lu, D.: DualAST: Dual Style-Learning Networks for Artistic Style Transfer. In: CVPR (2021)

6. Chen, J., Shen, Y., Gao, J., Liu, J., Liu, X.: Language-Based Image Editing with Recurrent Attentive Models. In: CVPR (2018)
7. Chen, Y.L., Hsu, C.T.: Towards Deep Style Transfer: A Content-Aware Perspective. In: BMVC (2016)
8. Cheng, M.M., Zheng, S., Lin, W.Y., Warrell, J., Vineet, V., Sturgess, P., Crook, N., Mitra, N., Torr, P.: ImageSpirit: Verbal Guided Image Parsing. In: ACM Transactions on Graphics (2013)
9. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing Textures in the Wild. In: CVPR (2014)
10. El-Nouby, A., Sharma, S., Schulz, H., Hjelm, D., Asri, L.E., Kahou, S.E., Bengio, Y., W.Taylor, G.: Tell, Draw, and Repeat: Generating and Modifying Images Based on Continual Linguistic Instruction. In: ICCV (2019)
11. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging Frequency Analysis for Deep Fake Image Recognition. In: ICML (2020)
12. Fu, T.J., Wang, X.E., Grafton, S., Eckstein, M., Wang, W.Y.: SSCR: Iterative Language-Based Image Editing via Self-Supervised Counterfactual Reasoning. In: EMNLP (2020)
13. Fu, T.J., Wang, X.E., Grafton, S., Eckstein, M., Wang, W.Y.: M3L: Language-based Video Editing via Multi-Modal Multi-Level Transformer. In: CVPR (2022)
14. Gal, R., Patashnik, O., Maron, H., Chechik, G., Cohen-Or, D.: StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. In: arXiv:2108.00946 (2021)
15. Gao, C., Gu, D., Zhang, F., Yu, Y.: ReCoNet: Real-time Coherent Video Style Transfer Network. In: arXiv:1807.01197 (2018)
16. Gatys, L.A., Ecker, A.S., Bethge, M.: A Neural Algorithm of Artistic Style. In: arXiv:1508.06576 (2015)
17. Gatys, L.A., Ecker, A.S., Bethge, M.: Texture Synthesis Using Convolutional Neural Networks. In: NeurIPS (2015)
18. Gatys, L.A., Ecker, A.S., Bethge, M., Hertzmann, A., Shechtman, E.: Controlling Perceptual Factors in Neural Style Transfer. In: CVPR (2017)
19. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Networks. In: NeurIPS (2014)
20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: NeurIPS (2017)
21. Huang, H., Wang, H., Luo, W., Ma, L., Jiang, W., Zhu, X., Li, Z., Liu, W.: Real-Time Neural Style Transfer for Videos. In: CVPR (2017)
22. Huang, X., Xing, L., Dernoncourt, F., Paul, M.J.: Multilingual Twitter Corpus and Baselines for Evaluating Demographic Bias in Hate Speech Recognition. In: LREC (2020)
23. Huang, X., Belongie, S.: Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. In: ICCV (2017)
24. Huh, M., Liu, A., Owens, A., Efros, A.A.: Fighting Fake News: Image Splice Detection via Learned Self-Consistency. In: ECCV (2018)
25. Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y., Song, M.: Neural Style Transfer: A Review. In: arXiv:1705.04058 (2017)
26. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In: ECCV (2016)
27. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and Improving the Image Quality of StyleGAN. In: CVPR (2020)

28. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: ICLR (2015)
29. Kwon, G., Ye, J.C.: CLIPstyler: Image Style Transfer with a Single Text Condition. In: CVPR (2022)
30. Laput, G., Dontcheva, M., Wilensky, G., Chang, W., Agarwala, A., Linder, J., Adar, E.: PixelTone: A Multimodal Interface for Image Editing. In: CHI (2013)
31. Li, B., Qi, X., Lukasiewicz, T., Torr, P.H.S.: ManiGAN: Text-Guided Image Manipulation. In: CVPR (2020)
32. Li, X., Liu, S., Kautz, J., Yang, M.H.: Learning Linear Transformations for Fast Arbitrary Style Transfer. In: CVPR (2019)
33. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal Style Transfer via Feature Transforms. In: NeurIPS (2017)
34. Li, Y., Liu, M.Y., Li, X., Yang, M.H., Kautz, J.: A Closed-form Solution to Photorealistic Image Stylization. In: ECCV (2018)
35. Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., Sun, Z., Li, Q., Ding, E.: AdaAttN: Revisit Attention Mechanism in Arbitrary Neural Style Transfer. In: ICCV (2021)
36. Liu, X., Lin, Z., Zhang, J., Zhao, H., Tran, Q., Wang, X., Li, H.: Open-Edit: Open-Domain Image Manipulation with Open-Vocabulary Instructions. In: ECCV (2020)
37. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. In: arXiv:1907.11692 (2019)
38. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep Photo Style Transfer. In: CVPR (2017)
39. Nam, S., Kim, Y., Kim, S.J.: Text-Adaptive Generative Adversarial Networks: Manipulating Images with Natural Language. In: NeurIPS (2018)
40. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In: arXiv:2112.10741 (2021)
41. Park, D.Y., Lee, K.H.: Arbitrary Style Transfer with Style-Attentional Networks. In: CVPR (2019)
42. Park, T., Zhu, J.Y., Wang, O., Lu, J., Shechtman, E., Efros, A.A., Zhang, R.: Swapping Autoencoder for Deep Image Manipulation. In: NeurIPS (2020)
43. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In: ICCV (2021)
44. Qiao, T., Zhang, J., Xu, D., Tao, D.: MirrorGAN: Learning Text-to-image Generation by Redescription. In: CVPR (2019)
45. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-Shot Text-to-Image Generation. In: arXiv:2102.12092 (2021)
46. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative Adversarial Text to Image Synthesis. In: ICML (2016)
47. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: EMNLP (2019)
48. Salehi, P., Chalechale, A., Taghizadeh, M.: Generative Adversarial Networks (GANs): An Overview of Theoretical Model, Evaluation Metrics, and Recent Developments. In: arXiv:2005.13178 (2020)
49. Salvo, R.D.: Large Scale Ground Truth Generation for Performance Evaluation of Computer Vision Methods. In: VIGTA (2013)
50. Samanta, B., Ganguly, N., Chakrabarti, S.: Improved Sentiment Detection via Label Transfer from Monolingual to Synthetic Code-Switched Text. In: ACL (2019)

51. Samghabadi, N.S., Patwa, P., PYKL, S., Mukherjee, P., Das, A., Solorio, T.: Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In: LREC (2020)
52. Sanakoyeu, A., Kotovenko, D., Lang, S., Ommer, B.: A Style-Aware Content Loss for Real-time HD Style Transfer. In: ECCV (2018)
53. Shi, L., Shuang, K., Geng, S., Su, P., Jiang, Z., Gao, P., Fu, Z., de Melo, G., Su, S.: Contrastive Visual-Linguistic Pretraining. In: arXiv:2007.13135 (2020)
54. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: ICLR (2015)
55. Somavarapu, N., Ma, C.Y., Kira, Z.: Frustratingly Simple Domain Generalization via Image Stylization. In: arXiv:2006.11207 (2020)
56. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the Inception Architecture for Computer Vision. In: CVPR (2016)
57. Wang, P., Li, Y., Vasconcelos, N.: Rethinking and Improving the Robustness of Image Style Transfer. In: CVPR (2021)
58. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: CNN-Generated Images are Surprisingly Easy to Spot...for Now. In: CVPR (2020)
59. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncel, E.P.: Image Quality Assessment: From Error Visibility to Structural Similarity. In: TIP (2004)
60. Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., Duan, N.: GODIVA: Generating Open-DomaIn Videos from nAtural Descriptions. In: arXiv:2104.14806 (2021)
61. Wu, C., Timm, M., Maji, S.: Describing Textures using Natural Language. In: ECCV (2020)
62. Wu, L., Wang, Y., Shao, L.: Cycle-Consistent Deep Generative Hashing for Cross-Modal Retrieval. In: TIP (2018)
63. Xia, W., Yang, Y., Xue, J.H., Wu, B.: TediGAN: Text-Guided Diverse Face Image Generation and Manipulation. In: CVPR (2021)
64. Xian, W., Sangkloy, P., Agrawal, V., Raj, A., Lu, J., Fang, C., Yu, F., Hays, J.: TextureGAN: Controlling Deep Image Synthesis with Texture Patches. In: CVPR (2018)
65. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., Hes, X.: AttnGAN: FineGrained Text to Image Generation with Attentional Generative Adversarial Networks. In: CVPR (2018)
66. Yi, Z., Zhang, H., Tan, P., Gong, M.: DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In: ICCV (2017)
67. Yoo, J., Uh, Y., Chun, S., Kang, B., Ha, J.W.: Photorealistic Style Transfer via Wavelet Transforms. In: ICCV (2019)
68. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-Attention Generative Adversarial Networks. In: PMLR (2019)
69. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.: StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. In: ICCV (2017)
70. Zhang, Z., Wang, Z., Lin, Z., Qi, H.: Image Super-Resolution by Neural Texture Transfer. In: CVPR (2019)
71. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In: ICCV (2017)