Only-Style: Stylistic Consistency in Image Generation without Content Leakage

Tilemachos Aravanis[1] Panagiotis Filntisis[2,3] Petros Maragos[1, 2, 3] George Retsinas[2, 3]
[1]School of Electrical & Computer Engineering, National Technical University of Athens, Greece
[2]Robotics Institute, Athena Research Center, 15125 Maroussi, Greece
[3]HERON - Center of Excellence in Robotics, Athens, Greece

Abstract

Generating images in a consistent reference visual style remains a challenging computer vision task. State-of-the-art methods aiming for style-consistent generation struggle to effectively separate semantic content from stylistic elements, leading to content leakage from the image provided as a reference to the targets. To address this challenge, we propose Only-Style: a method designed to mitigate content leakage in a semantically coherent manner while preserving stylistic consistency. Only-Style works by localizing content leakage during inference, allowing the adaptive tuning of a parameter that controls the style alignment process, specifically within the image patches containing the subject in the reference image. This adaptive process best balances stylistic consistency with leakage elimination. Moreover, the localization of content leakage can function as a standalone component, given a reference-target image pair, allowing the adaptive tuning of any method-specific parameter that provides control over the impact of the stylistic reference. In addition, we propose a novel evaluation framework to quantify the success of style-consistent generations in avoiding undesired content leakage. Our approach demonstrates a significant improvement over state-of-the-art methods through extensive evaluation across diverse instances, consistently achieving robust stylistic consistency without undesired content leakage. Project-Page

1. Introduction

State-of-the-art text-to-image (T2I) models [3, 7, 34, 36, 40] demonstrate impressive results in transforming text into compelling visual outputs. However, such models do not hand the user control over specific visual stylistic results, often producing widely varying interpretations of the same textual descriptor, as shown in the first row of Fig. 1.

For this reason, several works aim to provide visual thematic consistency across different generated concepts, as they were created or performed in the same manner or technique (e.g., from the same artist). Even though these methods achieve the desired stylistic alignment of a reference image with a target one, they frequently exhibit content leakage. In other words, unintended semantic elements from the reference image subject, appear in the target image. Such cases are evident in the second row of Fig. 1.

Figure 1. **Only-Style**: The top row shows images independently generated by a text-to-image model using the style descriptor *"in vintage poster style"*. Applying a state-of-the-art method (here StyleAligned [13]) to align these images stylistically with the first image (the car) leads to unintended content leakage, causing visual elements of the car to infiltrate the other images. *Only-Style* addresses this by first localizing the semantic content of the reference subject in the target images (third row) and then guiding the alignment process to eliminate this undesired effect (fourth row).

The main motivation of this work is to find a semantically meaningful way to remove this content leakage while preserving consistency in style. To achieve this, we introduce a method to control the transfer of the reference subject patches in the target image, through a simple scaling of their representations. This scaling operation is adaptively tuned via the localization of the reference subject in the target image to determine whether the transfer of the reference subject features should be further restrained (see third row of Fig. 1). The final stylistic alignment is illustrated in the bottom row of Fig. 1. As it can be observed, the resulting target images exhibit no content leakage - thus we dubbed our method *Only-Style*.

Additionally, we present two novel evaluation methods to measure the impact of content leakage in style consistent image generation, both coarsely via an encoder-based metric (CL) and fine-grained via Large Vision-Language Models (LVLMs). The former quantifies the semantic correlation between the target image and the reference subject, thereby assessing content leakage, while the latter detects even subtle leakage cases by prompting LVLMs. We benchmark many state-of-the-art methods, showing that content leakage is a main challenge shared across all of them. To our knowledge, although content leakage has been recognized as an issue in stylistic consistency, no prior work has proposed a way to measure the appearance of content leakage cases — a very useful tool to understand the efficacy of methods that promote stylistic consistency.

Our main contributions are: • We introduce fine-grained control over the leakage of reference semantic elements to the target image, tailored for attention sharing approaches [13]. • We propose a novel method for subtle leakage localization, applicable to any style consistent generation approach. • We design an end-to-end method that achieves style-consistent image generation by adaptively scaling down the contribution of the reference subject, addressing the problem of content leakage. • We propose an evaluation framework that quantifies content leakage using two distinct metrics (coarse and fine-grained version). We release this benchmark to serve as a standardized baseline, addressing a notable gap in the community and facilitating more consistent comparisons of style alignment methods.

2. Related Work

Text-to-Image diffusion models. Diffusion models [15, 41, 43, 44] have transformed the field of image generation, producing highly diverse and visually striking outputs. Further, text-conditioned diffusion models [30, 37] enable the generation process to be guided by natural language prompts, leveraging these powerful generative capabilities.

Controlling the attention in T2I diffusion models. The attention mechanism is the common underlying ingredient within neural network backbones in T2I diffusion models. Recent works have explored how the self-attention and cross-attention layers can be harnessed to define both the layout and semantic content of text-generated images [4, 12, 32, 45]. Additionally, attention mechanisms have been widely applied for editing text-generated images [1, 6, 29, 33, 46]. Building on insights from these methods, we utilize the attention layers to disentangle content and style, addressing the challenge of style-consistent generation.

Style Transfer. Style transfer is a long-standing challenge in computer vision [5, 14] that refers to the process of transforming the visual style of an input image while preserving its content. Neural Style Transfer leverages deep features from pretrained networks to alter the style of a target image based on a reference [10, 20]. Moreover, GAN-based techniques have been developed to transfer images across different stylistic domains [18, 21, 31, 52].

Consistent style Generation. With the advent of diffusion models stylization research has focused on generating target text-specified concepts using either real or synthesized stylistic reference images. Different approaches have emerged to tackle this task: • One family of approaches involves training diffusion models to incorporate conditioning from the output representation of a pretrained image encoder[48–51], such as CLIP[35]. However these methods require significant computational resources to train this conditioning and tend to drive the model away from its training distribution. Following this paradigm and conceptually close to our work, InstantStyle [48], in a coarse attempt to reduce content leakage, injects the CLIP image embedding of the stylistic reference, subtracted by the CLIP text embedding of the reference subject, into specific blocks within the diffusion model. • Another line of recent works developed optimization techniques over one or more images that let the model capture certain visual features [8, 9, 11, 22, 39, 42] such as a style interpretation. For example B-LoRA [8] trains specific LoRAs within the diffusion backbone to capture separately the content and style of an image. • Closer to our work, to circumvent the computationally intensive pretraining or fine-tuning process per instance, recent approaches utilize self-attention layers of the model's backbone to allow communication between images within a batch, and thus the transfer of stylistic features from a single reference to other images [13, 19]. Building upon these state-of-the-art approaches, our method addresses the persistent challenge of content leakage.

3. Proposed Method: Only-Style

3.1. Overview of Only-Style

In the forthcoming analysis, we consider the following setup[1]: • The goal is the generation of two images I_{ref} and

[1] Our method can be easily extended to support multi-image and multi-subject generation (see Suppl. Mat. and Fig. 7)

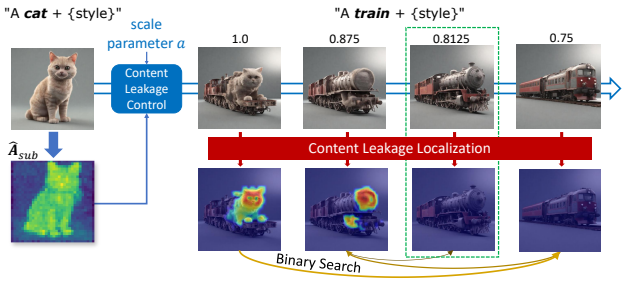


Figure 2. **Overview of *Only-Style*:** By localizing semantic content leakage, we adaptively tune a scaling parameter α that controls style sharing in image patches containing the reference subject (subject map $\hat{\mathbf{A}}_{sub}$), resulting in the optimal value that eliminates leakage while preserving stylistic consistency.

I_{tgt} with visually "aligned" style, but driven from different prompts P_{ref} and P_{tgt}. • The considered prompts have a specific structure of $\{subject\}$ + $\{style\}$. Specifically, we have the textual descriptions of the reference subject S_{ref} (e.g., "a cat") and the target subject S_{tgt} (e.g., "a train"), which are combined with the desired stylistic description P_{stl} (e.g., "in realistic 3D render").

Content leakage occurs when attributes of S_{ref} visually "leak" into patches of I_{tgt} leading to unwanted semantic content overlap. The core idea behind *Only-Style* is to detect these image patches associated with the reference subject and *adaptively* reduce their contribution to the shared style generation, as shown in Fig. 2. The algorithm consists of three main steps that are performed at inference time:

- *Content Leakage Control* (Sec. 3.3): First, we identify the patches in I_{ref} that are relevant to the reference subject. This way, their contribution to the shared style generation can be reduced by simply scaling them down.
- *Content Leakage Localization* (Sec. 3.4): Next, we detect the patches in I_{tgt} that are more relevant to S_{ref} than to S_{tgt}, denoting content leakage.
- *Adaptive Scaling* (Sec. 3.5): Finally, we combine the steps above to determine the optimal scaling, eliminating content leakage while retaining the stylistic alignment, via a binary search process.

3.2. Preliminaries

Attention in T2I Diffusion Models. State-of-the-art T2I diffusion models [3, 34, 40] typically use a U-Net [38] architecture as the backbone[2]. These image-to-image architectures are augmented with transformer blocks, each one of them consisting of a *self-attention layer* followed by a *cross-attention* layer. The latter contextualizes the deep image features with the text token embeddings. The proposed method is employed on these transformer blocks.

[2]some models, such as [7], use a Transformer [47] backbone - which is in line with the proposed framework since it relies on attention layers

Following the typical attention layer conventions [47], deep features are projected into queries $\mathbf{Q} \in \mathbb{R}^{n \times d_k}$, keys $\mathbf{K} \in \mathbb{R}^{m \times d_k}$, and values $\mathbf{V} \in \mathbb{R}^{m \times d_v}$. The output of the attention layer is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} = \mathbf{A}\mathbf{V}, \quad (1)$$

where $\mathbf{A} \in \mathbb{R}^{n \times m}$ is the output of the softmax operator, referred to as the *attention probabilities* and essentially describing the correlation between $\mathbf{Q}$ and $\mathbf{K}$. In self-attention, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are all derived from the same image features $\mathbf{F}$, while in cross-attention, $\mathbf{Q}$ comes from the image features and $\mathbf{K}$, $\mathbf{V}$ come from the text token embeddings.

3.3. Content Leakage Control

The first step in mitigating content leakage is to regulate the influence of the reference image. Most existing approaches incorporate hyperparameters that serve this purpose (see Fig.5 and the relevant discussion). Aiming for a finer control, *Only-Style* builds upon an *attention-based* method that achieves consistent style generation, StyleAligned [13]. Our objective is to leverage the attention mechanism to selectively scale down reference subject patches while maintaining style consistency, even under scaling adjustments.

Specifically, this control module consists of two steps: **1)** detection of image patches in I_{ref} that visually correspond to the subject S_{ref}, and **2)** scaling down the contribution of these subject patches only, according to a given scale parameter a. The first step requires the inference of the reference image I_{ref} according to the reference prompt $P_{ref} = S_{ref} + P_{stl}$, and is performed within the *cross-attention* layers of the transformer blocks. The second step is then performed on the *self-attention layers*, by scaling down the reference keys $\mathbf{K}_{ref}$ of the shared attention mechanism employed in [13] (see suppl. material) during the generation of I_{tgt}.

Detecting Subject Patches. As evident in Fig. 1, content leakage in I_{tgt} originates from the transfer of patches in I_{ref} that are semantically close to the reference subject S_{ref}. Our key observation is that cross-attention layers, which serve as a semantic explanation in T2I models[12], can be leveraged to annotate these patches in I_{ref}.

Thus, to access and control the "leakage" of a subject, an aggregated attention visual map $\mathbf{a}_{sub} \in \mathbb{R}^n \equiv \hat{\mathbf{A}}_{sub} \in \mathbb{R}^{H \times W}$ is required, referred to as *subject map*, where $\mathbf{a}_{sub}$ is the flattened version across all the patches, while $\hat{\mathbf{A}}_{sub}$ is reshaped to match the image's spatial structure. Specifically, we use the cross attention probabilities ($\mathbf{A}^{l,t}$, see Eq. 1) at iteration t and in the layer l. The considered layers are the bottleneck layers of the U-Net backbone, known to contain rich semantic information [23, 32].

Given a set of bottleneck attention layers B and iterations $t \in [1, T]$, we compute the averaged cross attention

probabilities with respect to the subject token S_{ref} as:

$$\mathbf{a}_{sub} = \frac{1}{T|B|}\Big(\sum_{t}\sum_{l\in B}\mathbf{A}^{l,t}\Big)\mathbf{e}_s \tag{2}$$

where $\mathbf{e}_s$ isolates the column corresponding to the subject.

The output of this step is a binary mask $\mathbf{R} \in \mathbb{R}^{H\times W}$, identifying patches relevant to the subject. Thus, given the subject map $\hat{\mathbf{A}}_{sub}$, we aim to separate the patches content-related patches (*source of leakage*) from unrelated ones. Since it is impossible to a priori specify a good thresholding value across all cases (slightly different text prompts lead to different attention values, even for the same subject token), we perform the separation via a K-means clustering method with two centroids. See Suppl. Mat. for details.

Controlling Content Leakage. Following [13], content leakage can be mitigated by reducing the contribution of reference key features $\mathbf{K}_{ref}$ in shared self-attention layers.

However, scaling all the patches is not optimal, since style contribution can be affected too (see ablation on Suppl. Mat.). Instead, we selectively scale only the key features of "content patches," as determined by the subject mask $\mathbf{R}$. Using a single scalar parameter $\alpha \in [0, 1]$, we scale the key features at each iteration t and layer $l \in B$ as:

$$\hat{\mathbf{K}}_{ref} = (1 - \mathbf{R}) \odot \mathbf{K}_{ref} + \alpha\mathbf{R} \odot \mathbf{K}_{ref} \tag{3}$$

Intuitively, by reducing α, this weighting makes the attention distribution on the reference subject patches more uniform, resulting in a global stylistic alignment rather than a polarised local "semantic" transfer. As shown in Fig. 2, decreasing α progressively reduces content leakage in a semantically explainable manner.

3.4. Content Leakage Localization

Reducing the influence of reference subject patches may lead to stylistic misalignment, as shown in Fig. 4. This necessitates finding a scaling parameter high enough for accurate style transfer yet low enough to minimize content leakage. To achieve this, we must measure content leakage in target images to establish a lower bound for scaling.

To this end, we introduce a patch-level content leakage localization method at inference, applied in two consecutive diffusion iterations. Our method relies on a simple premise: *determining whether a target image patch contains more information about the reference than the target subject.* To implement this, we need to define the following: **1)** how to extract robust and faithful representations $\mathbf{v}$ for both subjects, **2)** how to use them to detect leakages.

Extracting Subject Representations. The CLIP token embeddings that guide the generation via the cross-attention, are not expressive enough to localize subtle reference subject that overlap with the target. To improve this, we use cross-attention maps $\hat{\mathbf{A}}_{sub}$ (Sec. 3.3), averaged on a single iteration, to pool one representation per subject before each self-attention layer. Since directly using $\hat{\mathbf{A}}_{sub}$ does not reliably localize the most relevant features to the subject, we refine this via clustering and percentile thresholding, forming a binary mask $\mathbf{M}_{sub}$ of subject-relevant patches[3]. The refined subject-relevant attention map $\tilde{\mathbf{A}}_{sub}$ is then extracted as:

$$\tilde{\mathbf{A}}_{sub} = (\mathbf{M}_{sub} \odot \hat{\mathbf{A}}_{sub}) / \sum(\mathbf{M}_{sub} \odot \hat{\mathbf{A}}_{sub}) \tag{4}$$

In the iteration following the extraction of $\tilde{\mathbf{A}}_{sub}$, each layer l uses the feature map $\mathbf{F}^l$ before the self-attention layer to extract a per-layer visual representation $\mathbf{v}^l \in R^d$ that best describes the subject in layer l:

$$\mathbf{v}^l_{sub} = \sum_i \sum_j [\mathbf{F}^l \odot \tilde{\mathbf{A}}_{sub}]_{ij} \tag{5}$$

This gives us two sets of representation vectors: $\{\mathbf{v}^l_{ref}\}$ corresponding to S_{ref} in I_{ref} and $\{\mathbf{v}^l_{tgt}\}$ for S_{tgt} in I_{tgt}.

Detecting Leakages. The semantic relevance of each patch in I_{tgt} to the subject can then be computed using the cosine similarity scores between the target image features $\mathbf{F}^l_{tgt}$ and the subject representations $\mathbf{v}^l_{sub}$. For each subject, we aggregate this similarity across bottleneck layers $\{B\}$, resulting in the similarity map $\mathbf{C}_{sub} \in \mathbb{R}^{H\times W}$. Formally:

$$[\mathbf{C}_{sub}]_{ij} = \frac{1}{|B|}\sum_{l\in B} cos([\mathbf{F}^l_{tgt}]_{ij}, \mathbf{v}^l_{sub}), \tag{6}$$

where $i \in [1, H]$ and $j \in [1, W]$ denote spatial patch positions, and cos is cosine similarity. Thus, we obtain the similarity map $\mathbf{C}_{ref}$ of S_{ref} and $\mathbf{C}_{tgt}$ of S_{tgt} (see Fig. 3).

A patch p_{ij} is marked as content leakage (binary value L_{ij}) if it contains more of S_{ref} than S_{tgt}:

$$\begin{aligned}\mathbf{L}_{ij} = &\big([\mathbf{C}_{ref}]_{ij} \geq [\mathbf{C}_{tgt}]_{ij} + t_{leak}\big) \wedge \\ &\big(([\mathbf{C}_{tgt}]_{ij} \geq t_{rel}) \vee ([\mathbf{C}_{ref}]_{ij} \geq t_{rel})\big)\end{aligned} \tag{7}$$

where t_{leak} determines the minimum difference for leakage detection, and t_{rel} filters out irrelevant/background patches.

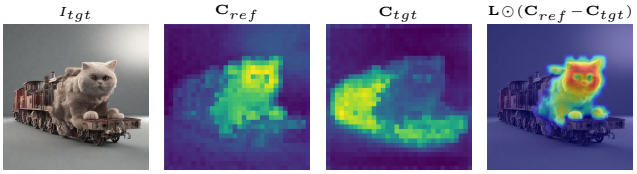


Figure 3. **Similarity maps** of the original reference subject (cat) and the target subject (train). By combining these maps we can effectively localize content leakage in the target image.

[3]Check Suppl. Mat. for more details on this step.

Figure 4. **Adaptive vs Fixed Scaling**: Fixing the scaling parameter that controls shared attention does not yield consistent results across all instances, often failing to prevent content leakage or accurately align the desired style with the target subject.

Finally, we can obtain the overall leakage value as the logical addition of $\mathbf{L}_{ij}$: $L_o = \bigvee_{ij} \mathbf{L}_{ij}$. Both thresholds remain fixed across all experiments (t_{leak}=0.1 & t_{rel}=0.4).

A key property of generative diffusion models is that structure and semantics emerge early in the process [32]. Since content leakage is inherently semantic, we can apply the proposed process early in the denoising stage, bypassing full generation. Thus, to *increase efficiency*, we perform the proposed localization approach at $t = T/2$, as our experiments indicate that leakage is observable by this stage.

Generalization. The proposed localization approach can be applied in the output of any style consistency diffusion pipeline, used as a standalone component. Specifically, given the images I_{ref} and I_{tgt} along with their subjects S_{ref} and S_{tgt} we can use DDIM inversion to obtain latents z_T, z_{T-1} for each image. Then, we simulate the final two diffusion steps to localize content leakage as described in this section. This has minimal computational overhead and can serve as a post-processing step for any style consistency method.

3.5. Adaptive Scaling

As mentioned before, we would like to choose the maximum value of scale α that resolves content leakage, as faithfully aligning the style of the reference "subject" with the target one is also necessary in many cases, and it is intuitively natural that lower values of scale, lower that alignment. To avoid a linear search, we exploit the monotonicity

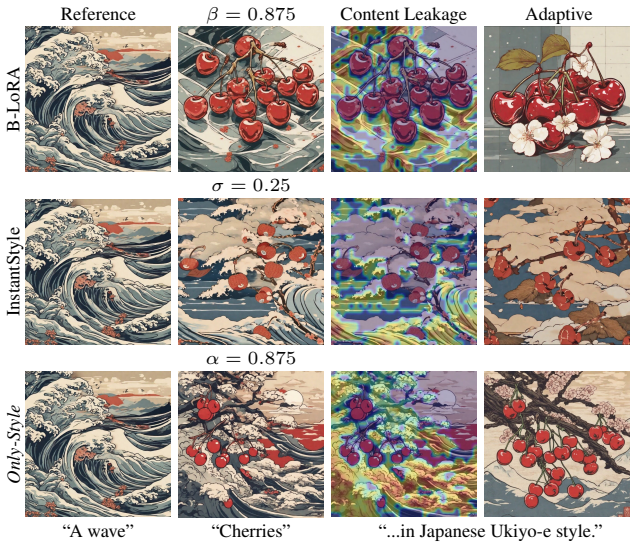


Figure 5. **Hyperparameter tuning to mitigate content leakage.** We apply the leakage detection of Sec.3.4 to adaptively tune hyperparameters presented in B-LoRA[8] and InstantStyle[48]. While these methods reduce leakage, they distort the reference style. *Only-Style* preserves both style and content integrity.

of α with respect to the binary leakage indicator L_o, which answers the decision problem: *"Is there any content leakage on I_{tgt}?"*. Given this property, we apply binary search on α to efficiently minimize leakage. This process has a multiplicative computational overhead of $\Theta(|\log(p)|)$, requiring $|\log(p)|$ style aligning generations, where p is the enforced precision of α.

4. Experiments and Evaluation

4.1. Experimental setup

Implementation. We implement our method on top of StyleAligned [13] which uses Stable Diffusion XL (SDXL) [34] at its core. The proposed leakage control process (for a single scale α) takes 31 seconds in a NVIDIA GeForce RTX 3090, only 2 seconds more than the base StyleAligned generation. We set a fixed precision value of $p = 0.03125$ for the binary search of Sec. 3.5, which means that the whole process involves 4 half generations (see Sec. 3.4), each one of them determining the binary indication of leakage, and one whole, to generate the final style-consistent pair. Thus, our method runs for approximately 1.5 minutes per style alignment instance. Details on time requirements of sota in the Suppl. Mat.

Evaluation prompt set. We create an evaluation set of 100 prompts, by extracting the 100 creative style descriptors used in the evaluation set of StyleAligned [13], and then employing ChatGPT to generate highly diverse objects (4 per style) that could appear in the specific style context. The unified style prompt looks like the following: {*'A clock', ..., 'A cupcake'*} *in abstract rainbow colored flow-*

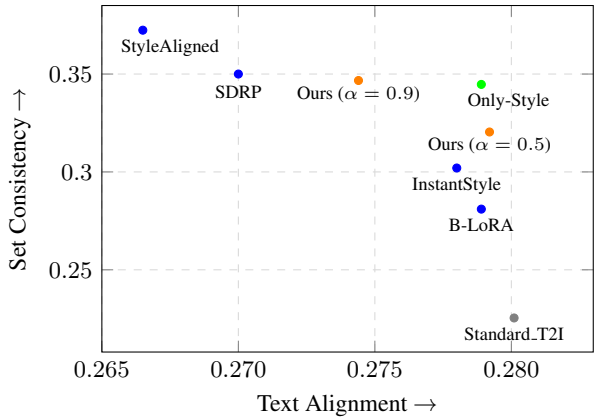


Figure 6. **Text Alignment vs Stylistic Set Consistency**: We compare three state-of-the-art methods (blue marks), a baseline without stylistic alignment (grey mark), our two ablation variants (orange marks) and *Only-Style* (green mark) in terms of text alignment (CLIP similarity) and set consistency (DINO similarity).

ing smoke wave design. The textual descriptions of the objects are kept minimal without modifiers such as adjectives and prepositional phrases, as the linguistic identification of the subject text tokens within a prompt, that is necessary for the localization of content leakage, remains out of the scope of our work. The set is presented in the Suppl. Mat.

Comparison with state-of-the-art methods. We compare our work with the following state-of-the-art style-consistent generation methods, implemented on top of generative diffusion models. Apart from StyleAligned (SA) [13], we considered InstantStyle (IS) [48] as an adapter-based method that focuses on avoiding content leakage, B-LoRA [8] and StyleDrop (SDRP), as two optimization-based baselines that yield state-of-the-art results. For B-LoRA [8] we utilize only the style adaptation and employ it for text-based style consistent generation. For the first three we utilize their official implementations, while we implement SDRP on top of SDXL. All methods use SDXL as their base model. Comparisons with additional baselines can be found in the Suppl. Mat.

4.2. Ablation study

Adaptive vs Fixed scaling. To highlight the critical role of adaptivity in style alignment, we fix the scaling parameter α into distinct values, essentially performing only the approach presented in Sec. 3.3. Specifically, we select two different fixed values ($\alpha = 0.9$ and $\alpha = 0.5$) and qualitatively compare the results with those of *Only-Style*, as shown in Fig. 4 - quantitative results with these fixed value alternatives will be presented in upcoming sections. It is evident that in both cases, the fixed scaling parameter is either unnecessarily low, ruining stylistic alignment, or not low enough to erase the effect of content leakage. On the contrary, *Only-Style* faithfully removes content leakage while maintaining the desired stylistic alignment. More ablation studies can be found in the Suppl. Mat.

4.3. Comparisons

Qualitative Comparisons. Figure 7 presents qualitative comparisons between *Only-Style* and the considered baselines. The first four rows showcase results from our evaluation prompt set, the fifth row depicts an example of an intricate multi-subject reference and target scene, while the bottom three rows include real stylistic reference images, a common application scenario in the literature. Concerning the case of real images, we employ a DDIM-based inversion technique to transfer the style of real images to the generated images of different target prompts, following the paradigm of StyleAligned [13]. Please refer to the Suppl. Mat. for details on the multi-subject extensions, as well as more qualitative results.

Tuning method-specific hyperparameters. Many methods that tackle style-consistent image generation have introduced hyperparameters to control the impact of the stylistic reference to the target. For example B-LoRA [8] introduces a scalar $\beta \in [0, 1]$ to reduce the influence of the style-LoRA adapter δW: $W = W_0 + \beta \cdot \Delta W$. On the other hand, InstantStyle [48] addresses leakage by modulating the subtraction of the subject CLIP text embedding from the reference image CLIP embedding using a scalar $\sigma \in [0, 1]$: $\text{CLIP}_{\text{img}}(I_{\text{ref}}) - \sigma \cdot \text{CLIP}_{\text{txt}}(S_{\text{ref}})$. We compare the effects of these parameters with our approach in Fig. 5. Additionally, we apply the proposed adaptive tuning algorithm—leveraging the generalized leakage localization method presented in Sec. 3.4—to determine the maximum values of these parameters that prevent content leakage. While this effectively eliminates content leakage, we observe that such parameters degrade the stylistic alignment with the reference. In contrast, *Only-Style* preserves the intended stylistic alignment while mitigating content leakage.

Text Alignment and Stylistic Set Consistency. Following [8, 13, 19], we use CLIP [35] cosine similarity to measure the *text alignment* between each target image I_{tgt} and the text description of the target subject S_{tgt}, while as *stylistic set consistency* we measure the cosine similarity between DINO [2] embeddings of the generated target images I_{tgt} with their reference I_{ref}. The results prompt set can be seen in Figure 6, where we compare our method to the four state-of-the-art baselines we mentioned, a baseline generation method without any stylistic alignment (Standard T2I [34]) and the two fixed scale variants of Sec. 3.3. *Only-Style* exhibits a notable balance between retaining the style of the reference image (set consistency) and faithfully depicting the target subject (text alignment). Despite an expected drop in set consistency compared to StyleAlign, which generates aligned images at the cost of exhibiting leakages, our method achieves almost identical text align-

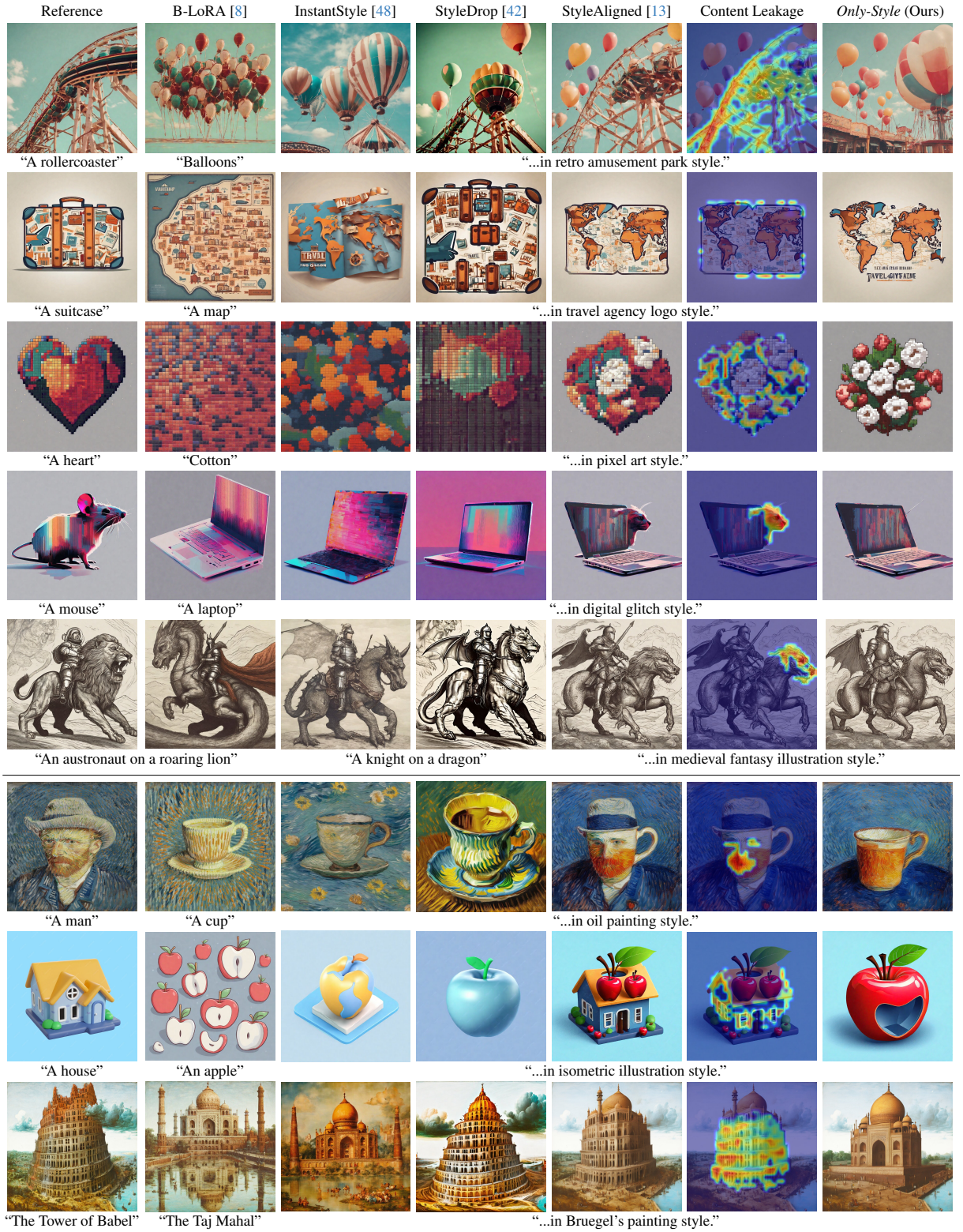


Figure 7. **Qualitative results**. We compare Only-Style against StyleAligned [13], InstantStyle [48], B-LoRA [8]. and StyleDrop [42]. In the next-to-last column we also highlight the content leakage observed in StyleAligned, which is localized and effectively mitigated by our method. First four rows are examples from our evaluation set, fifth row showcases an intricate multi-subject case, while the last three rows correspond to real reference images.

Table 1. **Content Leakage (CL) Metric Results.** We calculate CL scores across various methods, quantifying content leakage as the cosine similarity between CLIP embeddings of the target image and the reference subject's text description. Lower is better (less leakage).

full leakage	SA [13]	SDRP [42]	IS [48]	B-LoRA [8]	Ours ($\alpha = 0.5$)	Ours ($\alpha = 0.9$)	*Only-Style*	no leakage
0.28	0.231	0.227	0.220	0.223	0.214	0.219	**0.215**	0.21

Table 2. **Content Leakage Measurements using LVLM-Based Prompting.** Our method shows very low content leakage, closely matching the performance of standard T2I. For all questions, the numbers denote success rate.

Question	SA [13]	SDRP [42]	IS [48]	B-LoRA [8]	Ours ($\alpha = 0.5$)	Ours ($\alpha = 0.9$)	*Only-Style*	Standard T2I [34]
"Are there any $\{S_{ref}\}$ visual features in this $\{S_{tgt}\}$ image?" (Q1)	0.470	0.553	0.647	0.643	0.663	0.603	**0.683**	**0.703**
"Is there any $\{S_{ref}\}$ in this image?" (Q2)	0.583	0.687	0.793	0.786	0.823	0.770	**0.830**	**0.833**
"Is there any $\{S_{tgt}\}$ in this image?" (Q3)	0.850	0.903	0.94	0.947	0.943	0.930	**0.957**	**0.963**

ment with standard T2I, supporting our claim on keeping the semantics of the target subject intact. On the other hand InstantStyle [48] and B-LoRA [8] preserve alignment with the target subject but compromise stylistic consistency.

Additionally, we observe that fixed scaling results in lower stylistic consistency, as it indiscriminately scales instances that might not exhibit any leakage, while the text alignment can be negatively affected by a high scale value (0.9) that allows leakage cases. Even though this experiment offers a strong indication of the effectiveness of our method, the capability to eliminate content leakage is not captured properly by the above metrics. This is because the cosine similarity of the image DINO embeddings is favored by semantic content leakage (details on Suppl. Mat.).

Content Leakage. The aforementioned metrics fail to quantify content leakage in a straightforward way. To address this, we introduce a novel metric, called *CL*, which is defined as the cosine similarity between the CLIP embeddings of the target image I_{tgt} and the text description of the reference image subject S_{ref}, namely quantifying the correlation of the reference subject with the target image. Results of the *CL* metric can be found in Table 1. Standard T2I is used as a lower bound for the metric, which denotes the "no leakage" case, since no style sharing takes place. Likewise, we use standard T2I to generate images of S_{ref} and compute *CL* again in order to define the upper bound, denoting the "full leakage" case, when we expect to detect the subject in the image. As we can see, Only-Style obtains a score very close to the "no leakage" case, showing minor bias towards generating the reference subject. On the contrary, other methods exhibit considerably higher CL scores, indicating non-trivial leakage of the reference subject. As expected, InstantStyle (IS) [48], that tries to address the leakage issue, has the best performing score out of the compared methods, but still is out-performed by *Only-Style*.

LVLM-based Quantitative Evaluation Protocol. The aforementioned metrics are not adept in capturing fine-grained details. For example, subtle content leakage may not be penalized by the cosine similarity between the CLIP representations. For this reason, we introduce a novel evaluation protocol based on Large Vision-Language Models. Specifically, we employ LLaVA [24–27] in order to unveil content leakage by prompting the model to identify the desired target subject and possibly the undesired reference subject in the generated target image I_{tgt}. Specifically we provide the LVLM the generated images I_{tgt} along with following questions. **Q1:** *"Are there any $\{S_{ref}\}$ visual features in this $\{S_{tgt}\}$ image?"* - if the answer is positive, it suggests the existence of undesired semantic features of the reference subject. **Q2:** *"Is there any $\{S_{ref}\}$ in this image?"* - a more robust question that naturally exposes only severe content leakage cases. **Q3:** *"Is there any $\{S_{tgt}\}$ in this image?"* - check if the desired target subject is not rendered at all in the target image.

We present the results of the aforementioned evaluation in Table 2. We explicitly ask the LVLM to *"Choose one: Yes or No"* along with the question and the image subjected in stylistic alignment in order to measure the success rate of the respective method in the question we pose. We observe, again, significant content leakage across state-of-the-art methods and almost identical to standard T2I performance from our method, suggesting almost no leakage. Please refer to the Suppl. Mat. for qualitative examples indicating the performance of the proposed framework as well as benchmarking of additional methods that highlights the presence of content leakage in style-consistent generation.

User Study. We conducted a user study where participants were shown randomized triplets consisting of a reference image and two target images generated by Only-Style and one competitor method. Participants were asked to select their preferred image based on the following criteria: stylistic alignment, text alignment, and overall image quality. An option "Cannot Decide" was also provided. We collected 800 pairwise method comparisons across 100 users and show the results in Table 3. Only-Style was significantly preferred over all other baselines, indicating the effectiveness of resolving content leakage in terms of human preference. More information provided in the Suppl. Mat.

5. Conclusion

We introduced *Only-Style*, a novel approach designed to pinpoint and mitigate unwanted semantic content leakage in

Table 3. **User study:** 'a/b' indicates that Ours (left) was preferred *a* times, while the competing method was chosen *b* times. Only-Style was the preferred method by the participants.

	StyleAligned [13]	IS [48]	B-LoRA [8]	SDRP [42]
Only-Style	**357**/137	**319**/210	**419**/155	**321**/202

style-consistent generation. Extensive experiments demonstrate that *Only-Style* prevents content leakage, ensuring stylistic consistency in target images with the reference style. Finally, we proposed a framework to quantitatively assess this issue within style alignment methods, providing a structured approach to evaluate their effectiveness.

References

[1] Jingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, 2023. 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 6, 5

[3] Uiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, José Lezama, Lu Jiang, Ming Yang, Kevin P. Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. In *International Conference on Machine Learning*, 2023. 1, 3

[4] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. In *Proceedings of SIGGRAPH*, 2023. 2, 3

[5] Alexei A. Efros and William T. Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 341–346. ACM, 2001. 2

[6] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *Advances in Neural Information Processing Systems*, pages 16222–16239, 2023. 2

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, pages 12606–12633. PMLR, 2024. 1, 3

[8] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1549–1565. Springer, 2024. 2, 5, 6, 7, 8, 9

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 2

[10] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423. IEEE, 2016. 2

[11] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7289–7300, 2023. 2

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 2, 3

[13] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

[14] Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and David H. Salesin. Image analogies. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 327–340. ACM, 2001. 2

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6840–6851, 2020. 2

[16] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 5

[17] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, 2017. 2

[19] Jaeseok Jeong, Junho Kim, Yunjey Choi, Gayoung Lee, and Youngjung Uh. Visual style prompting with swapping self-attention. In *arXiv preprint*, 2024. 2, 6, 1, 5

[20] Hongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, and Mingli Song. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 26(11):3365–3385, 2019. 2

[21] Oren Katzir, Dani Lischinski, and Daniel Cohen-Or. Cross-domain cascaded deep translation. In *Proceedings of the Eu-*

ropean Conference on Computer Vision (ECCV), pages 619–634. Springer, 2020. 2

[22] Nupur Kumari, Bowen Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[23] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 3

[24] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 8

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024.

[27] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 8

[28] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 7

[29] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6038–6047, 2023. 2

[30] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, 2021. 2

[31] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 319–345. Springer, 2020. 2

[32] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23051–23061, 2023. 2, 3, 5, 1, 6

[33] Or Patashnik, Rinon Gal, Daniel Cohen-Or, Jun-Yan Zhu, and Fernando De la Torre. Consolidating attention features for multi-view image editing, 2024. 2

[34] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. 1, 3, 5, 6, 8, 7

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 2, 6

[36] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8821–8831. PMLR, 2021. 1

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. 3

[39] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023. 2, 4, 5, 6

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, pages 36479–36494, 2022. 1, 3

[41] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 2256–2265, 2015. 2

[42] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, Yuan Hao, Irfan Essa, Michael Rubinstein, and Dilip Krishnan. Styledrop: Text-to-image generation in any style. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 7, 8, 9, 6

[43] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[44] Yang Song, Jascha Sohl-Dickstein, Durk P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 2

[45] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Trans. Graph.*, 43, 2024. 2, 1, 5, 6

[46] Tarek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, 2023. 2

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3

[48] Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. InstantStyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 2, 5, 6, 7, 8, 9

[49] Zhouxia Wang, Xintao Wang, Liangbin Xie, Zhongang Qi, Ying Shan, Wenping Wang, and Ping Luo. Styleadapter: A unified stylized image generation model. *International Journal of Computer Vision*, 132:1–20, 2024.

[50] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. CSGO: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*, 2024. 4, 5, 6

[51] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 4, 5, 6

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232. IEEE, 2017. 2

Only-Style: Stylistic Consistency in Image Generation without Content Leakage

Supplementary Material

6. Methodology Details

6.1. Preliminaries: StyleAligned

As we discussed in the main manuscript, recent state-of-the-art style alignment methods in image generation [13, 19] leverage the self-attention layers of T2I models during inference to facilitate communication between images within a batch, thereby aligning their styles. We will provide further details on the operations involved in these methods and the underlying intuition, focusing on StyleAligned [13], which our method builds upon.

StyleAligned employs an attention sharing operation between a stylistic reference image (typically the first one within a batch) and the target images (other images within the same batch). This operation is only applied to the self-attention layers of the attention-augmented UNet backbone. On such an attention layer of the model's backbone, the queries $\mathbf{Q}_{tgt}$ and keys $\mathbf{K}_{tgt}$ of the target image are normalized using the queries $\mathbf{Q}_{ref}$ and keys $\mathbf{K}_{ref}$ of the reference, with the adaptive instance normalization operation (AdaIN) [17], which essentially aligns the target features with respect to the first and second moments of the reference features. Formally, we have:

$$\text{AdaIN}(\mathbf{X}, \mathbf{Y}) = \sigma(\mathbf{Y}) \left(\frac{\mathbf{X} - \mu(\mathbf{X})}{\sigma(\mathbf{X})} \right) + \mu(\mathbf{Y})$$

$$\hat{\mathbf{Q}}_{tgt} = \text{AdaIN}(\mathbf{Q}_{tgt}, \mathbf{Q}_{ref}), \;\; \hat{\mathbf{K}}_{tgt} = \text{AdaIN}(\mathbf{K}_{tgt}, \mathbf{K}_{ref})$$

Then, to further promote sharing, the attention operation is applied to concatenated versions of the keys and the values that include both reference and target features. This way, the sharing is performed in a "natural" way, where features from both reference and target images are mingled together, essentially providing style context from the reference images to the target one. More specifically, the target queries are replaced by the normalized ones $\hat{\mathbf{Q}}_{tgt}$, the target keys are replaced by the concatenation of the reference keys $\mathbf{K}_{ref}$ with the normalized target ones $\hat{\mathbf{K}}_{tgt}$ and finally the target values are replaced by the concatenation of the reference values $\mathbf{V}_{ref}$ with the target ones $\mathbf{V}_{tgt}$. The concatenation is performed at a token level, duplicating the context length in the attention layer. Following the notation of [13], the substituted shared self-attention layer is denoted as $\text{Attention}(\hat{\mathbf{Q}}_{tgt}, \mathbf{K}_{rt}, \mathbf{V}_{rt})$, where:

$$\mathbf{K}_{rt} = \begin{bmatrix} \mathbf{K}_{ref} \\ \hat{\mathbf{K}}_{tgt} \end{bmatrix}, \; \mathbf{V}_{rt} = \begin{bmatrix} \mathbf{V}_{ref} \\ \mathbf{V}_{tgt} \end{bmatrix}$$

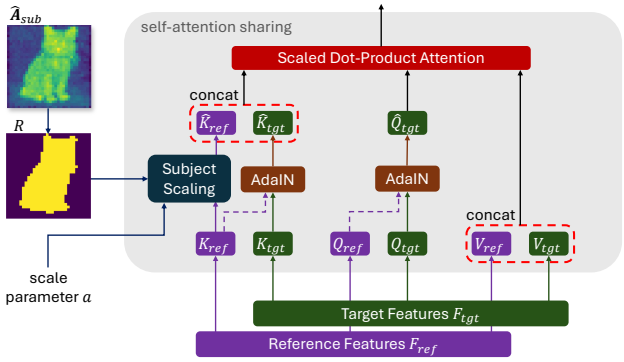


Figure 8. **Content Leakage Control**: Content leakage is mitigated by applying a weighting of the localized reference subject Key representations $\mathbf{K}_{ref}$, in every self-attention module that is used to align the style of a reference image with a target.

Note that this concatenation does not affect the size of the output, since the patch length of the queries is not affected.

The concatenation of the target features with the reference ones at a token level allows a minimal contextualization of the target image features with the reference, effectively aligning the two images. Meanwhile, applying AdaIN to the target keys using the reference boosts the attention similarity scores between the target features and the reference, facilitating a smoother attention flow from the reference to the target.

6.2. Extracting the Subject Mask R

As we discussed in Sec. 3.3 of the main manuscript, given the subject map $\hat{\mathbf{A}}_{sub} \in \mathbb{R}^{H \times W}$ (illustrated in Fig. 9), we aim to separate the patches into two distinct groups: one that is semantically related to the reference subject (and is the source of content leakage) and one unrelated.

Specifically, we consider the one-dimensional semantic representations of the image patches in $\hat{\mathbf{A}}_{sub}$ and use a K-means clustering method with two centroids to separate them, fixed across all of our experiments. Retrieving the patches grouped in the cluster with the maximum value centroid gives us the annotated subject of the image. This is equivalent to a binarization approach with a threshold depending on the image and its subject map $\hat{\mathbf{A}}_{sub}$ [45][4], as opposed to a fixed threshold approach across all images [32]

[4] A similar mask extraction was employed in [45] to extract a subject mask and then preserve the identity of this subject across multiple images, following a "dual" direction of aligning subjects and not style.

Figure 9. **Visualization of the intermediate results in the extraction of mask $\mathbf{R}$.** We cluster the aggregated cross-attention probabilities $\hat{\mathbf{A}}_{sub}$ using K-means with two centroids and then apply morphological closing to fill small gaps in the foreground.

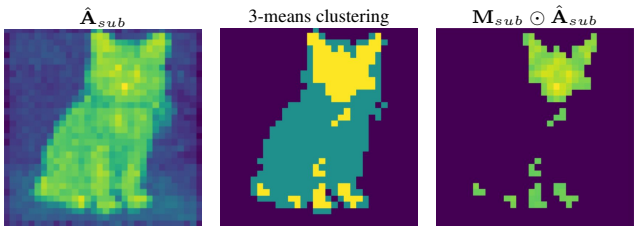


Figure 10. **Visualization of intermediate steps in extracting the mask $\mathbf{M}_{sub}$.** Using K-means clustering with 3 centroids, we segment the subject map $\hat{\mathbf{A}}sub$ to identify semantically rich subject patches (yellow-labeled $\mathbf{M}_{sub}$). Cross-attention values from these patches (third image) are then used to compute a weighted average of image representations during inference, yielding the subject representation.

which typically under performs (see Suppl. Sec. 2.1). To ensure that all the subject patches are obtained, we apply a denoising morphological closing in the binary subject mask, filling small holes and gaps in the foreground. The resulting binary mask $\mathbf{R} \in \mathbb{R}^{H \times W}$, takes true values if the corresponding patch is deemed relevant to the subject. The intermediate results of this process are illustrated in Fig. 9

Then, we use this binary subject mask to scale down the influence of the reference key features $\mathbf{K}_{ref}$ on the shared self-attention layers. As outlined in Eq. 3 of the main manuscript, we apply a uniform scalar value across all subject patches for scaling, following a "hard" decision rationale instead of using a "soft" scaling via the cross-attention probabilities $\hat{\mathbf{A}}_{sub}$ for those patches. Such "hard" choice allows the scaling parameter to be set to $\alpha = 1$ when no leakage is detected, effectively replicating the base StyleAligned [13] process in our implementation. In other words, we wanted to keep the functionality of StyleAligned as it is if no leakage is detected, rather than modifying the subject contribution every time regardless the leakage.

6.3. Leakage Control over StyleAligned

The scaling of the content patches is performed using the following equation, as it was derived in Sec. 3.3 of the main manuscript.

$$\hat{\mathbf{K}}_{ref} = (1 - \mathbf{R}) \odot \mathbf{K}_{ref} + \alpha \mathbf{R} \odot \mathbf{K}_{ref} \tag{8}$$

This way, following the notation of [13], we effectively control the self-attention distribution $\mathbf{A}$, between $\hat{\mathbf{Q}}_{tgt}$ and the updated $\hat{\mathbf{K}}_{rt} = [\hat{\mathbf{K}}_{ref}\, \hat{\mathbf{K}}_{tgt}]^{\top}$, thus controlling the transfer of the value representations $\mathbf{V}_{rt}$, and more precisely their subset that corresponds to the reference subject patches, in the target image. Note that when α=1, we have the exact same behavior with StyleAligned [13].

The proposed functionality of the scaling operation over the shared attention mechanism of [13] is depicted in Fig. 8.

6.4. Extracting the Subject Description Mask M

As outlined in Sec. 3.4 of the main manuscript, we focus on isolating a subset of the subject map $\hat{\mathbf{A}}_{sub}$ to pool the representations of image patches, thereby extracting a representation of the image's subject. This is achieved again using a binary mask $\mathbf{M}_{sub}$, which contains true values for patches whose representations should be included in the pooling operation. This subject description mask $\mathbf{M}_{sub}$, differs from the previously defined subject mask $\mathbf{R}$ in its granularity. Here, we are interested in more fine-grained localization of patches that are relevant to the subject and can help build robust pooled representations.

To extract this mask we perform again a K-means clustering of the subject map $\hat{\mathbf{A}}_{sub}$, using three clusters this time, one grouping the background patches, one grouping the poor semantic patches, and one grouping the patches with rich semantic information. We only use the latter to represent the respective subject, making sure that, the patches do not exceed 10% of the total image patches in order to retrieve a compact representation and not average-out important semantic features. This is performed via percentile thresholding if the resulting cluster with the maximum value centroid exceeds the 10^{th} percentile. Note that if the number of subject patches exceeds the 10%, the clustering operation is redundant, since one can apply percentile thresholding directly on the values of $\hat{\mathbf{A}}_{sub}$. Nonetheless, the clustering step is crucial in cases of small objects, as the percentile thresholding would also annotate background patches. Again, this is equivalent to a binarization with a threshold dependent on $\hat{\mathbf{A}}_{sub}$, but following a "stricter" criterion compared to $\mathbf{R}$. The intermediate results of this process are illustrated in Fig. 10. The proposed mask extraction was deemed helpful in practice, providing robust subject descriptions and thus helping the localization of content leakage, and no further exploration was performed on alternative ways to extract $\mathbf{M}_{sub}$.

6.5. Extensions

Multi-Image Extension. To create a set of style-consistent images using the same stylistic reference image, we follow

Figure 11. **Examples including multi-reference and multi-target subjects.** Only-Style can be directly extended to remove content leakage in multi-reference and multi-target subject scenes.

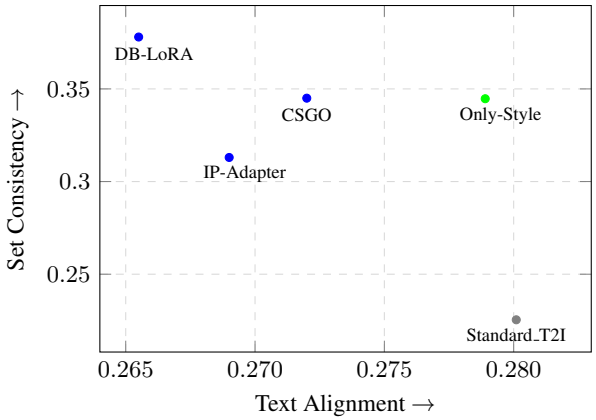


Figure 12. **Text Alignment vs Stylistic Set Consistency**: We compare three additional state-of-the-art methods (blue marks), a baseline without stylistic alignment (grey mark) and *Only-Style* (green mark) in terms of text alignment (CLIP similarity) and set consistency (DINO similarity).

the StyleAligned [13] approach by extending the batch with multiple target images. Specifically, the target images attend to the first image in the batch, which serves as the reference. Our end-to-end method can be applied independently to each target image by replicating the process described in the main manuscript. This involves defining a unique scaling parameter α for each target image and using a binary search algorithm to optimize the scaling by localizing content leakage for each such image. Importantly, this approach preserves batch parallelism, as both content leakage control and localization rely on tensor operations that can be executed in parallel. An example of a stylistically aligned image set is illustrated in Fig. 1 of the main manuscript.

Multi-subject Extension. In the main manuscript, we analyzed the single-subject scenario, where both the reference and the target prompt contained one subject. Nonetheless, our method can easily be extended in multi-subject scenarios.

For multiple reference subjects, our approach can be generalized by replicating the process outlined in the main manuscript. Here, we assume that each subject can have an optimal scaling value independent of the values selected for the other reference subjects - such an assumption stems that in theory the subject masks should be disjoint and scale a different part of the common reference image. Thus, exactly as in the multi-image scenario, we duplicate the batch and independently apply the end-to-end method to each subject, disregarding the others. Finally, we combine the optimal scaling parameters α for each reference subject to generate the resulting image, ensuring that no subject experiences leakage. It is important to note that this process requires extending the batch to include as many images as there are reference subjects, as well as performing an extra final generation of the optimal scaling set. These operations increase computational overhead, both time-wise (the extra generation step leads to a $\times 6$ overhead, including the binary search, to the standard StyleAligned for this batched multi-reference case) and memory-wise (memory requirements are multiplied by the number of reference subjects). We illustrate some indicative examples on Fig. 11. Our experimentation with multiple subjects shows that usually only the visually dominant reference subject leaks in the target image, as text-to-image models frequently focus on one subject in multi-subject scenarios [4].

For multiple target subjects, we just need to perform the content leakage localization (Sec. 3.4 of the main manuscript) of the reference subject with each target one, distinctly. Essentially we check if any patch of the generated target image contains more information about the reference subject than each of the target ones. It is worth noting that this requires extracting a subject representation for each target subject, which adds minimal computational overhead, since it is only performed at the last iteration of the generation process (see Sec. 3.4 of the main manuscript).

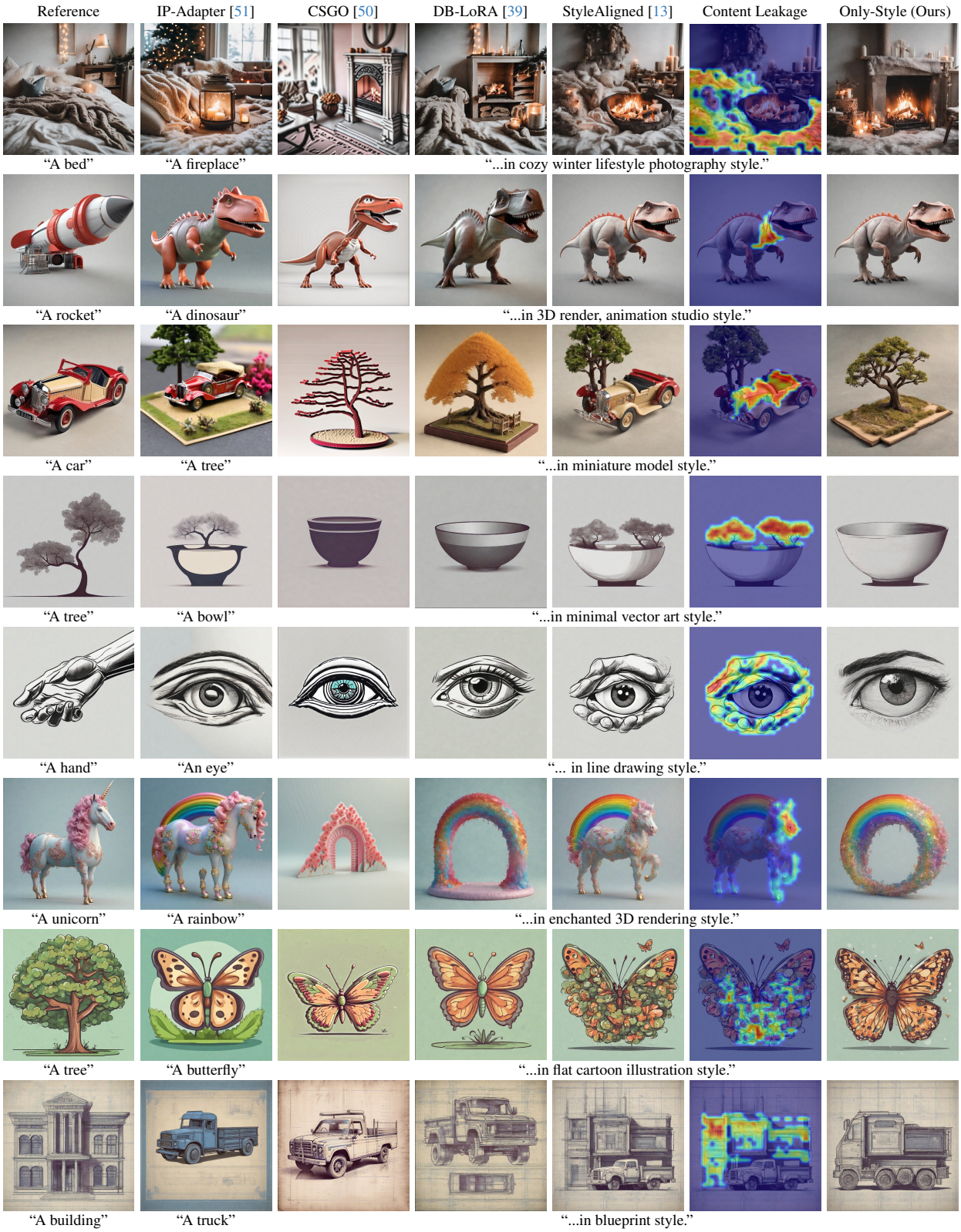


Figure 13. **Additional Qualitative results**. We compare Only-Style against StyleAligned [13], IP-Adapter [51], CSGO [50] and DB-LoRA [39]. In the next-to-last column we also highlight the content leakage observed in StyleAligned, which is localized and effectively mitigated by our method.

Metric	IP-Adapter	CSGO	DB-LoRA	*Only-Style*
CL (↓)	0.232	0.223	0.229	**0.215**
Q1 Success (↑)	0.467	0.614	0.607	**0.683**
Q2 Success (↑)	0.542	0.908	0.737	**0.830**
Q3 Success (↑)	0.886	0.732	0.857	**0.957**

Table 4. Quantitative comparison between IP-Adapter [51], CSGO [50], DB-LoRA [39], and *Only-Style*, in the metrics discussed in the main manuscript that quantify content leakage.

Method	**Set Consistency** (DINO ↑)
StyleAligned	0.372 ± 0.22
Consistory (fixed object)	0.326 ± 0.19
Standard T2I (fixed object)	0.218 ± 0.2
Standard T2I (fixed style)	0.225 ± 0.21
Only-Style	0.345 ± 0.2

Table 5. **Detailed Quantitative Results on Stylistic Set Consistency**. We evaluate the generated image sets in terms of set consistency (DINO embedding similarity). $\pm X$ denotes the standard deviation of the score across the evaluation set.

7. Additional Results

7.1. Additional Comparisons

To further highlight the effectiveness of the proposed approach, we additionally compare with the following state-of-the-art methods for style consistent image generation, namely IP-Adapter [51], CSGO [50] and Dreambooth [39], using the LoRA [16] variant (*DB-LoRA*). The first two are adapter-based methods that introduce additional layers to condition the diffusion model on the CLIP image representation of the stylistic reference, similar to InstantStyle [48]. The latter is an optimization-based method, which first finetunes the model on the reference image of a specific style by learning a compact set of adaptations (LoRA) that capture the visual characteristics of that style and then when generating new images, these learned LoRA weights are applied to transfer the original style to different subjects. All considered methods use SDXL as their base model as well. We provide both quantitative comparisons, based on the metrics outlined in the main manuscript, in Figures 4 and 12, as well as qualitative results in Figure 13, evaluated on our test prompt set. Notably, these methods also exhibit significant content leakage across all quantitative metrics assessing leakage, in contrast to *Only-Style*, emphasizing how frequently the problem occurs.

7.2. Discussion on Stylistic Set Consistency

As we discussed in the main manuscript, we follow state-of-the-art style alignment methods [13, 19] and evaluate *stylistic set consistency* within a style aligned image set, as the pairwise cosine similarity between DINO [2] embeddings of the generated target images I_{tgt} with their stylistic reference images I_{ref}. However, although the aforementioned

Figure 14. **Consistent subject in different styles.** We employ Standard T2I [34] to generate images of the same subject in different styles (first row). Since the identity of the subject is not preserved within different generations, it does not accurately simulate the effect of content leakage. To achieve this we employ a subject identity preservation method, ConsiStory [45], rendering the same object in different stylistic descriptors (second row).

metric promotes the stylistic consistency between images, it also promotes semantic and structural consistency, which is undesired in stylistic alignment.

We argue that this is because the metric is favored by semantic content leakage of the reference image subject in the target image. To quantitatively showcase this phenomenon, we employ two baselines that consist of generated sets of images in diverse styles but consistent depicted subjects. We reverse the logic of our evaluation prompt set (different objects in the same style) and generate the same object in different styles. For example: *A bear 'in Scandinavian folk art style.', 'in bohemian style.', 'in tribal tattoo style.'*

First, we employ the standard text-to-image model and generate images of an object in different styles. Note that the object generated in different styles is not the same for different generations (e.g., different kinds of bears are generated as shown in Fig. 14). This does not exactly simulate the problem of content leakage, which refers to the leak of semantic attributes of the specific visual interpretation of the reference object across the target images. To address this problem, mimicking the effect of content leakage, we employ a state-of-the-art subject identity preservation method, ConsiStory [45]. This method generates the same object (e.g., the same bear as illustrated in Fig. 14) across different styles, effectively consisting of a content leakage baseline w.r.t. the aforementioned evaluation process.

We observe that semantic consistency, expressed by the baselines we introduced, is favored as much as stylistic consistency within the *stylistic set consistency* metric. Specifically, the fixed-subject-in-different-styles variant of standard text-to-image generation achieves a set consistency score comparable to the different-subjects-in-a-fixed-style variant. Furthermore, when the identity of the generated subjects is preserved across styles using the ConsiStory approach, the pairwise set consistency achieves a level close

to state-of-the-art style alignment methods line *Only-Style* and StyleAligned [13], even though the stylistic alignment is diminished on purpose. This suggests that reducing unwanted content leakage while ensuring stylistic alignment can be penalized by this metric, which fails to fully reflect the effectiveness of our approach.

7.3. Additional Ablation Studies

Insufficiency of Fixed Thresholding.

To access and control content leakage we rely on the binary mask $\mathbf{R}$ to scale down only subject-related patches (see Sec. 3.3 of main manuscript and Sec. 1.2 of Supp. Material). As we described in Sec. 1.2 of this manuscript, the proposed extraction of $\mathbf{R}$ effectively calculates a different threshold for each $\hat{\mathbf{A}}_{sub}$ of the reference image. The same rationale was followed by [45], as opposed to the fixed threshold assumption of [32]. The fixed threshold alternative can be motivated by the fact that the $\hat{\mathbf{A}}_{sub}$ map corresponds to the aggregated cross-attention probabilities and thus a suitable probability-motivated threshold can work for all cases. Nonetheless, such an approach is inadequate in practice, as different text prompts result in varying attention probabilities. This stems from the variability of text-tokens within the prompt, which leads to distinct cross-attention distributions that cannot be modeled in advance.

We visually illustrate the effectiveness of our approach and highlight the insufficiency of fixed thresholding in Fig. 15. For the fixed thresholding case, we tune the threshold to faithfully capture the image's subject in the first column and fix it across all the other instances. As shown, the fixed threshold often fails, either being too high or too low, whereas our method consistently captures the visual elements of the object across all generated instances.

Impact of Subject Detection.

To motivate the annotation of the reference subject patches, we visually illustrate the effect of scaling all the reference image patches, essentially setting $\mathbf{R} = \mathbf{1}_{H \times W}$. We compare the results of our fixed scaling pipeline ($\alpha = .5$), presented within the ablation study in Sec. 4.2, with and without this fine-grained choice of patches, and visually display the results in Fig. 16. It is obvious that scaling agnostically the reference image patches ruins the stylistic alignment of the target image with respect to the reference. Moreover, in many cases the structure and semantics of the image are ruined as well. Note that this approach (i.e., scaling all patches) has been employed in [13] in order to mitigate the transfer of extremely popular reference image assets, which can result in disregarding the target prompt.

CLIP text embeddings vs Subject Representations. As discussed in Section 3.4 of the main manuscript, we use a patch-level localization method during inference to annotate reference subject features in the target image, which indicates content leakage. Given the semantic nature of this problem, a natural starting point is to explore the layers responsible for determining the semantics in text-to-image (T2I) generation. These semantics are primarily guided by the cross-attention layers. Thus, an intuitive initial experiment involves performing the cross-attention mechanism between the target image features and the reference subject's textual description, identifying dominant cross-attention values in the aggregated subject map. Patches in the target image that exhibit content leakage are then defined as those that "attend" significantly more to the reference subject token than to the target subject token. However, the CLIP token embeddings used in the cross-attention mechanism are not always sufficiently expressive to localize subtle visual features of the reference subject, especially when these features overlap with those of the target subject in the generated image.

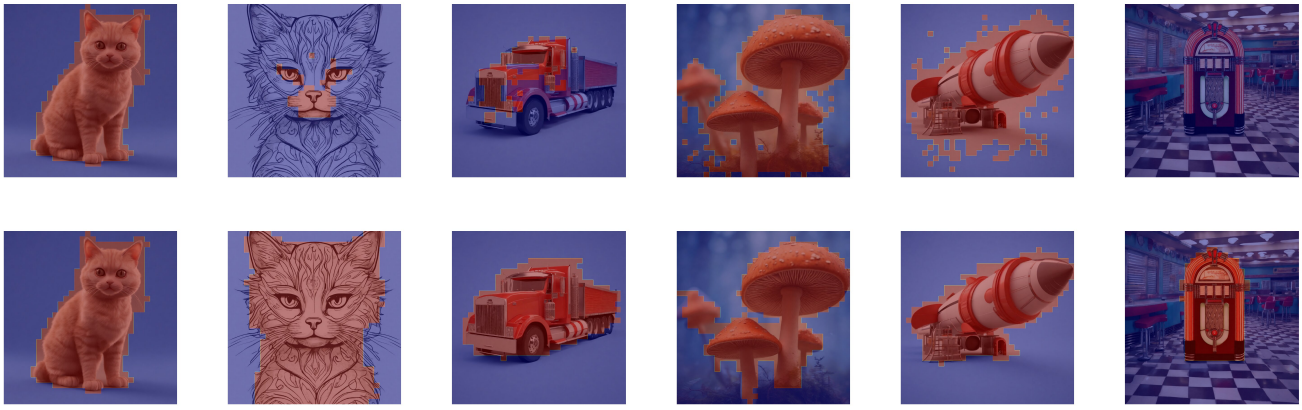
To showcase this limitation and motivate our subject representation extraction, we perform the localization using the cross-attention values, as we described above, and visually illustrate the results in Figure 17. It becomes clear that while this approach can work in cases that the leakage is semantically evident or the CLIP representations of the subjects are expressive enough to distinguish the reference from the target one (e.g., top row of Fig. 17), it fails to systematically localize the subtle content leakage features in the target image. This is because the visual representation features of our approach are by definition more descriptive of the per-case generated image and can accurately detect patches that are correlated with either the reference or the target subject. On the contrary, the textual CLIP features used in the cross-attention mechanism are limited to a more general semantic representation of the subject that can be hurtful in the context of accurate leakage detection.

7.4. Time requirements of state-of-the-art methods

In Table 6, we present the requirements in terms of time for the state-of-the-art style alignment methods evaluated. The reported time reflects the duration each method requires to generate a stylistically aligned set of two images. For optimization-based methods like B-LoRA [8], StyleDrop (SDRP) [42] DB-LoRA [39], we account for both the fine-tuning process on the reference image and the final inference to produce the stylistically aligned target. StyleAligned [13] generates a batch of two images, using the first as the reference and the second as the target.

Adapter-based methods, such as IP-Adapter [51], CSGO [50], and InstantStyle [48], operate by encoding the reference image and subsequently generating the target while integrating the reference information through cross-attention layers. However, these methods necessitate large-scale training to effectively enable this image conditioning within a diffusion model.

For our method, we first infer only the reference im-



(a) “Cat” (b) “Cat” (c) “Truck” (d) “Mushroom” (e) “Rocket” (f) “Jukebox”

Figure 15. **Insufficiency of Fixed Thresholding:** Binarization of $\hat{\mathbf{A}}_{sub}$ for different images, indicating the ability to correctly localize the subject, either via fixed thresholding (top row) or via the proposed approach (bottom row).

Method	Pretraining	Opt.	Time Requirement
IP-Adapter	✓	✗	0 min 14 sec
InstantStyle	✓	✗	0 min 16 sec
CSGO	✓	✗	0 min 20 sec
B-LoRA	✗	✓	11 min 13 sec
DreamBooth-LoRA	✗	✓	8 min 42 sec
SDRP	✗	✓	13 min 09 sec
StyleAligned	✗	✗	0 min 29 sec
Only-Style	✗	✗	1 min 46 sec

Table 6. **Time requirements of different Style Consistent Generation Methods**. We report for each method the time required to generate a stylistically aligned set of two images on an NVIDIA RTX 3090. All methods are implemented on top of SDXL. “Pretraining” denotes methods that use large scale training to incorporate image conditioning. “Opt.” denotes methods that require per instance optimization to capture a style.

age to detect the reference subject, followed by a binary search to determine the optimal scaling factor α, which results in the final stylistic alignment. As discussed in the main manuscript, we set a binary search precision of $p = 0.03125$, requiring the generation process to be repeated five times. All methods are implemented on top of the SDXL [34] framework and evaluated in a NVIDIA GeForce RTX 3090.

7.5. LVLM-based Evaluation Protocol

We also display multiple qualitative results of our evaluation framework in Figure 19, to better showcase the purpose of the questions we pose to evaluate content leakage (see Sec. 4.3 of the main manuscript). Given only the target image and the respective question, we observe that the large multimodal model can understand even subtle content leakage features. Moreover, it can unveil cases where the prompt specified target subject is not rendered at all, due to severe content leakage or dominance of background stylistic features (bottom two rows).

Notably, the LVLM systems cannot always provide correct answers for such an intricate task as the content leakage detection of fine-grained visual features. We showcase such failure cases in Fig. 18. First, the LVLM frameworks are prone to hallucinations [28], sometimes forcing the response to fit the question. For example, we come across a few object hallucinations, especially when we prompt the model to identify “visual features” which are subtle by definition (top row of Fig. 18). Moreover, content leakage refers to the presence of the reference image subject in the target image, where the generated target image is not semantically consistent to the target prompt anymore. Nonetheless, the generated target image may include visual features related to the reference subject that are in line with the requested style and do not affect the correct generation of the target subject, without displaying any content leakage. In such cases, the LVLM can correctly detect visual features of the reference subject that, however, do not correspond to a leakage case. This is particularly evident when it is semantically natural for the reference and the target subject to co-occur in a stylistic alignment scenario (bottom rows of figure 18).

However, these limitations do not consistently favor one method over another, so the mean success rates reported on our evaluation dataset serve as a reliable indicator of content leakage for each method.

7.6. Details on the User Study

In the study, users were shown a stylistic reference image alongside two target images, one generated by *Only-Style* and one by a competitor method. The images were ac-

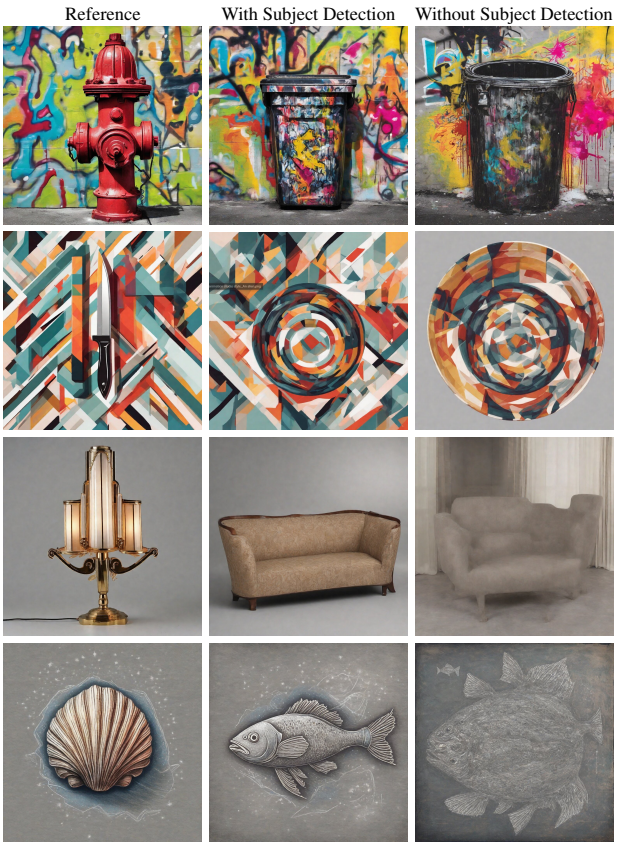


Figure 16. **Impact of Subject Detection.** We compare the results of the scaling method presented in Sec. 3.3, with and without the fine-grained choice on the reference image subject patches. Specifically, we fix the scaling parameter across our experiment ($\alpha = .5$), controlling the transfer of, on the one hand, the reference subject image patches (proposed - middle column), and on the other hand, all reference image patches (right column). We observe that scaling all patches ruins the stylistic alignment (top two rows), or exhibits destructive results (bottom two rows).

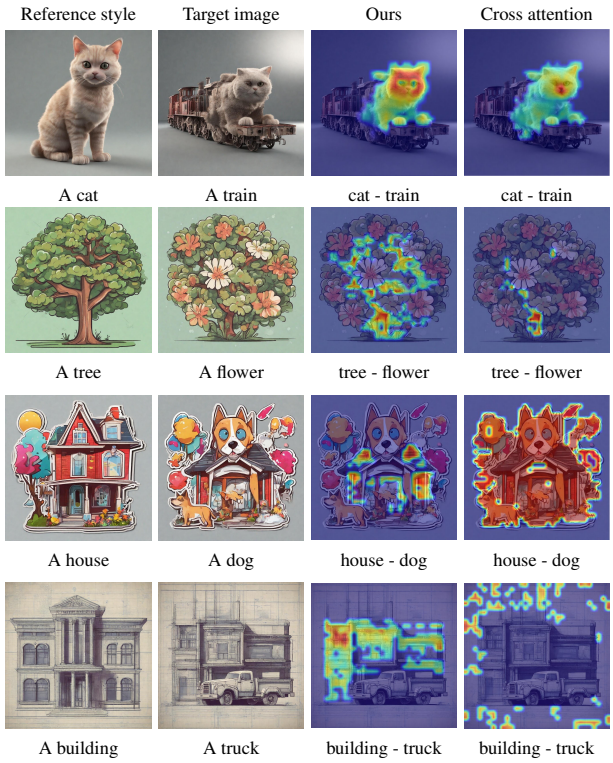


Figure 17. **CLIP text embeddings vs Subject Representations.** The first two columns are the reference and the target images, while the next two rows visualize the localization difference between the target and the reference, as defined by $\mathbf{L} \odot (\mathbf{C}_{ref} - \mathbf{C}_{tgt})$ (see Sec. 3.4 of the main manuscript). Our content leakage localization method, based on the extraction of subject representations on the image feature space, faithfully localizes the content leakage, if exists. On the contrary, the cross attention scores between image features and textual CLIP features of the subject token, even though semantically explainable, are not a trustworthy metric to perform this localization.

Competitor	Our Method	Competitor Method	Tie/Undetermined
StyleAligned [13]	357	137	306
IS [48]	319	210	271
B-LoRA [8]	419	155	226
SDRP [42]	321	202	277

Table 7. **Absolute Numbers of our Perceptual User Study**. A total of 800 pairwise comparisons were performed against each competitor method.

companied by their generating text prompts. Participants were asked to select their preferred target image based on the following criteria, stylistic alignment to the reference, alignment with the target image prompt and overall image quality, an option cannot decide was also provided, as illustrated in the example of Fig. 20. The question aimed to provide an overall evaluation of the factors contributing to successful stylistic alignment. Detailed results of our perceptual User Study with human participants are presented in table 7. As can be observed by the number of undetermined responses, participants often faced challenges in selecting a preferred method due to the conflicting evaluation criteria (style alignment versus text alignment) they were asked to consider simultaneously. Nonetheless, *Only-Style* was significantly preferred over all other baselines. It is worth noting that the significant number of undetermined responses against StyleAligned is due to instances where StyleAligned does not exhibit leakage, resulting in our method producing an identical target image.

8. Limitations

Although *Only-Style* consistently localizes the semantic content of the reference image within the target and removes it while preserving stylistic alignment, it exhibits the fol-

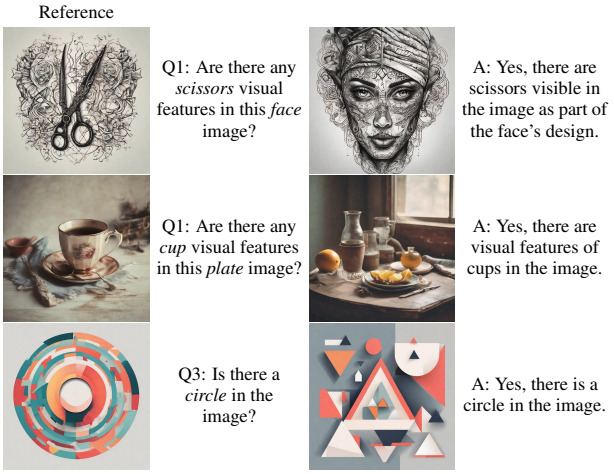


Figure 18. **Failure cases of the LVLM evaluation framework.** The target image in the first row is generated with *Only-Style* while the images in the bottom two rows are standard text-to-image generations. The subject of each image can be inferred from the respective questions.

lowing limitations.

- **Localization Accuracy:** Since our goal is to reveal the semantic visual features of the reference subject that "leaked" in the target image, we want the subject representations $\mathbf{v}_{sub}$ (see Sec. 3.4 of the main manuscript) to focus solely on the semantic features of the image subject. However, in some cases, the retrieved representations also capture stylistic features alongside the semantic ones. This results in the unintended identification of stylistic features from the reference subject within the target image as content leakage.
- **Monotonicity Assumption:** The proposed binary search for determining the optimal scaling operates under the assumption that lower values of the scaling parameter α correspond to reduced content leakage, while higher values increase it. While this monotonicity assumption relies on a straight-forward intuition ("if we reduce the contribution of the reference subject patches, we will will reduce leakage phenomena") and has been experimentally validated, it lacks a formal theoretical guarantee, especially given the complexity of the diffusion backbone. Moreover, potential non-accurate localization of the leakage (due to the way that we measure leakage - see 1st limitation) can also disrupt the monotonicity assumption, even though we have not encountered such problem in practice.
- **Computational Complexity:** Finally, one already discussed issue is the increased overhead of the proposed method compared to the vanilla StyleAligned approach. This overhead mainly stems from the iterations of the binary search. Thus, further reducing the complexity is one of the major directions for future research.

9. Future work

As a potential future enhancement we believe that it is worth exploring the ability to adaptively change the scaling parameter α during a single style alignment generation - adopting scheduling tactics or more sophisticated mechanisms.

Moreover, in a different direction, it is imperative to further establish and validate well-suited metrics, such as the proposed LVLM evaluation protocol. The main goal of such an effort is to minimize metric-induced biases (to avoid issues we met while using the set consistency metric for example). Towards this end, we can extend the concept of LVLM acting as "critics" beyond the content leakage detection.

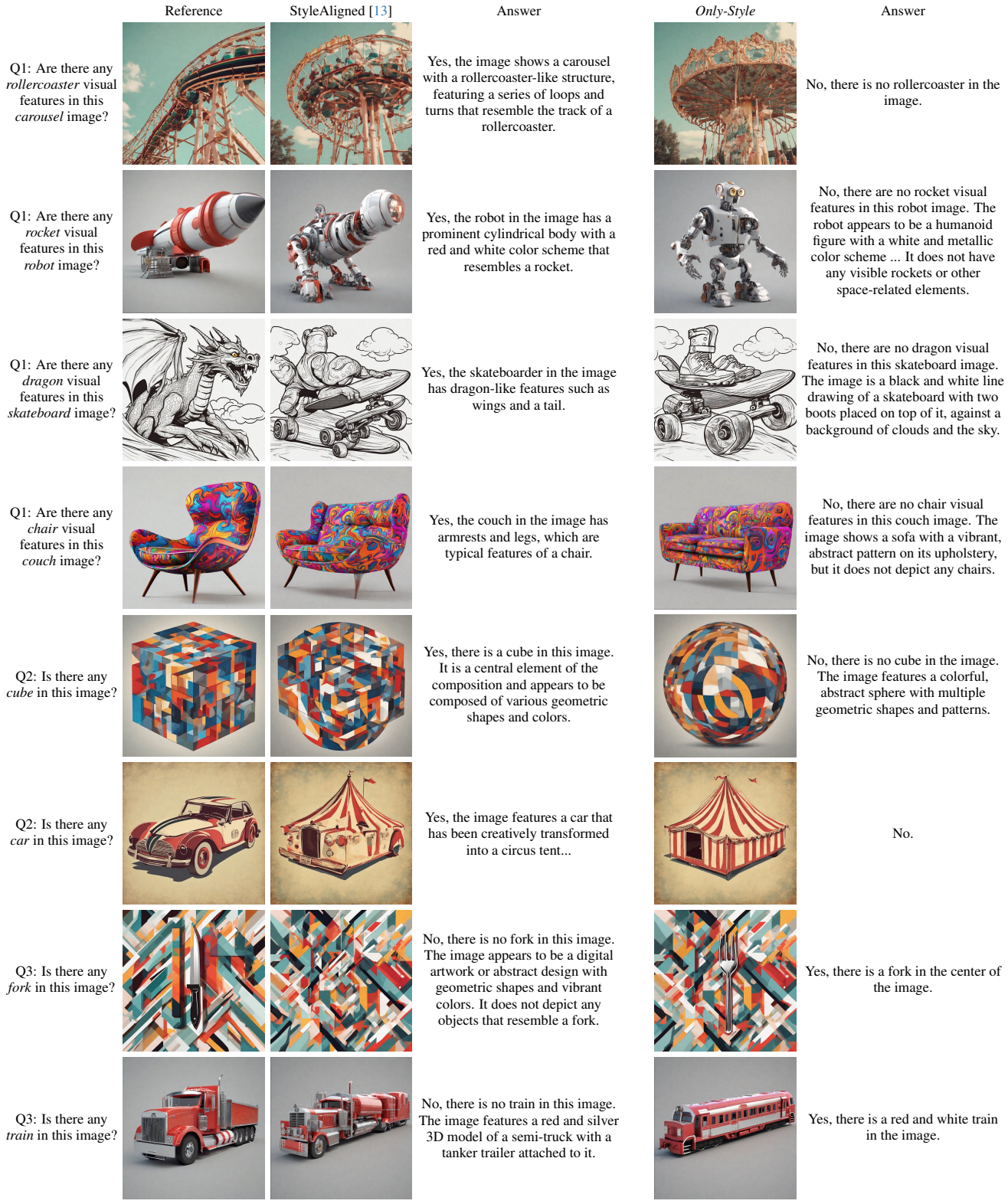


Figure 19. **Qualitative examples of our LVLM-based evaluation protocol**. We present results from StyleAligned [13], a method prone to content leakage, and *Only-Style* that mitigates this undesired effect.

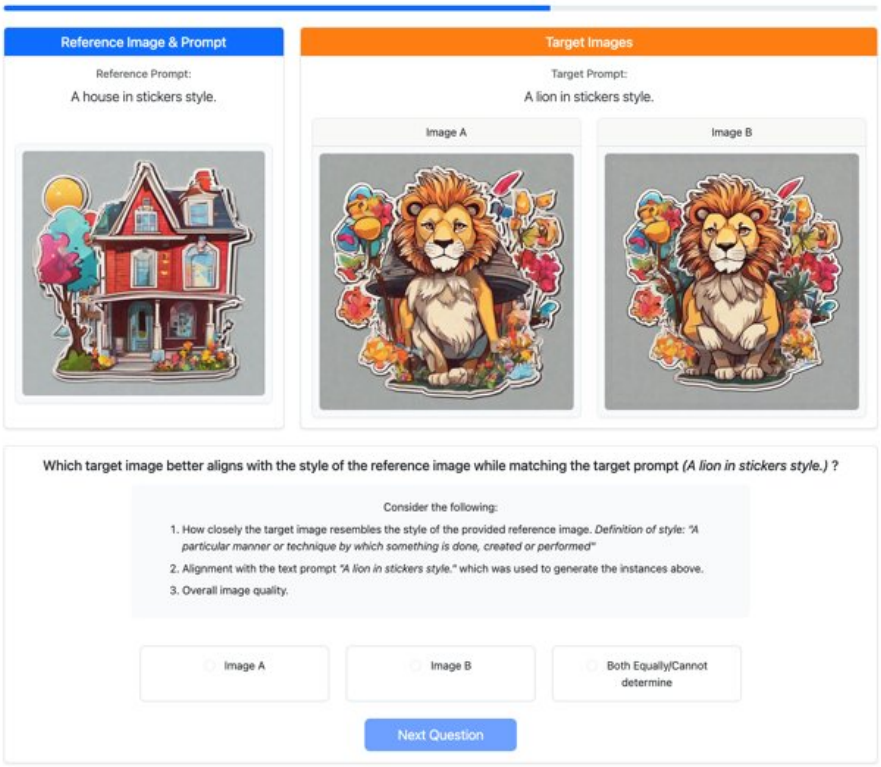


Figure 20. **An example screenshot of a question from the conducted perceptual User Study.**

Evaluation prompt set:

A house, A dog, A lion, A hippo in stickers style.
A kite, A skateboard, A canoe, A hammock in watercolor painting style.
A hand, A leaf, An eye, A feather in line drawing style.
A dragon, A teapot, A skateboard, A cactus in cartoon line drawing style.
A truck, A boat, A train, A car in 3D rendering style.
A mushroom, A dragon, A dwarf, A fairy in glowing style.
A bottle, A wine glass, A teapot, A cup in glowing 3D rendering style.
A bear, A frisbee, A ball, A torch in kid crayon drawing style.
A couch, A table, A bird, A fish in wooden sculpture style.
An elephant, A zebra, A rhino, A giraffe in oil painting style.
A tree, A flower, A mushroom, A butterfly in flat cartoon illustration style.
A clock, A chameleon, A candle, A cupcake in abstract rainbow colored flowing smoke wave design.
A fork, A spoon, A knife, A glass in melting golden 3D rendering style.
A train, A van, An airplane, A bicycle in minimalist round BW logo style.
A stop sign, A traffic light, A cone, A lighthouse in neon graffiti style.
A car, A bear, A circus tent, A clown in vintage poster style.
A wine glass, A cup, A bowl, A pitcher in woodblock print style.
A surfboard, A wave, A dolphin, A palm in retro surf art style.
A swan, An umbrella, A boat, An airplane in minimal origami style.
A robot, A spaceship, A drone, Godzilla in cyberpunk art style.
A scissors, A bug, A face, A rose in tattoo art style.
A lamp, A chair, A sofa, A mirror in art deco style.
A plant, A bed, A wave, A sunbed in vintage travel poster style.
A rollercoaster, A wheel, A carousel, Balloons in retro amusement park style.
A rocket, A dinosaur, A robot, An alien in 3D render, animation studio style.
A jukebox, A milkshake, A bench, A record player in 1950s diner art style.
A bird, A fox, A cactus, A deer in Scandinavian folk art style.
A dragon, A potion, A sword, A shield in fantasy poison book style.
A giraffe, An elephant, A flamingo, A parrot in Hawaiian sunset paintings style.
A guitar, A balloon, A drum, A microphone in paper cut art style.
A car, A vase, A camera, A watch in retro hipster style.
A suitcase, A ship, A train, A map in vintage postcard style.
A mask, A feather, A tent, A sword in tribal tattoo style.
A wave, A mountain, A cherry, A crane in Japanese ukiyo-e style.
A castle, A knight, A dragon, A wizard in fantasy book cover style.
A fireplace, A blanket, A cup, A book in hygge style.
A stone, A rake, A leaf, A lantern in Zen garden style.
A star, A planet, A comet, The moon in celestial artwork style.
A zebra, A giraffe, A horse, A lion in medieval fantasy illustration style.
A unicorn, A fairy, A castle, A rainbow in enchanted 3D rendering style.
A suitcase, A globe, A plane, A map in travel agency logo style.
A cup, Beans, A croissant, A teapot in cafe logo style.
A book, An owl, A globe, A lantern in educational institution logo style.
A screwdriver, A wrench, A hammer, A toolbox in mechanical repair shop logo style.
A stethoscope, A pill, A syringe, A thermometer in healthcare and medical clinic logo style.
A cloud, A heart, A balloon, A blossom in doodle art style.
A knife, A spoon, A fork, A bowl in abstract geometric style.
A kangaroo, A skyscraper, A lighthouse, A bridge in mosaic art style.
A butterfly, A flamingo, A flower, The sun in paper collage style.
A sunflower, A saxophone, A compass, A guitar in origami style.
A fire hydrant, A trash can, A mailbox, A streetlamp in abstract graffiti style.
A bench, A wolf, A can, A dragon in street art style.
A leaf, A clock, A cloud, A star in mixed media art style.
A snowboard, Skis, A helmet, A ski pole in abstract expressionism style.
A mouse, A keyboard, A laptop, A monitor in digital glitch art style.
A chair, A couch, A mirror, A lamp in psychedelic art style.
A clock, A vase, A painting, A torch in street art graffiti style.

A shoe, A phone, A bottle, A rose in pop art style.
A key, A bird, A door, A lock in minimalist surrealism style.
A cube, A sphere, A pyramid, A circle in abstract cubism style.
A woman, A bicycle, A camera, A bat in abstract impressionism style.
A chair, A table, A lamp, A bookshelf in post-modern art style.
A cat, A car, An android, A drone in neo-futurism style.
A lollipop, A ladder, A star, A rocket in abstract constructivism style.
Lava, Smoke, Water, Fire in fluid art style.
A butterfly, A bug, A blade, A moth in macro photography style.
A burger, A pizza, A salad, A soda in professional food photography style for a menu.
A cup, A wine glass, A plate, A bottle in vintage still life photography style.
A car, A cat, A tree, A bus in miniature model style.
A tent, A campfire, A backpack, A sleeping bag in outdoor lifestyle photography style.
A cat, A train, A serpent, A fish in realistic 3D render.
A record, A cassette, A microphone, A guitar in retro music and vinyl photography style.
A bed, A chair, A fireplace, A table in cozy winter lifestyle photography style.
A candle, A blossom, A light, A vase in bokeh photography style.
A circle, A triangle, A square, A hexagon in minimal flat design style.
A tree, A bird, A bowl, A corn in minimal vector art style.
A cloud, Waves, A blade, A sun in minimal pastel colors style.
A kitten, A tree, A house, A fence in minimal digital art style.
A fish, A bat, A star, A seashell in minimal abstract illustration style.
A mountain, A river, A cloud, A bush in minimal monochromatic style.
A wolf, A skull, A horse, A raven in woodcut print style.
A seashell, A fish, A hand, A starfish in chalk art style.
A heart, A moon, A satellite, Cotton in pixel art style.
A superhero, A villain, A city, A spaceship in comic book style.
A rocket, A planet, A spaceship, A dragon in vector illustration style.
A house, A car, A tree, A cat in isometric illustration style.
A computer, A phone, A camera, A tablet in wireframe 3D style.
A leaf, A cloud, A fish, A wave in paper cutout style.
A building, A bridge, A truck, A leopard in blueprint style.
A hero, A monster, A spaceship, A robot in retro comic book style.
A flowchart, An advertisement, A map, A graph in infographic style.
A microscope, A crystal, A flag, A telescope in geometric shapes style.
A cat, A dog, A bird, A rabbit in cartoon line drawing style.
A flower, A tree, A river, A mountain in watercolor and ink wash style.
A mushroom, A clock, A fish, A key in dreamy surreal style.
A car, A clock, A pipe, A gear in steampunk mechanical style.
Clock, Globe, Map, A compass in 3D realism style.
A bus, A scooter, A car, A bicycle in retro poster style.
A flower, A feather, A bat, A cactus in bohemian hand-drawn style.
Panda, Rhino, Telescope, Hippo in vintage stamp style.