Supplementary Material

1. Methodology Details

1.1. Preliminaries: StyleAligned

As we discussed in the main manuscript, recent state-of-the-art style alignment methods in image generation [4, 7] 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 [4], 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) [6], 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 [4], the substituted shared self-attention layer is denoted as $\text{Attention}(\hat{\mathbf{Q}}_{tgt}, \mathbf{K}_{rt}, \mathbf{V}_{rt})$, where:

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

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

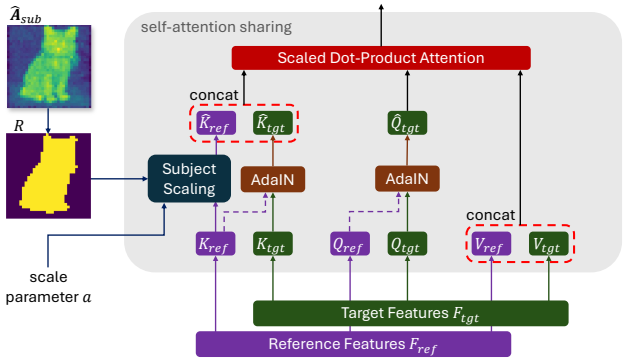

Figure 1. **Content Leakage Control**: Content leakage is mitigated by applying a weighting of the localized reference subject Key representations $\mathbf{K}_{ref}$, in every self-attention module that is used to align the style of a reference image with a target.

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.

1.2. 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. 2), 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 $\hat{\mathbf{A}}_{sub}$ [13][1], as opposed to a fixed threshold approach across all images [9] 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

[1] A similar mask extraction was employed in [13] 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 2. **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.

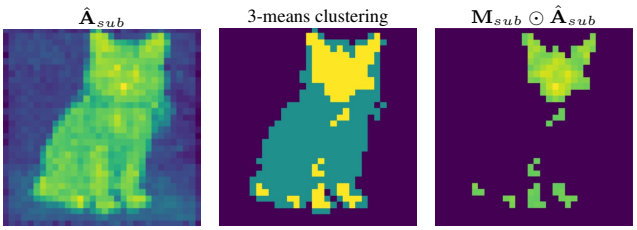

Figure 3. **Visualization of intermediate steps in extracting the mask $\mathbf{M}_{sub}$.** Using K-means clustering with 3 centroids, we segment the subject map $\hat{\mathbf{A}}sub$ to identify semantically rich subject patches (yellow-labeled $\mathbf{M}_{sub}$). Cross-attention values from these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

intermediate results of this process are illustrated in Fig. 2

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 [4] 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.

1.3. 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.

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

This way, following the notation of [4], 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 [4].

The proposed functionality of the scaling operation over the shared attention mechanism of [4] is depicted in Fig. 1.

1.4. 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. 3. 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 $\mathbf{M}_{sub}$.

1.5. Extensions

Multi-Image Extension. To create a set of style-consistent images using the same stylistic reference image, we follow the StyleAligned [4] 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

Figure 4. **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.

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

Metric	SA [4]	SDRP [12]	B-LoRA [3]	*Only-Style*
Text Align. ↑	0.276	0.2785	0.279	**0.284**
Set Cons. ↑	**0.298**	0.245	0.228	<u>0.282</u>
CL ↓	0.242	0.228	0.223	**0.209**
Q1 Success ↑	0.490	0.683	0.723	**0.767**
Q2 Success ↑	0.607	0.733	0.777	**0.843**
Q3 Success ↑	0.860	0.907	0.927	**0.947**

Table 1. Quantitative comparisons on complex prompts. *Only-Style* achieves state-of-the-art performance even under complex reference/target prompts.

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. 4. 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 [2].

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).

We also create a prompt set of 50 complex reference/target subject prompts accompanied by a style descriptor from our initial evaluation set and presenting quantitative results in Tab. 1. The trends of our initial prompts also hold here, confirming that complex prompts *do not limit our method*. The subject tokens are manually annotated within the complex prompt, but we note that some natural language processing models can automate this process.

2. Additional Results

2.1. 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 (IP) [16], CSGO [15], Visual Prompting (VP) [7] and Dreambooth [11], using the LoRA [5] 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 [14]. The third one utilizes a self-

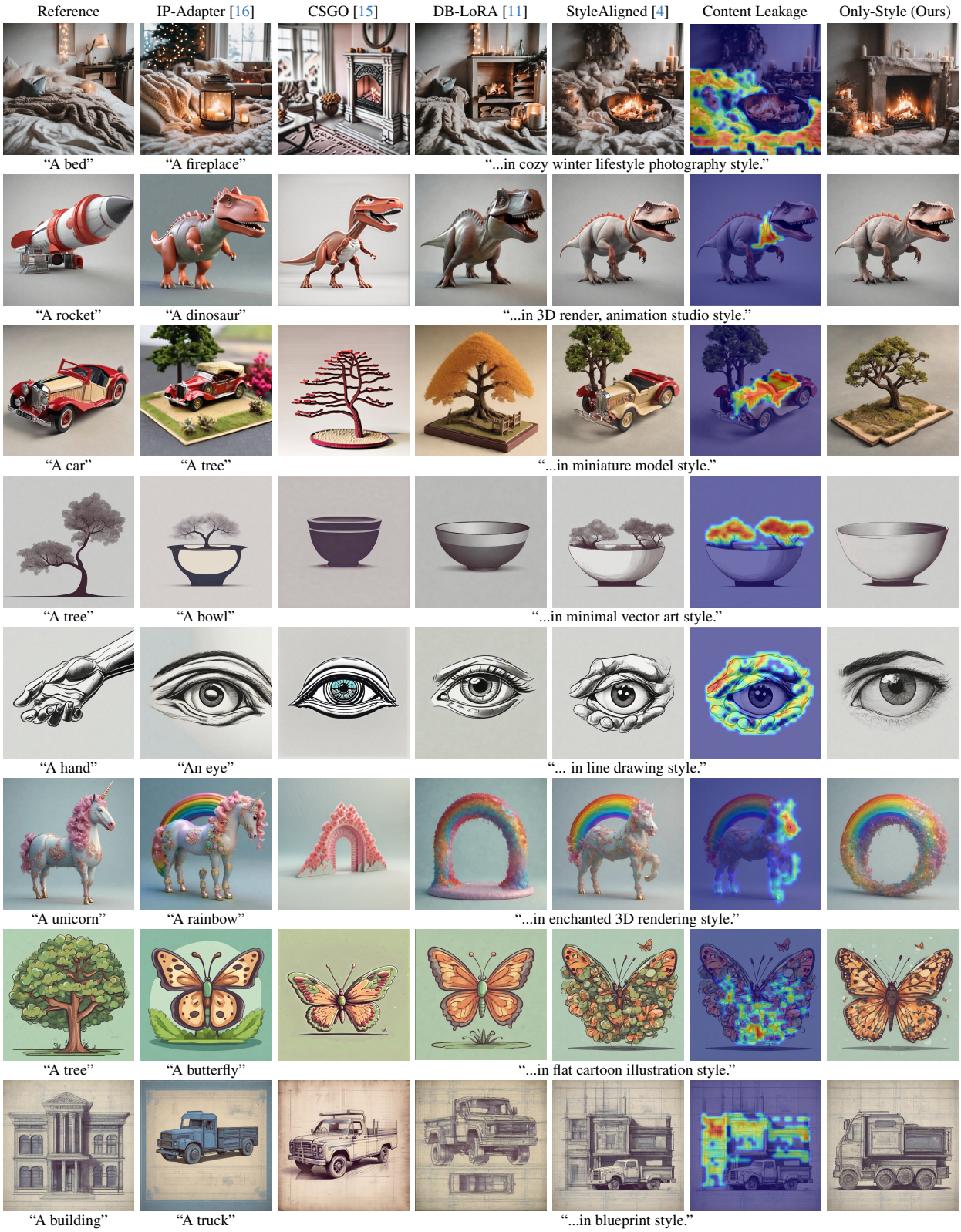

Figure 5. **Additional Qualitative results**. We compare Only-Style against StyleAligned [4], IP-Adapter [16], CSGO [15] and DB-LoRA [11]. In the next-to-last column we also highlight the content leakage observed in StyleAligned, which is localized and effectively mitigated by our method.

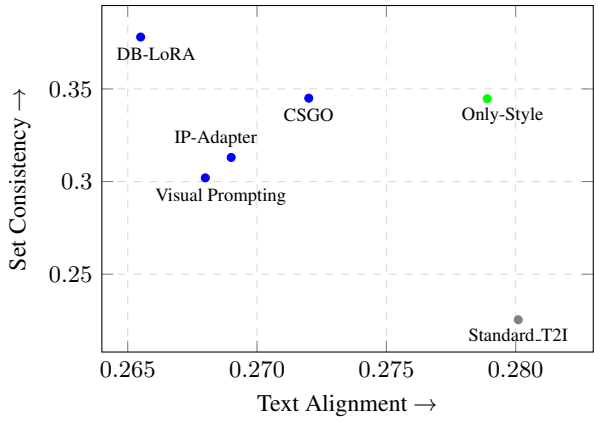

Figure 6. **Text Alignment vs Stylistic Set Consistency**: We compare four additional state-of-the-art methods (blue marks), a baseline without stylistic alignment (grey mark) and *Only-Style* (green mark) in terms of text alignment (CLIP similarity) and set consistency (DINO similarity).

attention based conditioning similar to StyleAligned [4]. 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 Fig. 6 and Tab. 2, as well as qualitative results in Fig. 5, evaluated on our test prompt set. Due to space limitations in Fig. 5, we skip Visual Prompting [7] as the weakest method quantitatively. 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.

2.2. Discussion on Stylistic Set Consistency

As we discussed in the main manuscript, we follow state-of-the-art style alignment methods [4, 7] and evaluate *stylistic set consistency* within a style aligned image set, as the pairwise cosine similarity between DINO [1] embeddings of the generated target images I_{tgt} with their stylistic reference images I_{ref}. However, although the aforementioned 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.

Metric	IP	VP	CSGO	DB-LoRA	*Only-Style*
CL (↓)	0.232	0.228	0.229	0.215	**0.215**
Q1 Success (↑)	0.467	0.540	0.607	0.683	**0.683**
Q2 Success (↑)	0.542	0.665	0.737	0.830	**0.830**
Q3 Success (↑)	0.886	0.903	0.857	0.957	**0.957**

Table 2. Quantitative comparison between IP-Adapter (IP) [16], Visual Prompting (VP) [7], CSGO [15], DB-LoRA [11], and *Only-Style*, across metrics quantifying content leakage.

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. 7). 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 [13]. This method generates the same object (e.g., the same bear as illustrated in Fig. 7) 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 [4], 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.

2.3. 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 [13], as opposed to the fixed threshold assumption of [9]. The fixed threshold alterna-

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 3. **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.

tive 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. 8. 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. 9. 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 [4] 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

Figure 7. **Consistent subject in different styles.** We employ Standard T2I [10] 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 [13], rendering the same object in different stylistic descriptors (second row).

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 10. 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. 10), 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.

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

In Table 4, 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 [3], StyleDrop (SDRP) [12] DB-LoRA [11], we account for

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

Figure 8. **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).

both the fine-tuning process on the reference image and the final inference to produce the stylistically aligned target. StyleAligned [4] 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 [16], CSGO [15], and InstantStyle [14], 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 image 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 [10] framework and evaluated in a NVIDIA GeForce RTX 3090.

2.5. LVLM-based Evaluation Protocol

We also display multiple qualitative results of our evaluation framework in Figure 12, 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. 11. First, the LVLM frameworks are prone to hallucinations [8], 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. 11). 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

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 4. **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.

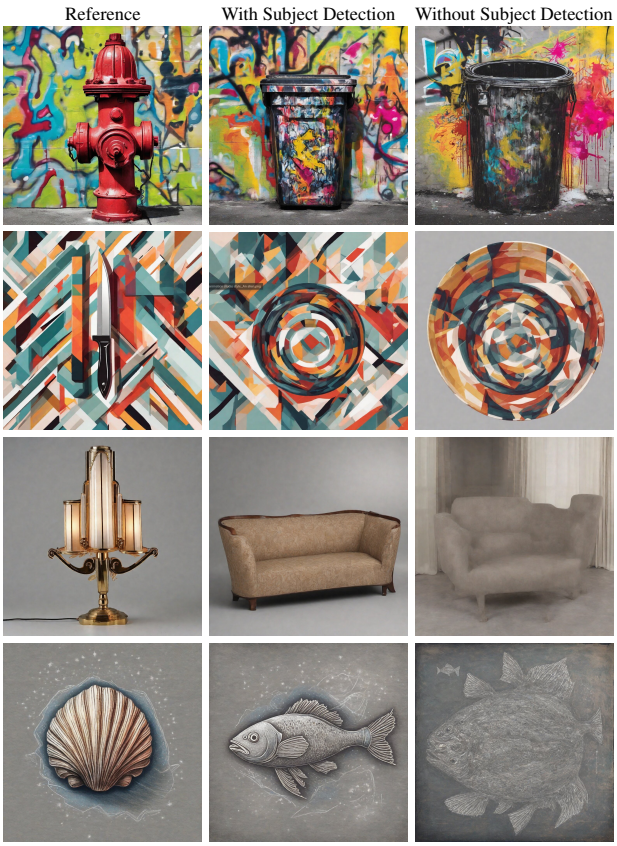

Figure 9. **Impact of Subject Detection.** We compare the results of the scaling method presented in Sec. 3.3, with and without the fine-grained choice on the reference image subject patches. Specifically, we fix the scaling parameter across our experiment ($\alpha = .5$), controlling the transfer of, 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).

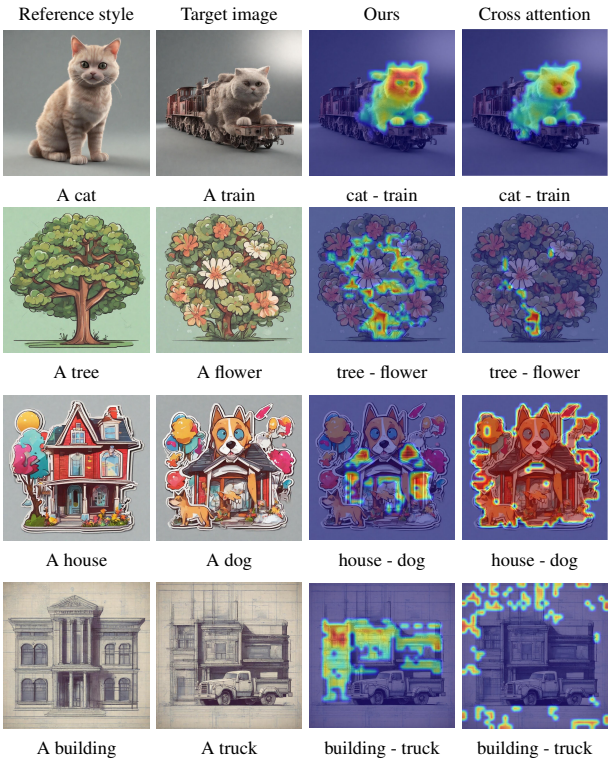

Figure 10. **CLIP text embeddings vs Subject Representations.** The first two columns are the reference and the target images, while the next two rows visualize the localization difference between the target and the reference, as defined by $\mathbf{L} \odot (\mathbf{C}_{ref} - \mathbf{C}_{tgt})$ (see 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.

natural for the reference and the target subject to co-occur in a stylistic alignment scenario (bottom rows of figure 11).

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.

2.6. 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 accompanied 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. 13. 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 5. 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.

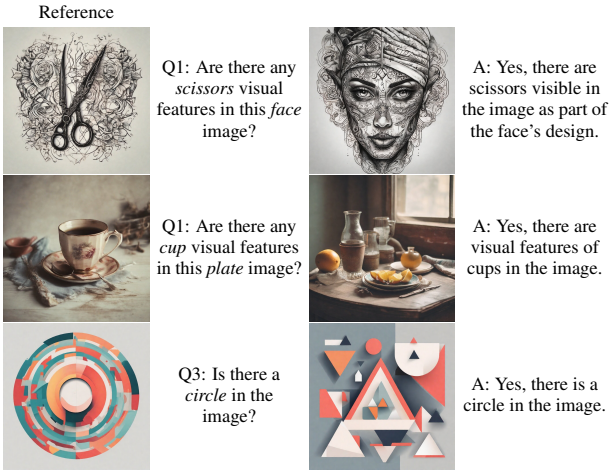

Figure 11. **Failure cases of the LVLM evaluation framework.** The target image in the first row is generated with *Only-Style* while the images in the bottom two rows are standard text-to-image generations. The subject of each image can be inferred from the respective questions.

Competitor	Our Method	Competitor Method	Tie/Undetermined
StyleAligned [4]	357	137	306
IS [14]	319	210	271
B-LoRA [3]	419	155	226
SDRP [12]	321	202	277

Table 5. **Absolute Numbers of our Perceptual User Study**. A total of 800 pairwise comparisons were performed against each competitor method.

3. 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 following 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.

4. 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.

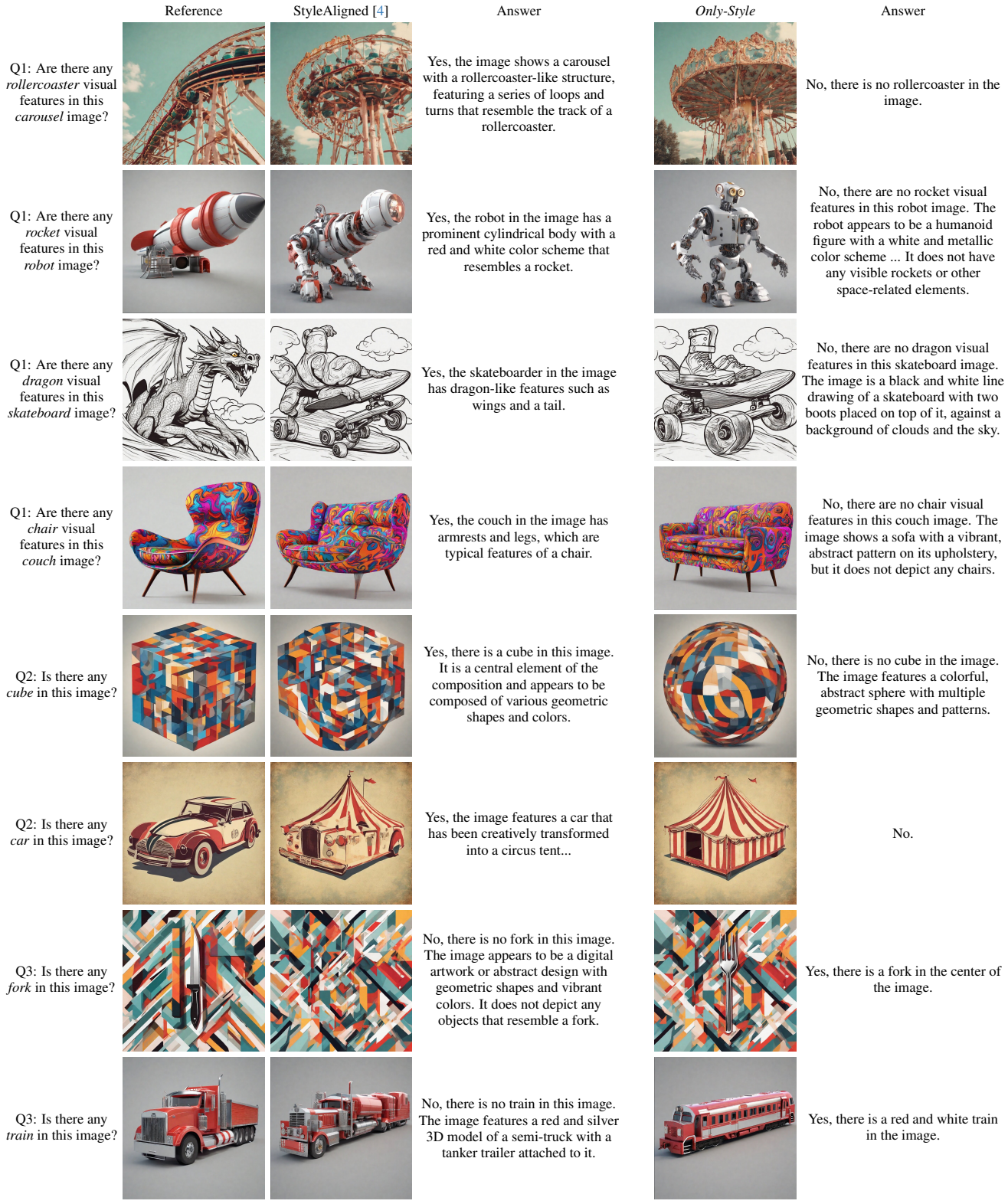

Figure 12. **Qualitative examples of our LVLM-based evaluation protocol**. We present results from StyleAligned [4], a method prone to content leakage, and *Only-Style* that mitigates this undesired effect.

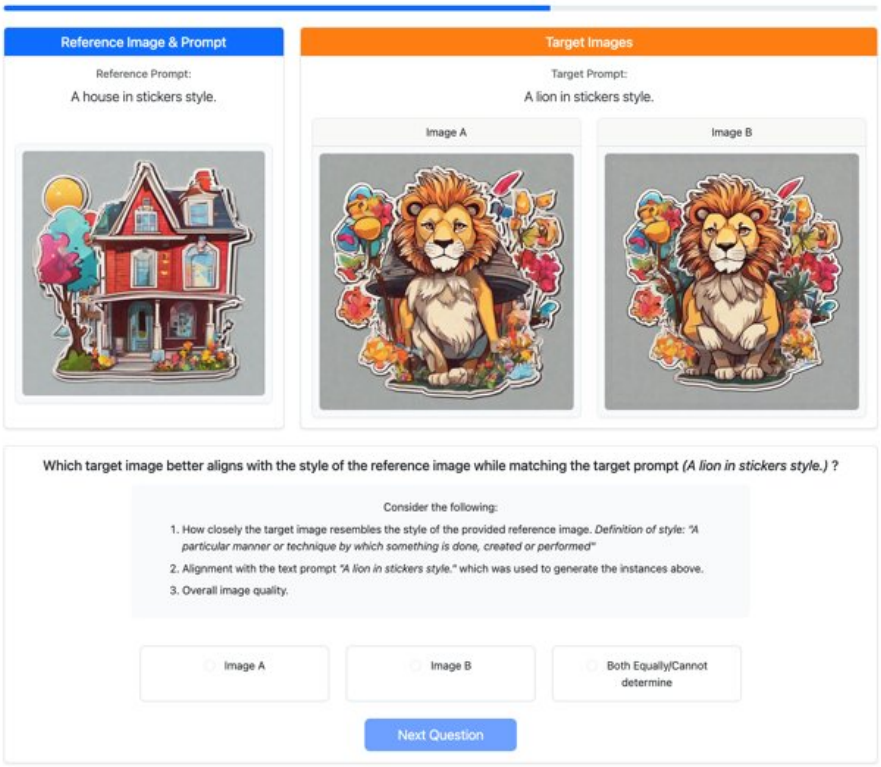

Figure 13. **An example screenshot of a question from the conducted perceptual User Study.**

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.

References

[1] 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. 5

[2] 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. 3

[3] 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. 3, 6, 9

[4] 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, 9, 10

[5] 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. 3

[6] 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

[7] Jaeseok Jeong, Junho Kim, Yunjey Choi, Gayoung Lee, and Youngjung Uh. Visual style prompting with swapping self-attention. In *arXiv preprint*, 2024. 1, 3, 5

[8] 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

[9] 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. 1, 5

[10] 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. 6, 7

[11] 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. 3, 4, 5, 6

[12] 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. 3, 6, 9

[13] 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. 1, 5, 6

[14] 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. 3, 7, 9

[15] 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. 3, 4, 5, 7

[16] 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. 3, 4, 5, 7