ConsiStyle: Style Diversity in Training-Free Consistent T2I Generation

YOHAI MAZUZ*, Tel Aviv University, Israel
JANNA BRUNER*, Tel Aviv University, Israel
LIOR WOLF, Tel Aviv University, Israel

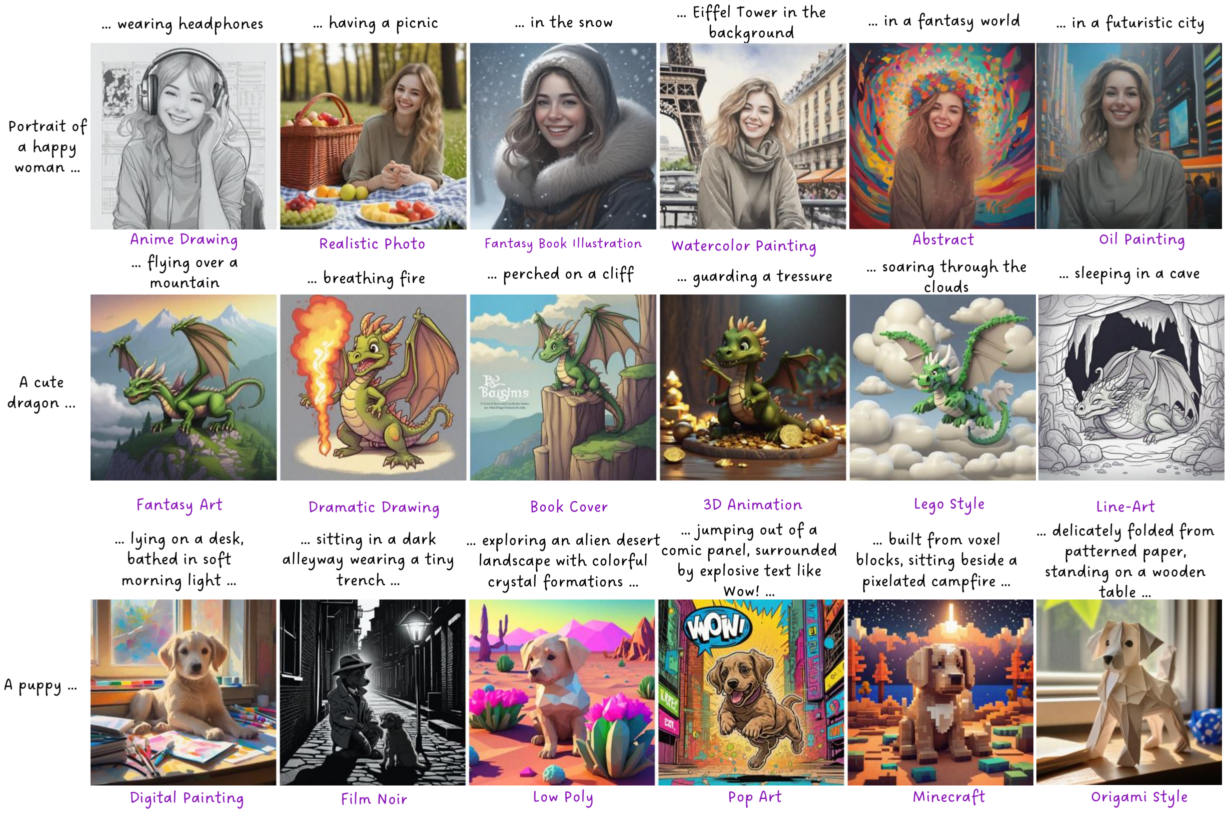

... wearing headphones
... having a picnic
... in the snow
... Eiffel Tower in the background
... in a fantasy world
... in a futuristic city
Portrait of a happy woman ...
Anime Drawing
Realistic Photo
Fantasy Book Illustration
Watercolor Painting
Abstract
Oil Painting
... flying over a mountain
... breathing fire
... perched on a cliff
... guarding a tressure
... soaring through the clouds
... sleeping in a cave
A cute dragon ...
Fantasy Art
Dramatic Drawing
Book Cover
3D Animation
Lego Style
Line-Art
... lying on a desk, bathed in soft morning light ...
... sitting in a dark alleyway wearing a tiny trench ...
... exploring an alien desert landscape with colorful crystal formations ...
... jumping out of a comic panel, surrounded by explosive text like Wow! ...
... built from voxel blocks, sitting beside a pixelated campfire ...
... delicately folded from patterned paper, standing on a wooden table ...
A puppy ...
Digital Painting
Film Noir
Low Poly
Pop Art
Minecraft
Origami Style

Fig. 1. We present **ConsiStyle** - a training-free optimization method that decouples style from subject-specific characteristics such as color, structure, patterns and unique markings. Our approach preserves subject consistency across various prompts while maintaining alignment with diverse style descriptions.

In text-to-image models, consistent character generation is the task of achieving text alignment while maintaining the subject's appearance across different prompts. However, since style and appearance are often entangled, the existing methods struggle to preserve consistent subject characteristics while adhering to varying style prompts. Current approaches for consistent text-to-image generation typically rely on large-scale fine-tuning on curated image sets or per-subject optimization, which either fail to generalize across prompts or do not align well with textual descriptions. Meanwhile, training-free methods often fail to maintain subject consistency across different styles. In this work, we introduce a training-free method that, for the first time, jointly achieves style preservation and subject consistency across varied styles. The attention matrices are manipulated such that Queries and Keys are obtained from the anchor image(s) that are used to define the subject, while the Values are imported from a parallel copy that is not subject-anchored. Additionally, cross-image components are added to the self-attention mechanism by expanding the Key and Value matrices. To do without shifting from the target style, we align the statistics of the Value matrices. As is demonstrated in a comprehensive battery of qualitative and quantitative experiments, our method effectively decouples style from subject appearance and enables faithful generation of text-aligned images with consistent characters across diverse styles.

Code will be available at our project page: jbruner23.github.io/consistyle.

CCS Concepts: • **Computing methodologies** → **Computer graphics**; **Machine learning**.

This is the author version of a paper accepted to SIGGRAPH Asia 2025.
The final published version appears in ACM Transactions on Graphics.

1 Introduction

In visual storytelling, from comics to animation to movies, the same character often traverses diverse stylistic worlds. Whether rendered as a hyper-realistic portrait, a minimal sketch, or even a pixel art

*Denotes equal contribution.

Authors' Contact Information: Yohai Mazuz, yohaimazuz@mail.tau.ac.il, Tel Aviv University, Israel; Janna Bruner, jannabruner@mail.tau.ac.il, Tel Aviv University, Israel; Lior Wolf, wolf@cs.tau.ac.il, Tel Aviv University, Israel.

figure in a parody sequence, the human eye can perceive subjects of different styles as the same character, see Fig. 1. However, since style is a crucial part of the overall appearance, maintaining identity while varying style poses a tremendous technical challenge.

Text-to-image diffusion models [5, 21, 24] have made significant progress in generating high-quality, stylized images from text prompts, enabling the creation of diverse and complex visuals. Yet, these models typically generate each image independently, making it difficult to preserve consistent subject identity across multiple images or prompts.

There are three factors we would like to control independently: (i) prompt-aligned scene and setting, (ii) prompt-aligned image style, and (iii) cross-image character consistency. These factors have been studied in various partial combinations. Hertz et al. [15] have shown that using attention sharing techniques, the style of generated images can be aligned without pre-training, while Alaluf et al. [2] show how to transfer the appearance of an object in one image to another by mixing attention components between the images. Character consistency has been studied either as a personalization problem [8, 22] or as a consistent generation problem. The former receives the target subject as a set of input images. The latter only requires that the generated subject is fixed among all generated images, and can be either based on finetuning (reducing the problem to that of personalization) [3, 12, 23] or by training-free approaches [26, 32]. Training-free approaches offer prompt-faithful generation, but fall short in maintaining subject consistency across diverse styles.

In this work, we address the problem of generating consistent characters across varying prompts and styles, proposing a training-free framework that aligns both semantic identity and visual style, as shown in Fig. 2. Our method consists of three main stages:

(1) Style extraction: We run SDXL [20] with the desired prompts and record the Value matrices from the self-attention layers. These serve as style anchors in the later diffusion process.
(2) Cross-image attention with style alignment: We modify the self-attention mechanism to allow each image to attend to the others during generation, encouraging subject consistency across the image set, we apply adaptive instance normalization to avoid style leak between images.
(3) Identity alignment: We compute feature correspondences using DIFT [25], and apply the resulting mappings to the Query and Key matrices only. In the early diffusion steps, the previously recorded Values are injected to guide the process toward the desired style distribution.

As far as we can ascertain, this is the first training-free method to jointly decouple style from identity while ensuring both prompt alignment and subject consistency across diverse styles.

Our empirical evaluations demonstrate that our method consistently outperforms prior approaches in both style and prompt alignment, while maintaining subject consistency comparable to existing methods.

2 Related Work

We aim to generate an array of images depicting consistent subjects across diverse styles, enabling more flexible and expressive prompt design. Following previous work, the main characteristics that are concerned with in style are the shapes, textures and colors.

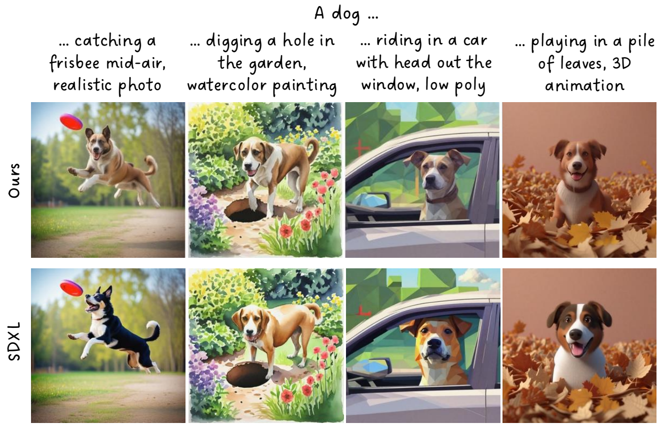

A dog ...
... catching a frisbee mid-air, realistic photo
... digging a hole in the garden, watercolor painting
... riding in a car with head out the window, low poly
... playing in a pile of leaves, 3D animation
Ours
SDXL

Fig. 2. Consistent character generation across diverse styles. Our method preserves key characteristics such as patterns and colors while adhering to the style specified in each prompt. In contrast, SDXL aligns with the prompt and style but fails to maintain consistency across different prompts.

Style Transfer and Style Alignment techniques aim to disentangle visual style from content, enabling the generation of diverse outputs while preserving underlying structure or semantics. Early approaches focused on transferring artistic styles onto photographs [11], whereas more recent methods have emphasized stronger style-content decoupling, preserving spatial structure and identity across a variety of stylistic domains [2, 4, 9]. In the context of style alignment, Hertz et al. [15] propose extending the self-attention mechanism to share attention across a set of images, promoting consistent style across generations. While these methods achieve state-of-the-art results in their respective domains, they are not directly suited to the problem we address. Our work builds upon these advancements, aiming to explicitly distinguish between style and character identity, and to preserve both consistently across varying prompts and visual domains.

Personalization methods aim to condition generative models on a specific subject, enabling consistent synthesis of that subject in new contexts. DreamBooth [23] and Textual Inversion [8] introduced approaches to personalize diffusion models using only a few images of a subject. Though effective, these methods require subject-specific training and often struggle with preserving fidelity across diverse prompts or styles.

Consistent Text-to-Image Generation maintaining consistency of a character or subject across multiple generated images remains a key challenge in text-to-image diffusion. Some methods leverage attention maps, reference encodings, or optimization strategies to enforce identity coherence across generations [3, 26, 32]. However, many of these approaches either depend on personalization or exhibit limited flexibility when prompts vary significantly in content or style. Ensuring prompt alignment while maintaining consistent identity across diverse visual appearances is still an open problem.

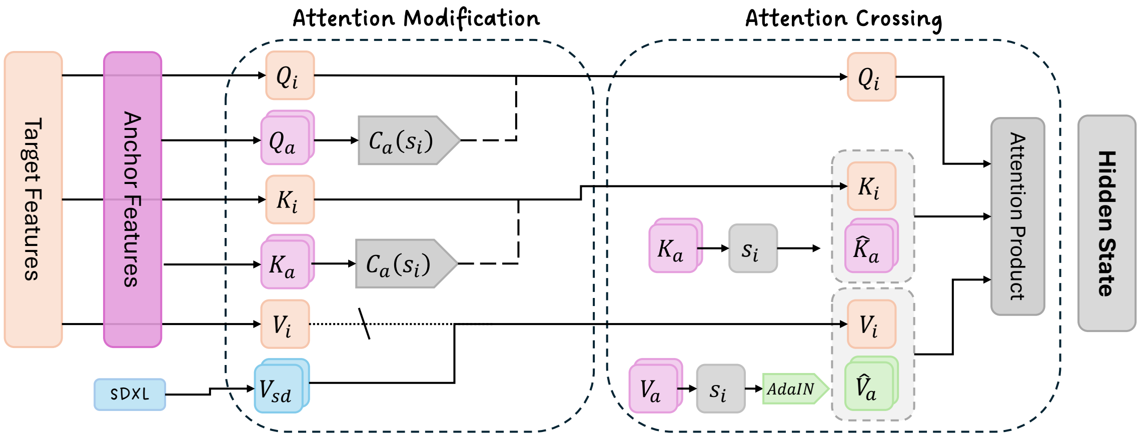


Fig. 3. Overview of our method, illustrating the attention modification and crossing components.

Image Harmonization aims to make composite images appear visually coherent by ensuring that the foreground object aligns stylistically with the background scene. This involves matching lighting, texture, and color distributions to produce seamless and natural-looking compositions. Although harmonization primarily focuses on visual realism rather than identity preservation, techniques from this field [6, 27, 33] inform design choices in style-consistent generation. In our method, extending the attention mechanism and injecting feature correspondences between foreground subjects can occasionally introduce artifacts, disrupting visual harmony between the foreground and background. To address this, we incorporate the Value matrices extracted from SDXL, this guides the generation process toward the original style and composition distribution, resulting in more naturally harmonized outputs.

3 Method

Modern text-to-image (T2I) diffusion models integrate transformer blocks within the U-net layers, allowing patches in the latent space to attend to one another. This process refines image coherence by enabling feature aggregation across the spatial dimensions.

Given input features $X \in \mathbb{R}^{B \times N \times d}$, where B is the batch size, $N = H \times W$ is the number of patches, and d is the feature dimension, self-attention employs three learnable linear projections to compute the query, key, and value matrices:

$$Q, K, V \in \mathbb{R}^{B \times N \times d}, \tag{1}$$

The self-attention mechanism captures the pairwise relationships between patches using the scaled dot-product formulation:

$$\text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d}}\right) V_i, \tag{2}$$

where $i \in [B]$ is the index of an image in the batch. This aggregates context from the entire image, enhancing feature representations for more consistent and contextually accurate outputs. The resulting attended features are then typically projected back to the original feature dimension before being passed to subsequent layers.

3.1 Method Overview

Maintaining subject consistency in training-free approaches for diffusion models remains a significant challenge. Existing methods often rely on injecting hidden states or cross-image key and value sharing within the self-attention layers to preserve subject identity. However, these approaches can inadvertently introduce style misalignment, as hidden states typically encode both semantic and visual features. For instance, injecting the hidden state of a colorful image into a grayscale context can result in unintended color transfer, leading to stylistic inconsistencies.

To address this, our approach focuses on isolating the semantic consistency of subjects while reducing unintended style entanglement. By precisely managing the flow of visual features and separating semantic content from stylistic elements, our method ensures accurate subject alignment without compromising intended appearance. This balance is accomplished through a combination of targeted attention mechanisms and adaptive normalization, designed to preserve structural integrity while maintaining style fidelity.

Our approach for improving style and subject consistency in text-to-image diffusion models is illustrated in Fig. 3. It is a multi-phase process (see Table 1 for a list of the symbols):

(1) **Initial generation** We first run a vanilla generation using the SDXL model. During this pass, we store the intermediate value features V_{sd}, which capture the fine-grained texture and color details needed for consistent style preservation in subsequent generations.
(2) **Correspondence computation** Next, we run a generation to compute a set of DIFT features [25], which are used to establish a correspondence mapping C_α for the subject indices s_α in each image α and only includes attention crossing, a component which allows images to attend to subjects of other images in the self-attention layer and does not rely on the correspondence mapping. The subject indices are obtained from the cross-attention layer using a threshold over the attention map matching the subject token query. The images generated for computing DIFT do not employ the Q and K

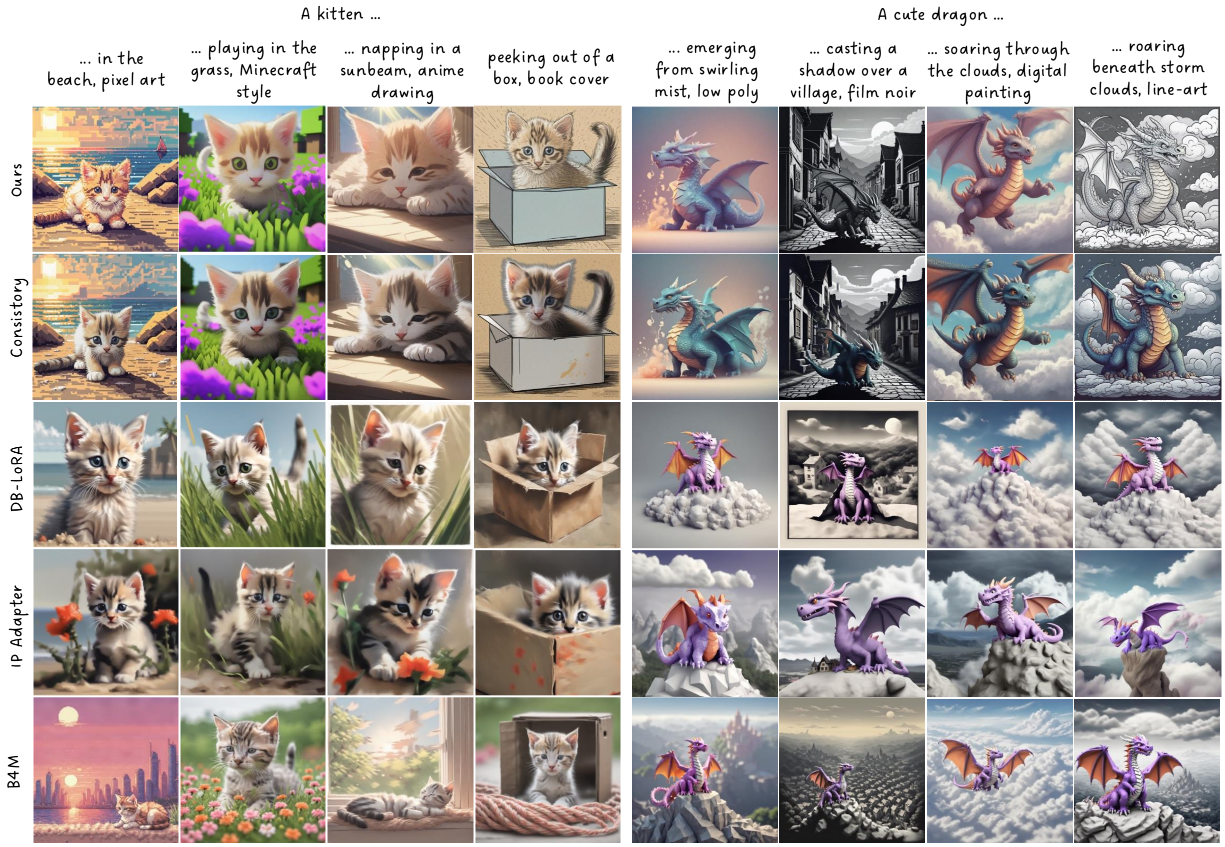

A kitten ...
... in the beach, pixel art
... playing in the grass, Minecraft style
... napping in a sunbeam, anime drawing
peeking out of a box, book cover
A cute dragon ...
... emerging from swirling mist, low poly
... casting a shadow over a village, film noir
... soaring through the clouds, digital painting
... roaring beneath storm clouds, line-art
Ours
Consistory
DB-LoRA
IP Adapter
B4M

Fig. 4. Qualitative comparison of our method with Consistory, DB-LoRA, IP-Adapter and B4M demonstrates its effectiveness across varying text descriptions, character consistency, and style alignment. Unlike other methods, our approach preserves character features and maintains consistent appearance while faithfully adhering to the specified style and textual descriptions.

modifications of Sec. 3.2 since these require the subject location information that is computed using DIFT. However, the component crossing of Sec. 3.3, in which K and V are enriched with key and value pairs from other images in the batch after applying AdaIn is applied to obtain some level of subject consistency while maintaining style.

(3) **Final generation** In the final pass, we perform a full generation that integrates all components: **i. Value-Preservation:** We reuse the stored values from the initial vanilla run, maintaining stylistic consistency, **ii. Attention Transfer:** During the initial phase, we employ the correspondence mapping to inject query and key features, for aligning the subject details across images, and **iii. Attention Crossing:** Throughout the final generation, we apply attention crossing to allow image queries to attend to the Keys and Values of other images in the batch, which improves subject consistency, while using AdaIN to prevent style-leak between images due to the different Value distributions.

Table 1. Symbol Definitions

Symbol	Description
$V_{\mathrm{sd},i}$	self-attention Values of i-th image in the SDXL model
B	the batch size
d, H, W	dimension, height and width in the latent feature space
$N = HW$	the number of total patches
i	the index of ith image in the batch
a	the anchor image index or indices
$Q_\alpha, K_\alpha, V_\alpha$	self-attention queries, keys, values for sample $\alpha \in [B]$
h_α	self-attention hidden state of image α
z_α	self-attention latent of image α
s_α	the subject indices in image α
$C_{\alpha'}(s_\alpha)$	mapping of patches between image α and image α'
$\mathcal{A}$	Adaptive Instance Normalization (AdaIn)

3.2 Transferring Style While Maintaining Appearance

To enhance the consistency of subjects across prompts, we focus on the selective transfer of keys and queries between subjects presented in the array of images at early stages, avoiding value exchange.

Table 2. Comparing our method to other zero-shot methods that achieve style or identity consistency, each with its own distinct goal. The component modification part shows the modification of the presentation of each generated image, while the component import part shows how the self-attention of the model is modified to have a cross attention component. The methods in the table are Consistory [26], Cross-Image Attention [2], StyleAligned [15] and IlluSign [4]. Our method is the only one that has an identity goal as well as a style goal, on top of the prompt faithfulness goal, which requires a much more nuanced solution. Prompt-to-Prompt [14] is not listed as its attention modifications are done in the cross-attention layer and not the self-attention layer.

Method	**Component Modification**				**Component Crossing**	
	h/z	**Q**	**K**	**V**	**K**	**V**
Consistyle (ours)	–	$Q_i[s_i] \leftarrow Q_a[C_a(s_i)]$	$K_i[s_i] \leftarrow K_a[C_a(s_i)]$	$V_i \leftarrow V_{\mathrm{sd},i}$	$K_i \leftarrow [K_i, K_j[s_j]]$	$V_i \leftarrow [V_i, \mathcal{A}(V_j[s_j], V_i)]$
Consistory	$h_i[s_i] \leftarrow h_a[C_a(s_i)]$	–	–	–	$K_i \leftarrow [K_i, K_j[s_j]]$	$V_i \leftarrow [V_i, V_j[s_j]]$
Cross-Image Atten.	$z_i \leftarrow \mathcal{A}(z_i, z_a)$	–	$K_i \leftarrow K_a$	$V_i \leftarrow V_a$	–	–
StyleAligned	–	$Q_i \leftarrow \mathcal{A}(Q_i, Q_a)$	$K_i \leftarrow \mathcal{A}(K_i, K_a)$	–	$K_i \leftarrow [K_i, K_a]$	$V_i \leftarrow [V_i, V_a]$
IlluSign	–	$Q_i \leftarrow Q_i + \frac{1}{2}Q_a$	$K_i \leftarrow K_a$	$V_i \leftarrow V_a$	–	–

This approach reduces the risk of unintended style propagation by preserving fine-grained details while maintaining subject structure and identity. To align with the subjects' structural differences, we use a correspondence mapping.

We aim to transfer style from the target image while adopting the semantic content of the source images. To achieve this, we first perform a vanilla pass through the SDXL model, during which we store the value matrices V_{sd} from the self-attention layer of the decoder at the highest-resolution transformer block $[64 \times 64]$, and only during the early diffusion steps, specifically at steps $\frac{n}{10}$ to $3\frac{n}{10}$ (where n is the number of steps). We also obtain subject masks $s_1, s_2, \dots$ using a threshold over the attention maps matching the subject token query in the cross-attention layer.

In the subsequent Consistyle pass, we apply the DIFT-based feature mapping between queries and keys, and inject the stored values at the corresponding layers and diffusion steps to:

$$\begin{aligned} K_i[s_i] &\leftarrow K_a[C_a(s_i)] \\ Q_i[s_i] &\leftarrow Q_a[C_a(s_i)] \\ V &= V_{sd} \end{aligned} \tag{3}$$

where $Q_i, K_i \in \mathbb{R}^{N \times h \times C}$, h are the attention heads and C number of features in the corresponding decoder layer $\in [32, 64]$, a is the anchor image index(ices), and C_a is the patch mapping induced from DIFT features [25] of the patches computed during a previous run of the model between an anchor image and the target image, note that if there are multiple anchor images, the most similar patch across the anchors is used.

3.3 AdaIN for Style Preservation in Attention Crossing

As direct attention components injection can align the subject's details across images, in order to converge the subjects to the same structure we use an attention crossing component in which Queries may attend to Keys and Values of the images in the batch. Although it leads to improved consistency, the incorporation of Values between different images can lead to style contamination, as Values inherently carry fine-grained texture and color details.

Our approach mitigates this by applying adaptive instance normalization (AdaIN) [17] to the values before cross-subject attention, effectively isolating semantic content from style-specific features, since the statistical distribution of features is a key aspect of style, and texture is often defined in terms of such statistics [13]. By matching feature statistics using AdaIN [17], we preserve the intended style: merely normalizing to constant values would distort the distribution and alter the resulting image style.

The $\mathcal{A}$ operator operation is defined as follows,

$$\mathcal{A}(x, y) = \sigma(y)\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \mu(y) \tag{4}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation functions, respectively. The subsequent attention crossing is thus, for $i \in [B]$

$$V_i \leftarrow \left[\mathrm{AdaIN}(V_1[s_1], V_i), \cdots, V_i, \cdots, \mathrm{AdaIN}(V_n[s_n], V_i)\right], \tag{5}$$

$$K_i \leftarrow [K_1[s_1], \cdots, K_i, \cdots, K_n[s_n]], \tag{6}$$

where K_l, V_l are the keys and values in the self-attention layer matching image l, and s_l are the mask indices of the subjects in the image.

3.4 Summarizing the differences from other methods

With a clear view of the method and the associated terminology, we revisit the comparison to previous work, now on a clear technical level. A comparison to the most similar contributions can be found in Table 2. The table separates the modification of the self-attention mechanism from the step of selectively importing content from other attention maps, creating a cross-attention scheme. We observe notable differences in the handling of self-attention component imports and modifications across various methods. Most approaches rely on direct Value imports from the anchor image. For instance, StyleAligned and Consistory both utilize direct Value imports, with Consistory further incorporating hidden state injections for enhanced visual consistency. Cross-Image Attention [2] similarly relies on direct Value modification, emphasizing precise texture and color transfer.

In contrast, our approach avoids direct Value imports to prevent the style misalignment typically associated with direct appearance transfer. Instead, we employ targeted Value modifications, aligning values to the original *SDXL* features V_{sd} for improved style alignment. Additionally, we apply AdaIN to regulate the statistical properties of imported Values, ensuring smoother integration into the target domain. Unlike Consistory and Cross-Image Attention, we do not use the hidden values of the feature embedding for the

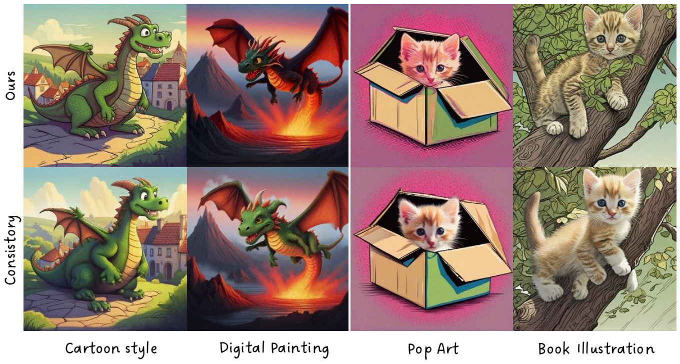


Fig. 5. *Harmonization.* Our method preserves the desired style, seamlessly integrating characters into stylized contexts such as cartoons or illustrations. It adapts both the appearance and the setting, e.g., casting firelit shadows on a dragon or applying a pinkish tone to a kitten in a similar environment.

self-attention intervention, and instead modify Q, K, and V selectively, giving V a different treatment. StyleAligned modified only Q and K, using AdaIn that in our method is applied for the V of the dictionary expansion part (the crossing of attentions).

There are also several differences, which are not captured in Table 2. For example, the timing of the attention intervention varies across methods. Key and Query modifications as well as the V_{sd} injection in Consistyle, are confined to early generation steps, while other modifications and imports are applied across all time steps as specified in Sec. 3.2. Also, the Keys and Queries modification is done only on the text-guided part of the batch in the self-attention layer, where the V_{sd} injection and attention crossing components are also applied on the non-guided elements, which seems to yield a slight improvement.

4 Experiments

We evaluate our method using three core metrics: prompt alignment, consistency, and style alignment, comparing it against four state-of-the-art (SOTA) baselines designed for consistent image generation. The first baseline is a training-free Consistory approach [26], the second is an encoder-based method, IP-Adapter [30], and the third and fourth are personalization training-based methods, DreamBooth-LoRA (DB-LoRA) [23] and Break-for-Make (B4M) [29]. For training details of our baseline models, please refer to the supplementary material. Unlike Consistory, which directly modifies the diffusion process, both IP-Adapter and DB-LoRA utilize a single reference image for personalization.

4.1 Implementation Details and Latency

Our method was evaluated on an A100 GPU (40GB) with a batch size of 5. SDXL and IPAdapter requires 12 sec per image. Consistory involves two passes, amounting to 24 sec per image, while our method requires three passes, totaling 36 sec per image. In contrast, training-based baselines involve substantially higher computational overhead. DB-LoRA requires around 10 minutes of training per subject. B4M entails 70 min per subject, 70 min per style, and an additional 70 min for each subject–style combination, resulting in 210 min for a single subject–style pair.

4.2 Qualitative Results

Our method is designed to accommodate a wide range of visual styles, including highly detailed photographic renderings, abstract illustrations and 3D aesthetics. In Fig. 4 and Fig. 8 we illustrate its ability to preserve style, maintain character consistency, and align with textual descriptions. As shown, B4M, IP-Adapter, and DB-LoRA exhibit strong character consistency, yet they fail to respect the intended style and often feature very little variation. Notably, in both subjects these three methods render the kitten using a realistic style and the dragon with a 3D animation style across all images, disregarding the specified stylistic variations due to overfitting to the reference image. In Fig. 8, B4M shows better alignment to the style but the other two baselines fail in this. Moreover, we observe that LoRA-based methods tend to suffer from overfitting and require an exhaustive hyperparameter search grid to maintain both content and style. Consistory demonstrates strong character consistency and faithful prompt alignment; however, it frequently fails to harmonize the subject with the stylistic setting. In Fig. 4, this issue is evident in the dragon images for the film noir and line-art styles—both intended to be black and white, yet Consistory generates colored images or subjects. This misalignment extends beyond color. For example, in the kitten images (first and fourth row), the subject exhibits a highly realistic texture that clashes with the surrounding stylistic context. Furthermore, as shown in Fig. 5, although the compositions may initially seem coherent, closer examination reveals integration flaws—characters often appear visually ‘stitched’ onto the background rather than naturally embedded within the scene. In contrast, our approach maintains both consistency and stylistic harmony, allowing characters to seamlessly integrate into their environments and preserving the intended aesthetic. The application of our method to multiple anchor images is demonstrated in Fig. 7.

4.3 Quantitative Results

To evaluate our approach, we constructed a dataset of 25 prompt groups, each consisting of a subject description paired with ten distinct prompts. The subjects were divided into four categories: humans, animals, fantasy creatures, and inanimate objects, generated using ChatGPT. In addition, we curated two style groups, each containing ten diverse styles obtained from an online resource [1]. By combining each prompt group with each style group, we obtained 500 images, organized into 50 unique sets. Each set corresponds to one experiment and consists of a batch of $B = 10$ images, where both prompts and styles vary, ensuring that no prompt or style is repeated within the same batch.

For our evaluation we employ several automated metrics following prior work [3, 18, 26]. Text alignment is measured using CLIP-Score [16], both with and without style descriptions. Subject consistency is evaluated using DreamSim [7] as proposed by [26] with background removal. For style alignment, we measure Gram Matrix distance [10], and for perceptual similarity we measure LPIPS [31], and DINO similarity [19]. These latter metrics (LPIPS and DINO) are

Table 3. Comparison of various methods along perceptual, content, and style metrics. (Mean ± SD)

Method	DreamSim ↓	CLIPScore ↑	CLIPScore, Styled ↑	LPIPS ↓	Gram L2 ↓	DINO ↑
Full Dataset (500 samples)						
Consistyle (ours)	0.40 ± 0.10	**32.84 ± 1.66**	**36.03 ± 1.69**	**0.21 ± 0.06**	**1.25 ± 0.69**	**0.85 ± 0.08**
Consistory	0.33 ± 0.12	32.75 ± 1.53	35.34 ± 1.67	0.27 ± 0.07	1.85 ± 1.13	0.78 ± 0.11
IP Adapter	**0.25 ± 0.08**	30.98 ± 2.03	32.07 ± 2.08	0.43 ± 0.07	2.96 ± 1.57	0.54 ± 0.17
DreamBooth-LoRA	0.28 ± 0.13	32.43 ± 1.74	34.33 ± 2.20	0.42 ± 0.07	2.60 ± 1.32	0.59 ± 0.15
Cross-Img	0.55 ± 0.12	30.68 ± 1.36	33.32 ± 1.52	0.33 ± 0.02	1.43 ± 0.70	0.85 ± 0.08
IlluSign	0.53 ± 0.12	30.98 ± 1.48	33.82 ± 1.31	0.38 ± 0.02	1.40 ± 0.67	0.82 ± 0.08
Subset (100 samples)						
Consistyle (ours)	0.46 ± 0.12	**33.00 ± 1.60**	**36.55 ± 1.18**	**0.30 ± 0.04**	**1.35 ± 0.9**	**0.86 ± 0.08**
B4M	**0.38 ± 0.17**	30.67 ± 2.03	33.30 ± 1.60	0.72 ± 0.02	2.75 ± 1.71	0.55 ± 0.15

Table 4. User study results showing pairwise preference percentages across three criteria. Each pair was rated for style alignment, subject consistency, and text alignment. Tie votes are split equally.

Question	Method A	Method B	A %	B %
Style	DB LoRA	Ours	25.2%	**74.8%**
Subject	DB LoRA	Ours	**64.5%**	35.5%
Text	DB LoRA	Ours	39.7%	**60.3%**
Style	Consistory	Ours	15.1 %	**84.8%**
Subject	Consistory	Ours	46.8%	**53.2%**
Text	Consistory	Ours	40.0%	**60.0%**
Style	DB LoRA	Consistory	21.3%	**78.7%**
Subject	DB LoRA	Consistory	**67.2%**	32.8%
Text	DB LoRA	Consistory	29.4%	**70.6%**

used to evaluate content fidelity between stylized images [28, 29], by comparing each stylized output to vanilla SDXL generations that serve as style references.

We conduct two sets of experiments: one using our full dataset of 500 images, and another using a subset of 100 images. The latter was necessary due to the computational cost of training B4M, which requires approximately four hours per image. The results of these experiments are presented in Table 3. As can be seen, our method outperforms all baselines in the prompt alignment and style alignment scores. However, we note a critical limitation in automated metrics for consistency when operating in style-diverse settings. Automated methods tend to over-penalize stylistic variation, potentially rewarding overly consistent outputs that lack stylistic diversity.

4.4 User Study

Since automatic metrics only partly correlated with human perception, especially when measuring subject consistency when varying style, we conducted a user study. The user study evaluates human preferences regarding style alignment, text alignment, and subject consistency. Each user is exposed to 12 random prompt and style combination. In each, the users are presented with pair of images generated by three methods: Consistory, Consistyle, and DB-LoRA. For each pairwise comparison participants answered three questions, one for each criterion. Users had the option to vote in favor of one set of images or to indicate that both methods performed equally well; in such cases, each method received half a vote. See supplementary material for more details.

The results are depicted in Table 4. As can be seen, our method outperforms the baselines in both style alignment and text alignment, and demonstrates higher subject consistency than the Consistory method. Meanwhile, DB-LoRA was preferred for subject consistency, as it maintains strong consistency across generations, although it often ignores style and textual descriptions, resulting in the lowest scores for those criteria.

4.5 Ablations

To better understand the impact of key design choices in our technique, we conduct a series of ablation studies, each isolating a critical component to assess its influence on consistency, style alignment, and overall image quality. The studied variants are:

(i) Consistory with an increased step budget (75 steps) to equate the runtime to our method,
(ii) Consistyle without query injection,
(iii) Consistyle without key injection,
(iv) Consistyle without both query and key injection,
(v) Consistyle with direct identity injection between images instead of subject-based DIFT mapping,
(vi) Consistyle without AdaIN in the attention crossing.

The quantitative outcomes are summarized in Tab. 5, with representative examples shown in Fig. 9 and Fig. 10. Variant (i) demonstrates that extending Consistory's runtime does not substantially change its performance: our method continues to surpass it across all metrics except DreamSim, consistent with earlier results. Variants (ii)–(v) confirm the importance of the attention injection design. Removing the query, key, or both worsens similarity scores, since the attention injection mechanism is designed to enhance subject consistency. At the same time, its absence slightly improves some style-alignment metrics, as fewer details are transferred across images.

In variant (v), we replace the DIFT-based mapping with a full direct identity injection from an anchor image. Since no patch-based

Table 5. Comparison results of the ablation study. Bold values denote the best results and underlined values indicate the second best.(Mean ± SD)

Ablation Method	DreamSim ↓	CLIPScore ↑	CLIPScore, Styled ↑	LPIPS ↓	Gram L2 ↓	DINO ↑
Consistory (original, 50 steps)	**0.33 ± 0.12**	32.75 ± 1.53	35.34 ± 1.67	0.27 ± 0.07	1.85 ± 1.13	0.78 ± 0.11
(i) Consistory (75 steps)	0.40 ± 0.12	32.79 ± 1.55	35.40 ± 1.67	0.42 ± 0.04	1.57 ± 0.30	0.77 ± 0.11
Consistyle (full)	0.40 ± 0.10	32.84 ± 1.66	**36.03 ± 1.69**	0.21 ± 0.06	1.25 ± 0.69	0.85 ± 0.08
(ii) Consistyle (no Q injection)	0.48 ± 0.11	32.75 ± 1.64	35.99 ± 1.61	0.28 ± 0.03	1.18 ± 0.23	0.87 ± 0.08
(iii) Consistyle (no K injection)	0.45 ± 0.12	32.73 ± 1.59	35.94 ± 1.71	0.32 ± 0.04	1.23 ± 0.25	0.84 ± 0.08
(iv) Consistyle (no QK injection)	0.48 ± 0.11	32.60 ± 1.66	35.80 ± 1.65	**0.18 ± 0.03**	**0.72 ± 0.11**	**0.94 ± 0.05**
(v) Consistyle (w/o DIFT)	0.43 ± 0.12	**32.88 ± 1.52**	35.74 ± 1.57	0.49 ± 0.03	2.19 ± 1.41	0.74 ± 0.14
(vi) Consistyle (no AdaIN)	0.42 ± 0.12	32.81 ± 1.59	35.66 ± 1.71	0.23 ± 0.03	1.39 ± 0.27	0.82 ± 0.09

mapping is available and subject segmentations vary in size, we inject keys and values from the entire image. In addition, since no distance measures are available to compare across different anchors, the injection is restricted to the first image in the batch. As shown in Tab. 5 and illustrated in Fig. 9, this variant leads to feature leakage across non-subject regions, resulting in degraded style-alignment scores. In particular, the image backgrounds shift noticeably in their tone compared to their original SDXL counterparts, unlike other variants where the backgrounds remain nearly unchanged.

Finally, variant (vi) highlights the role of AdaIN: without it, values are transferred directly across images, which weakens style alignment, as reflected by a pronounced drop in the Gram metric and visible artifacts in Fig. 10.

Overall, all ablated versions perform worse than our full model, underscoring the necessity of each component in achieving both strong consistency and style fidelity.

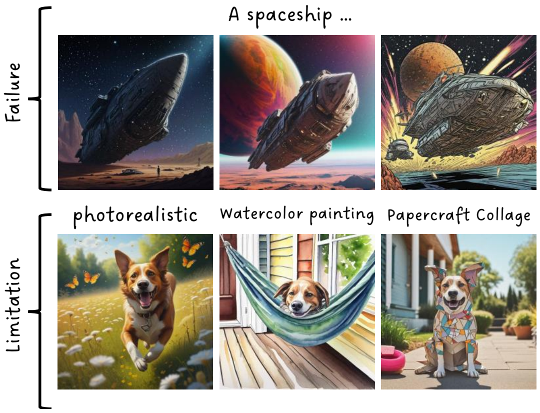


Fig. 6. Demonstration of the method's limitations. The first row illustrates inconsistencies in generating a complex object (spaceship), where high visual detail leads to variation across images. The second row highlights a failure to align with a distinct style-specifically, the Papercraft Collage style, evident in the face details.

5 Limitations

Our approach has several limitations, as can be seen in Fig. 6. First, similar to Consistory [26], it can produce suboptimal results when the correspondence mappings or cross-attention masks fail to accurately capture the intended relationships between images. Second, consistency can degrade for subjects with complex structures, such as large vehicles like boats and spaceships, where fine details are often challenging to maintain. This effect can extend to human subjects, where intricate facial features or clothing elements may vary across generations. The phenomenon is more prevalent when the initial subject interpretations differ significantly between images, such as generating a traditional wooden ship in one image and a modern yacht in another for the same "boat" prompt. While this issue can sometimes be mitigated by selecting different seeds, it remains a potential weakness in cases where precise subject alignment is critical. Finally, highly distinctive artistic styles, such as Papercraft Collage, Voxel Art or other niche 3D aesthetics, can occasionally lead to style misalignment, as these styles often introduce unique structural deformations or exaggerated textures that challenge the statistical alignment technique we employ.

We also observed artifacts in some generated images. To verify that these do not originate from our method, we manually labeled 400 samples as having either minor artifacts (e.g., small visual glitches) or severe ones (e.g., extra limbs). We found that minor artifacts occurred in 18.8% of SDXL images, 20.0% of ours, and 23.2% of Consistory; major artifacts appeared in 2.0%, 2.2%, and 4.0% respectively, which indicates that this stems from the base model.

6 Conclusion

We present Consistyle, a training-free approach for improving consistency while preserving style alignment in text-to-image generation. Our method leverages attention manipulation to enable controlled characteristic sharing between images, even in cases with significant appearance differences. This demonstrates that consistent image generation is feasible in style-diverse contexts, despite the typical entanglement of style and content that often challenges.

Acknowledgments

This work was supported by a grant from the Tel Aviv University Center for AI and Data Science (TAD). We would like to thank Yoad Tewel for his insightful feedback.

References

[1] Karlheinz Agsteiner. 2023. The 77 Styles of Stable Diffusion's SDXL. https://medium.com/@karlheinz.agsteiner/the-77-styles-of-stable-diffusions-sdxl-28303582fcf7. Accessed: 2025-08-24.

[2] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2023. Cross-Image Attention for Zero-Shot Appearance Transfer. arXiv:2311.03335 [cs.CV]

[3] Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2024. The Chosen One: Consistent Characters in Text-to-Image Diffusion Models. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) *(SIGGRAPH '24)*. Association for Computing Machinery, New York, NY, USA, Article 26, 12 pages. doi:10.1145/3641519.3657430

[4] Janna Bruner, Amit Moryossef, and Lior Wolf. 2025. IlluSign: Illustrating Sign Language Videos by Leveraging the Attention Mechanism. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*.

[5] Huiwen 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. 2023. Muse: Text-To-Image Generation via Masked Generative Transformers. *ArXiv* abs/2301.00704 (2023). https://api.semanticscholar.org/CorpusID:255372955

[6] Wenyan Cong, Junyan Cao, Li Niu, Jianfu Zhang, Xuesong Gao, Zhiwei Tang, and Liqing Zhang. 2021. Deep Image Harmonization by Bridging the Reality Gap. In *British Machine Vision Conference*. https://api.semanticscholar.org/CorpusID:232428225

[7] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. In *Advances in Neural Information Processing Systems*, Vol. 36. 50742–50768.

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. *ArXiv* abs/2208.01618 (2022). https://api.semanticscholar.org/CorpusID:251253049

[9] Junyao Gao, Yanchen Liu, Yanan Sun, Yinhao Tang, Yanhong Zeng, Kai Chen, and Cairong Zhao. 2024. StyleShot: A Snapshot on Any Style. *ArXiv* abs/2407.01414 (2024). https://api.semanticscholar.org/CorpusID:270870813

[10] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015).

[11] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 2414–2423. https://api.semanticscholar.org/CorpusID:206593710

[12] Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. 2023. Interactive Story Visualization with Multiple Characters. *SIGGRAPH Asia 2023 Conference Papers* (2023). https://api.semanticscholar.org/CorpusID:258960665

[13] R.M. Haralick. 1979. Statistical and structural approaches to texture. *Proc. IEEE* 67, 5 (1979), 786–804. doi:10.1109/PROC.1979.11328

[14] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. (2022).

[15] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. 2023. Style Aligned Image Generation via Shared Attention. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 4775–4785. https://api.semanticscholar.org/CorpusID:265608730

[16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 7514–7528. doi:10.18653/v1/2021.emnlp-main.595

[17] Xun Huang and Serge Belongie. 2017. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 1510–1519. doi:10.1109/ICCV.2017.167

[18] Gao Junyao, Liu Yanchen, Sun Yanan, Tang Yinhao, Zeng Yanhong, Chen Kai, and Zhao Cairong. 2024. StyleShot: A Snapshot on Any Style. *arXiv preprint arxiv:2407.01414* (2024).

[19] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. DINOv2: Learning Robust Visual Features without Supervision.

[20] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=di52zR8xgf

[21] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 10674–10685. https://api.semanticscholar.org/CorpusID:245335280

[22] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation. (2022).

[23] Simo Ryu. 2023. Low-rank Adaptation for Fast Text-to-Image Diffusion Fine-tuning. https://github.com/cloneofsimo/lora. Accessed: 2025-05-19.

[24] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *ArXiv* abs/2205.11487 (2022). https://api.semanticscholar.org/CorpusID:248986576

[25] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. 2023. Emergent Correspondence from Image Diffusion. *ArXiv* abs/2306.03881 (2023). https://api.semanticscholar.org/CorpusID:259089017

[26] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. 2024. Training-Free Consistent Text-to-Image Generation. arXiv:2402.03286 [cs.CV] https://arxiv.org/abs/2402.03286

[27] Yi-Hsuan Tsai, Xiaohui Shen, Zhe L. Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. 2017. Deep Image Harmonization. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 2799–2807. https://api.semanticscholar.org/CorpusID:1033001

[28] Ye Wang, Tongyuan Bai, Xuping Xie, Zili Yi, Yilin Wang, and Rui Ma. 2025. SigStyle: Signature Style Transfer via Personalized Text-to-Image Models. https://wangyephd.github.io/projects/sigstyle.html. arXiv preprint.

[29] Yu Xu, Fan Tang, Juan Cao, Yuxin Zhang, Oliver Deussen, Weiming Dong, Jintao Li, and Tong-Yee Lee. 2025. B4M: Breaking Low-Rank Adapter for Making Content-Style Customization. *ACM Transactions on Graphics (TOG)* 44, 4 (2025), 1–15. doi:10.1145/3728461

[30] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. (2023).

[31] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

[32] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. 2024. StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation. *NeurIPS 2024* (2024).

[33] Ziyue Zhu, Zhao Zhang, Zheng Lin, Ruiqi Wu, Zhi Chai, and Chunle Guo. 2022. Image Harmonization by Matching Regional References. *ArXiv* abs/2204.04715 (2022). https://api.semanticscholar.org/CorpusID:248085801

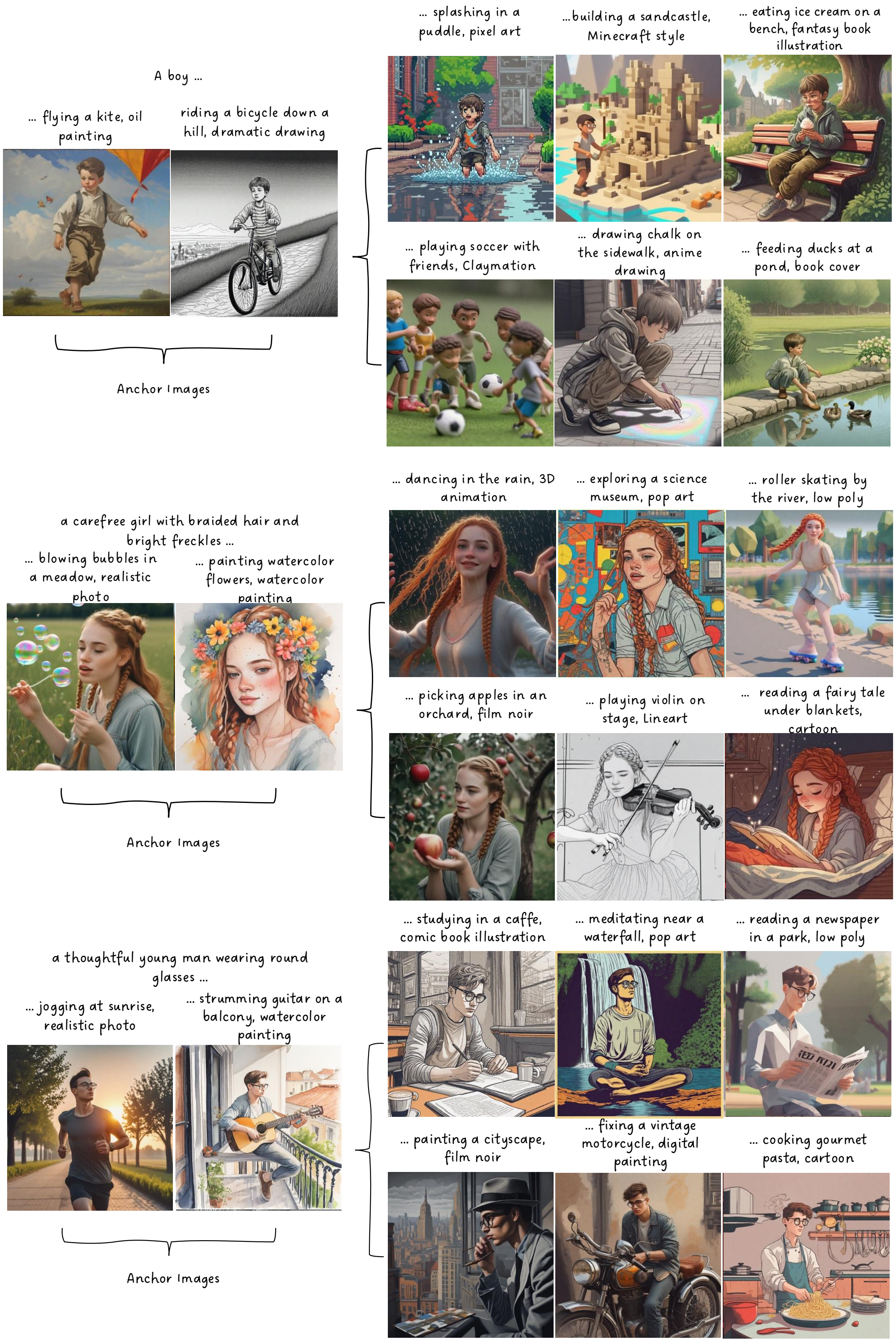

A boy ...
... flying a kite, oil painting
riding a bicycle down a hill, dramatic drawing
Anchor Images
... splashing in a puddle, pixel art
...building a sandcastle, Minecraft style
... eating ice cream on a bench, fantasy book illustration
... playing soccer with friends, Claymation
... drawing chalk on the sidewalk, anime drawing
... feeding ducks at a pond, book cover
a carefree girl with braided hair and bright freckles ...
... blowing bubbles in a meadow, realistic photo
... painting watercolor flowers, watercolor painting
Anchor Images
... dancing in the rain, 3D animation
... exploring a science museum, pop art
... roller skating by the river, low poly
... picking apples in an orchard, film noir
... playing violin on stage, Lineart
... reading a fairy tale under blankets, cartoon
a thoughtful young man wearing round glasses ...
... jogging at sunrise, realistic photo
... strumming guitar on a balcony, watercolor painting
Anchor Images
... studying in a caffe, comic book illustration
... meditating near a waterfall, pop art
... reading a newspaper in a park, low poly
... painting a cityscape, film noir
... fixing a vintage motorcycle, digital painting
... cooking gourmet pasta, cartoon

Fig. 7. Qualitative results highlighting the consistency, style alignment, and textual coherence of our method, guided by two anchor images.

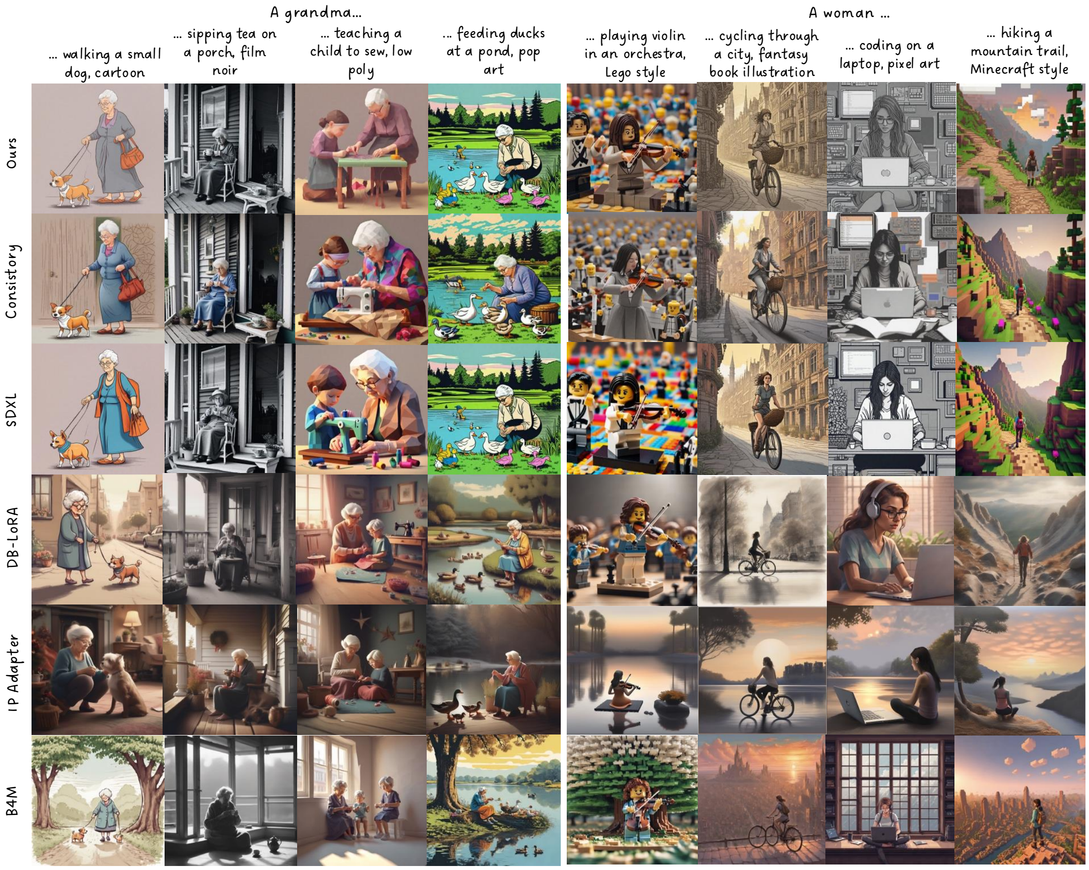

A grandma...
... walking a small dog, cartoon
... sipping tea on a porch, film noir
... teaching a child to sew, low poly
... feeding ducks at a pond, pop art
A woman ...
... playing violin in an orchestra, Lego style
... cycling through a city, fantasy book illustration
... coding on a laptop, pixel art
... hiking a mountain trail, Minecraft style
Ours
Consistory
SDXL
DB-LoRA
IP Adapter
B4M

Fig. 8. Additional comparison.

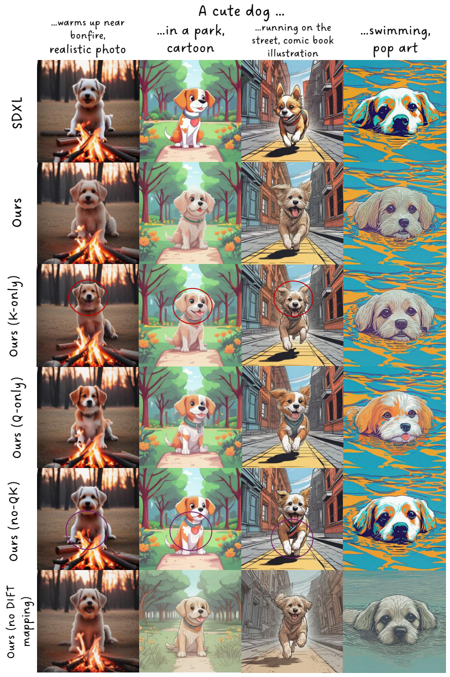


Fig. 9. Ablation study of the attention transfer component. We compare our full method vs variations with partial or no transfer of Keys and Queries. Inconsistent details such as color mismatch are marked.

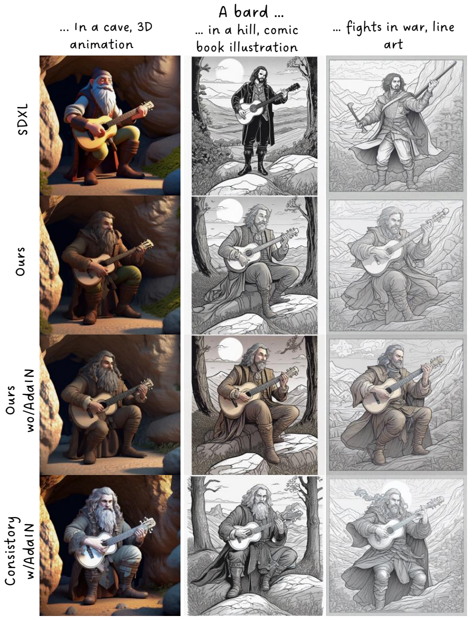

A bard ...
... In a cave, 3D animation
... in a hill, comic book illustration
... fights in war, line art
SDXL
Ours
Ours wo/AdaIN
Consistory w/AdaIN

Fig. 10. Effect of AdaIN on style alignment. We compare results across original outputs, Consistyle, Consistyle without AdaIN, and Consistory. Notice the color shifts, especially when AdaIN is removed or Consistory is used.

APPENDIX

A Training Details for DB LoRA and B4M

In this section, we provide a detailed overview of the training procedures used for the baselines we trained—DreamBooth LoRA (DB LoRA) and Break for Make (B4M).

A.1 DreamBooth LoRA

Workflow. Instance images were generated with the first prompt of the prompt group, combined with the appropriate style from the style group. It was paired with class images generated with random seeds on the concept token to improve results while mitigating overfitting. Instance prompts were constructed by substituting the concept token with a unique placeholder token.

Parameter Details. Models were trained for 250 steps at a resolution of 1024×1024, using a batch size of 1 with gradient accumulation over 3 steps. Mixed precision (`fp16`) and gradient checkpointing were applied to reduce memory usage. Optimization employed 8-bit Adam with a constant learning rate of 1×10^{-4} and no warmup. SNR weighting was applied with $\gamma = 5.0$, and seeds were clamped to the 32-bit integer range for reproducibility.

A.2 Break for Make

The training procedure for B4M consists of three phases:

Phase 1: Content LoRA. We first train a LoRA for the content reference (e.g., "a kitten"). For stable results, it is recommended to use at least three images. To this end, we generate one image with vanilla SDXL and two additional images with IP-Adapter in a "photo-realistic" style. These prompts are deliberately different from those used in our evaluation templates. To disentangle content from style, we additionally provide three style reference images, as suggested by the original authors, chosen from style categories not included in our evaluation set. Training is performed for 1000 steps, as recommended. However, we observe that overfitting may occur; in some cases, reducing the training to 500 steps yields better results.

Phase 2: Style LoRA. Next, we train a LoRA for style references. We generate three images (e.g., landscape, cat, etc.) with a specific style description (e.g., pixel art, line-art). As in Phase 1, we train for 1000 steps and provide three additional content images to encourage disentanglement between content and style.

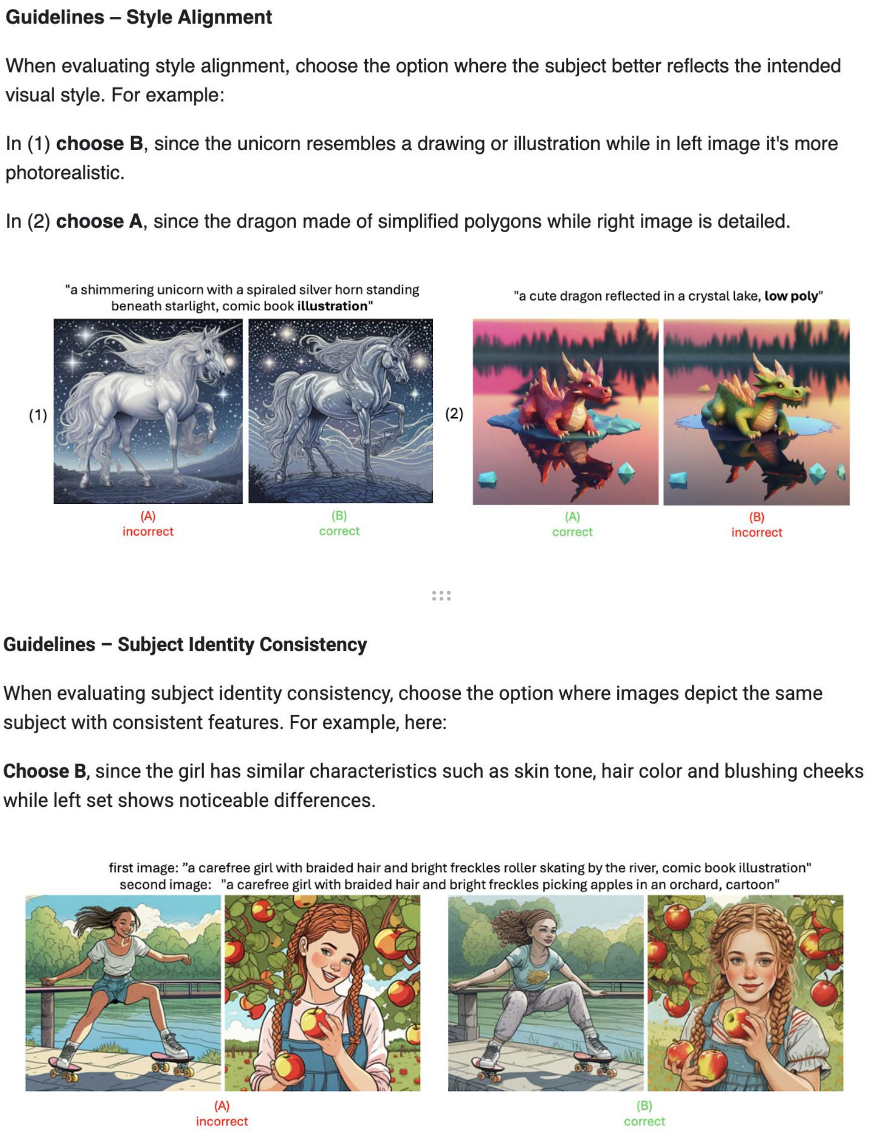
Phase 3: Content–Style LoRAs. In the final phase, we train separate LoRAs for each content–style pair, again for 1000 steps.

Inference. At inference time, images are generated by composing the learned tokens corresponding to content and style. For example, a prompt could be: *"snq woman coding on a laptop, w@z pixel art style."*

B User Study Instructions and Questions

In the user study, participants were asked to choose one set of images per method. Before answering any questions, they were instructed to carefully read the guidelines, which explained the relevant terminology and outlined what to look for when evaluating subject identity, text alignment, and style alignment. For each task, an example was provided to illustrate the evaluation criteria. The full set of guidelines is shown in Fig. 11.

Each participant was presented with 12 comparison sets, each containing four randomly selected images per method, as shown in Fig. 12. For each set, participants answered three questions corresponding to the three evaluation tasks. An example of the user study questions is shown in Fig. 13.



Guidelines – Style Alignment

When evaluating style alignment, choose the option where the subject better reflects the intended visual style. For example:

In (1) **choose B**, since the unicorn resembles a drawing or illustration while in left image it's more photorealistic.

In (2) **choose A**, since the dragon made of simplified polygons while right image is detailed.

"a shimmering unicorn with a spiraled silver horn standing beneath starlight, comic book **illustration**"

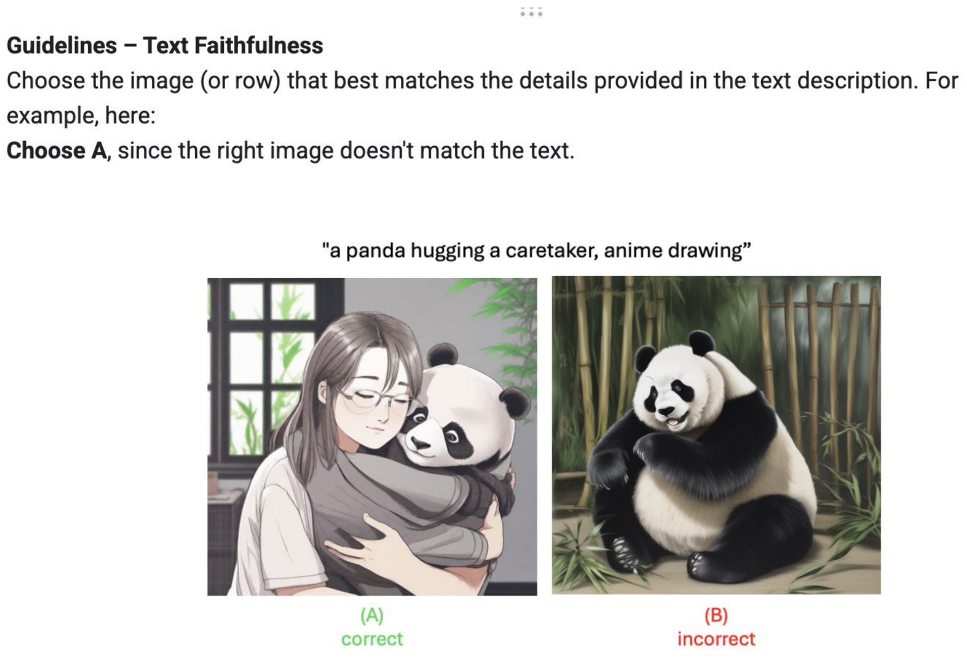
"a cute dragon reflected in a crystal lake, **low poly**"

Guidelines – Subject Identity Consistency

When evaluating subject identity consistency, choose the option where images depict the same subject with consistent features. For example, here:

Choose B, since the girl has similar characteristics such as skin tone, hair color and blushing cheeks while left set shows noticeable differences.

first image: "a carefree girl with braided hair and bright freckles roller skating by the river, comic book illustration"
second image: "a carefree girl with braided hair and bright freckles picking apples in an orchard, cartoon"



Guidelines – Text Faithfulness

Choose the image (or row) that best matches the details provided in the text description. For example, here:

Choose A, since the right image doesn't match the text.

"a panda hugging a caretaker, anime drawing”

(A)
correct

(B)
incorrect

Fig. 11. Guidelines from our user study.

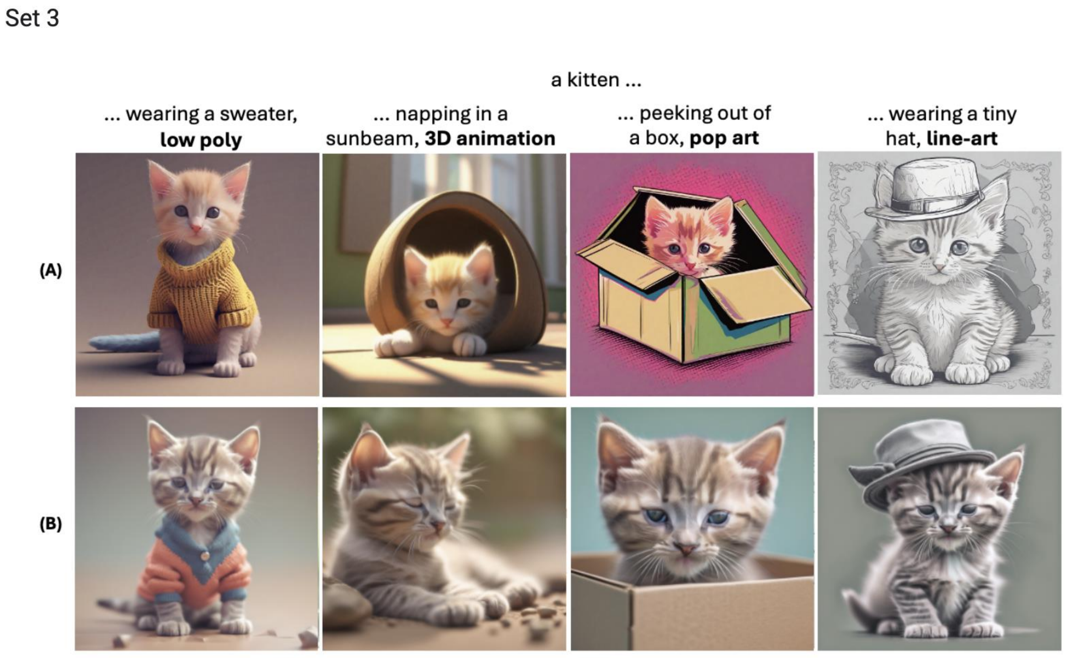
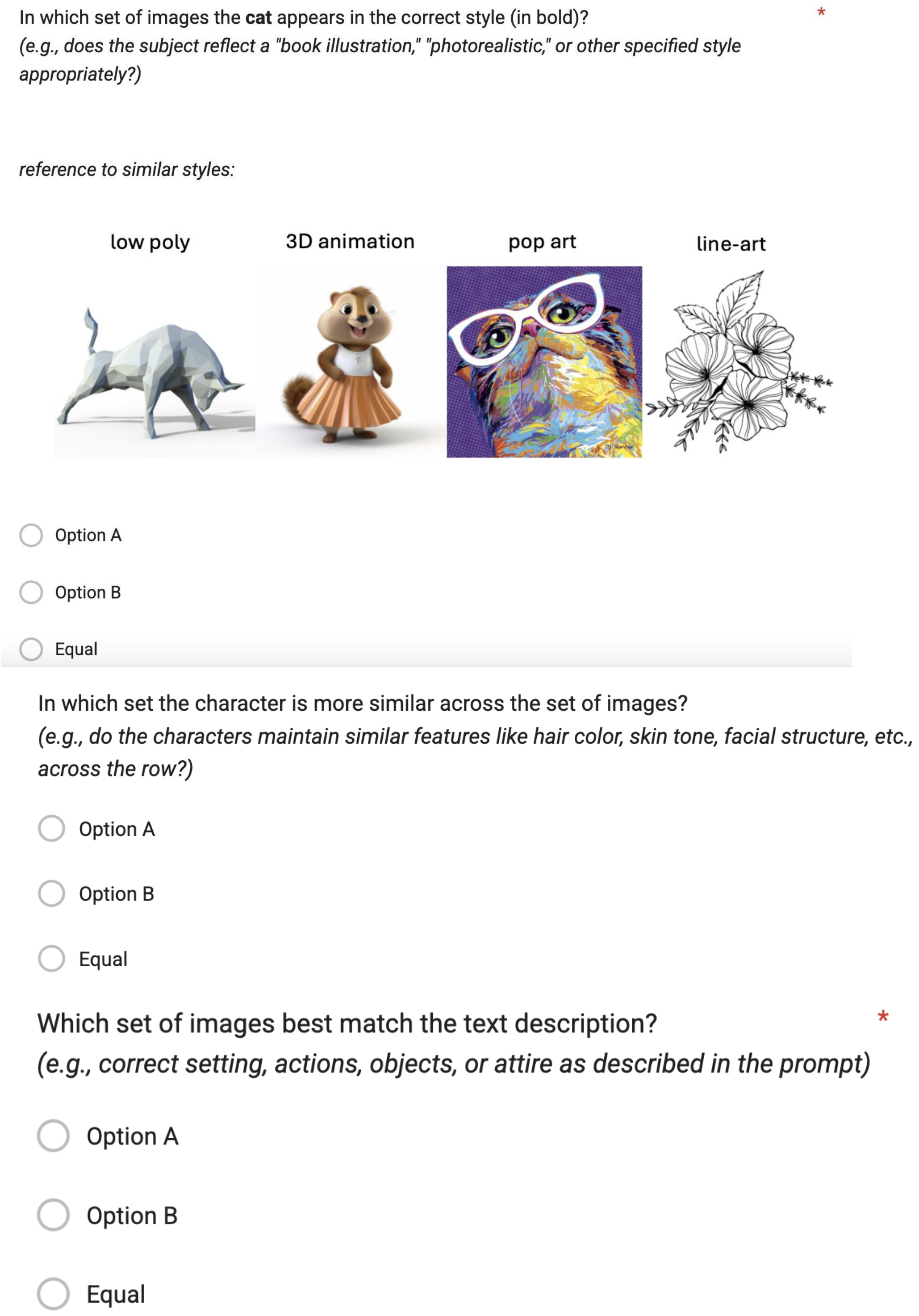

Set 3
a kitten ...
... wearing a sweater, **low poly**
... napping in a sunbeam, **3D animation**
... peeking out of a box, **pop art**
... wearing a tiny hat, **line-art**
(A)
(B)

Fig. 12. An example set of images used for comparison



In which set of images the **cat** appears in the correct style (in bold)? *

(e.g., does the subject reflect a "book illustration," "photorealistic," or other specified style appropriately?)

reference to similar styles:

○ Option A

○ Option B

○ Equal

In which set the character is more similar across the set of images?

(e.g., do the characters maintain similar features like hair color, skin tone, facial structure, etc., across the row?)

○ Option A

○ Option B

○ Equal

Which set of images best match the text description? *

(e.g., correct setting, actions, objects, or attire as described in the prompt)

○ Option A

○ Option B

○ Equal

Fig. 13. For each set of images, the following evaluation questions were asked in the user study.