Structure-Semantic Co-optimized Latent Diffusion Model for Fast Visual Anagram Synthesis

Xiang Gao, Yunpeng Jia*
School of Digital Media and Design Arts
Beijing University of Posts and Telecommunications
{gaoxiang1102, xibei156}@bupt.edu.cn

Abstract

Visual anagram is an intriguing form of art creation wherein a single image presents different conceptual interpretations under transformations such as flipping or rotation. Recent work has achieved visual anagram synthesis by leveraging pretrained text-to-image (T2I) diffusion models, yet still suffers from several key limitations including computational inefficiency, suboptimal aesthetic quality, and weak semantic fidelity and expressiveness. This work focuses on generating visual anagrams with substantially improved visual quality at minimal computational cost, thereby advancing intelligent creation of illusionary digital art. To increase image resolution while reducing time overhead, we adapt the cutting-edge parallel denoising algorithm from pixel-based T2I model to the adversarially distilled latent-based one, and accordingly propose a structure-semantic co-optimization (S2CO) framework to counteract the consequent visual degradation. As the core of our approach, S2CO framework comprises three key innovations: (i) null-text structure alignment optimization; (ii) semantic enhancement optimization; (iii) attention-guided noise fusion. Building upon these components, our method dubbed **S2CO-Anagram** is able to generate higher-resolution anagram images with noticeably superior visual harmony, aesthetic quality, and semantic faithfulness to the text prompt than related SOTA approaches, all while achieving substantially faster inference speed.

1 Introduction

Visual illusions occur when human perception deviates from physical reality. For decades, this intriguing phenomenon has captivated researchers across psychology, neuroscience, and computer vision. Recently, computationally synthesizing visual illusions has blossomed into a compelling area, encompassing algorithms that span classical optimization to modern deep generative models.

Early parametric approaches formulate visual camouflage [11, 63, 85], hybrid perception [62], geometric illusion [16, 9], motion illusion [20], and color illusion [74] as tractable optimization problems. These approaches typically require per-instance optimization and are limited in generalization ability.

With the proliferation of deep learning, GANs [33] achieve remarkable success in synthesizing illusions [32] and camouflages [34], marking a paradigm shift from hand-crafted parametric models to learnable generative models. More recently, techniques have leveraged generative prior of pretrained T2I diffusion models to synthesize optical illusion hidden images via score distillation [5], noise decomposition [30], and phase manipulation [28]. Practical applications fuse functional data with artistic expression, yielding scannable QR codes that double as aesthetic illusions [51, 50]. Collectively, these developments mark a new era in computational visual illusion synthesis.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



Figure 1: Our method enables fast synthesis of high-quality visual anagram images that conceptually conform to different text prompts after undergoing a transformation such as flipping and rotation, offering artists and designers an efficient tool to create optical illusion visual arts.

Among the various forms of optical illusions, visual anagrams, images that transform into distinct percepts upon specific transformations, has garnered increasing attention recently. Diffusion Illusions [5] optimizes an image based on the generative prior of T2I diffusion model, ensuring that its transformed views semantically align with different text prompts. However, this method yields limited resolution and quality, and the reliance on score distillation [65] optimization leads to prohibitive time consumption. SyncTweedies [45] builds a multi-view synchronous denoising architecture, achieving zero-shot anagram generation by fusing the predicted clean images across views at each DDIM [72] denoising step. Visual Anagrams [31], in contrast, achieves the same goal by averaging the multi-view noise predictions during the denoising process. Anagram-MTL [78] improves upon Visual Anagrams by introducing attention-based anti-segregation optimization and CLIP-based noise vector balancing, mitigating issues of "concept segregation" and "concept domination" respectively. However, these three methods rely on pixel-based T2I models, resulting in high computational overhead and limited image resolution. LookingGlass [7] realizes visual anagram generation based on Laplacian pyramid warping. This method achieves high image resolution by building upon latent-based diffusion model, though the repeated image warping and aggregation operations still incur prolonged inference time.

In this paper, we propose S2CO-Anagram, a lightweight latent diffusion model (LDM) for fast and high-quality visual anagram synthesis with example results displayed in Fig. 1. To achieve high resolution and inference speed, S2CO-Anagram inherits the multi-view parallel denoising paradigm [45, 31, 78] and builds upon SDXL-Turbo [1]—an adversarially distilled [69] LDM enabling high-resolution and few-step T2I synthesis. However, directly applying multi-view synchronous denoising to LDM fails to generate visually harmonious and semantically faithful results, a challenge we attribute to the following two problems. **(1) Structural Divergence**: the denoising of multi-view latents structurally diverges under different semantic guidance. For instance, "horse" and "penguin" differ inherently in shape and pose, driving their prompt-guided denoising processes toward different structures. However, averaging their structural signals via multi-view noise fusion forces the latent into a compromised structure that cannot faithfully capture either concept. Moreover, cross-view structural conflicts can lead to rigid, unnatural content fusion, producing less harmonious visual anagrams. These issues are more pronounced in latent-based T2I models especially those requiring few denoising steps. **(2) Semantic Dilution**: the multi-view latents are denoised toward different semantic directions. Averaging the predicted noises inevitably dilutes semantic information, undermining conceptual

faithfulness of the generated result to both text prompts. Consequently, we propose a structure-semantic co-optimization (S2CO) framework to tackle the aforementioned two problems.

To tackle the degradation in semantic fidelity and visual harmony caused by structural divergence, we propose **null-text structure alignment optimization**, which penalizes structural deviation at each denoising step by optimizing the null-text embedding in classifier-free guidance (CFG) with our proposed **View-Cycle Low-Frequency Consistency (VLC) Loss**. The VLC loss aligns spatial structure of the predicted clean latents across views in the DCT (Discrete Cosine Transform) domain, suppressing structural divergence along the synchronous denoising trajectory and thereby remarkably mitigating the corresponding issues of semantic vagueness and structural artifacts.

To mitigate semantic dilution problem caused by noise averaging, we propose semantic enhancement optimization and attention-guided noise fusion. The former approach explicitly enhances semantic distinctiveness by optimizing the input latent at each denoising step to maximally activate vision-text cross-attention maps associated to key concepts. The latter one leverages attention maps to guide spatially adaptive noise fusion, which preserves semantically salient features of both views more faithfully than simple noise averaging, and thereby alleviating semantic blurring.

To conclude, the key contributions of this work are summarized as follows:

- We propose S2CO-Anagram, a lightweight few-step LDM that efficiently generates high-quality visual anagrams via a novel structure-semantic co-optimization algorithm.
- A novel null-text structure alignment optimization is proposed to suppress structural divergence along the multi-view synchronous denoising process, promoting semantic faithfulness and visual harmony of the generated anagram images from a frequency-domain perspective.
- A semantic enhancement optimization which enhances conceptual fidelity and discernibility of the generated results by elevating cross-attention activations associated with key concepts.
- An attention-guided noise fusion scheme is proposed to promote semantic saliency and distinctiveness by leveraging cross-attention maps to guide spatially adaptive noise fusion.
- Our S2CO-Anagram outperforms related approaches both qualitatively and quantitatively in visual anagram generation, all while achieving substantially faster inference speed.

2 Related Work

2.1 Diffusion models and applications

. Since the advent of DDPM [37], diffusion models have soon emerged as the mainstream in generative modeling [15], and have achieved rapid advancements in inference speed [71, 57], conditional generation [68, 4, 81], structural control [84, 60, 10], and concept customization [53, 86, 40]. With LDM [67] dramatically lowing computational overhead by transferring diffusion model from high-dimentional pixel space to low-dimensional feature space, a plethora of visual applications built upon LDM have sprung up, such as text-guided image editing [59, 61, 6, 89], text-driven image-to-image (I2I) translation [73, 27, 23, 22], image colorization [82, 49], image inpainting [14, 54], and image restoration [76, 19, 92]. This work focus on leveraging off-the-shelf T2I diffusion model for high-quality and efficient visual anagram synthesis.

2.2 AI-driven visual art creation

. The intersection of AI and artistic creation has witnessed significant advancements in recent years. The seminal work of Neural Style Transfer [29] achieved stunning success in separating and recombining image content and style using convolutional neural network, reproducing artistic styles of renowned paintings on natural images. The algorithm was later substantially advanced in real-time efficiency [43, 38, 48], arbitrary-style generalization [64, 56, 87], and improved visual quality [52, 8, 46], spawning a wide range of artistic creation applications such as fashion design [41], text stylization [21], and decorated logo generation [3]. With the growing prevalence of GANs, a broader spectrum of visual art creation has emerged within the paradigm of cross-domain I2I translation [39, 91], spanning applications such as artistic painting stylization [80, 25, 24], image cartoonization [26, 42], and artistic font generation [77, 79]. Under the more recent diffusion-based paradigm, researchers have explored adapting large-scale T2I diffusion models for artistic

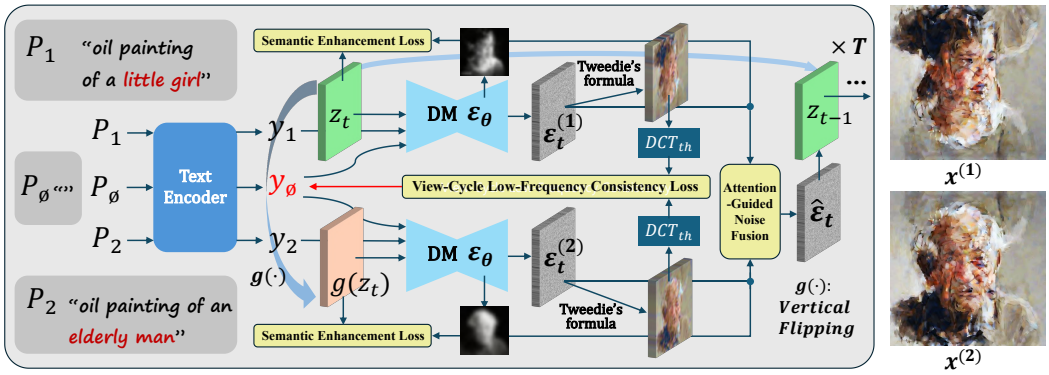


Figure 2: **Method overview**. Our method synchronously denoises two latent views linked by a specified transformation followed by fusing their noise estimations at each denoising step. Grounding on this process, a novel optimization framework comprising null-text structure alignment, semantic enhancement, and attention-guided noise fusion is proposed to enhance the generated results.

style transfer [88, 13, 17] or style-specific content creation [35, 90, 2], elevating the quality and controllability of artistic creation to an unprecedented level. Beyond image-based artistic creation, recent advances have extended artistic stylization into the 3D domain [18, 83, 58], achieving reference-based stylization of arbitrary 3DGS [44] scenes and thereby enabling artistic expression on immersive, freely navigable scenes. Further examples of diffusion-based visual art generation include aesthetic QR code creation [51], optical illusion hidden image generation [28], design image generation [75], artistic typography generation [55], and visual anagram synthesis [31]. This work focuses on generating visual anagrams with substantially improved visual quality at minimal inference time, thereby pushing the frontiers in the creation of illusionary digital art.

3 Proposed Method

This section first clarifies the overall model architecture, then elaborates on the key components of our method, and finally concludes with the full algorithm and implementation details. The preliminary backgrounds of diffusion model and classifier-free guidance (CFG) are described in the **Appendix**.

3.1 Overall Model Architecture

Given a pair of text prompts (P_1, P_2) and a geometric transformation $g(\cdot)$, our method synthesizes a pair of anagram images $(x^{(1)}, x^{(2)})$, where $x^{(1)}$ and $x^{(2)}$ are semantically aligned with P_1 and P_2, respectively, and satisfy $x^{(2)} = g(x^{(1)})$, $x^{(1)} = g^{-1}(x^{(2)})$.

The overall architecture of S2CO-Anagram is illustrated in Fig. 2. Our model is built upon pretrained LDM. The denoising network ϵ_θ is conditioned on the text embeddings (y_1, y_2) which are encoded from (P_1, P_2), the null-text embedding $y_\emptyset$ encoded from $P_\emptyset$="" is employed for CFG. At each denoising step t, the current denoising latent z_t and its transformed view $g(z_t)$ are input to ϵ_θ along with their paired text embeddings, yielding the predicted noises $\epsilon_t^{(1)}$ and $\epsilon_t^{(2)}$, respectively:

$$\epsilon_t^{(1)} = \omega\epsilon_\theta(z_t, t, y_1) + (1-\omega)\epsilon_\theta(z_t, t, y_\emptyset), \quad \epsilon_t^{(2)} = \omega\epsilon_\theta(g(z_t), t, y_2) + (1-\omega)\epsilon_\theta(g(z_t), t, y_\emptyset). \quad (1)$$

Where ω is the guidance strength in CFG [36]. The predicted noises $\epsilon_t^{(1)}$ and $\epsilon_t^{(2)}$ are integrated via the attention-guided noise fusion module, yielding the fused noise $\hat{\epsilon}_t$ that incorporates semantic concepts of both P_1 and P_2. Then, the next-step denoising latent z_{t-1} is computed and iterated as:

$$z_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}(z_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}_t)}{\sqrt{\bar{\alpha}_t}} + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\epsilon}_t, \quad (2)$$

where $\{\bar{\alpha}_t\}$ are DDPM [37] schedule parameters. The multi-view parallel denoising is repeated for T steps to yield the noise-free clean latent z_0. Let D denote the LDM VAE decoder, the final pair of

generated visual anagram images are obtained as:

$$x^{(1)} = \frac{D(z_0) + g^{-1}(D(g(z_0)))}{2}, \;\; x^{(2)} = \frac{D(g(z_0)) + g(D(z_0))}{2}. \tag{3}$$

The VAE decoder D is not transformation-invariant, i.e., $g(D(z_0)) \neq D(g(z_0))$, the averaging-based formulation of Eq. 3 is designed to maximize the transformation consistency between $x^{(1)}$ and $x^{(2)}$.

3.2 Null-Text Structure Alignment

Structural divergence occurs when the multi-view latents are denoised under different prompt guidance, that is, $g^{-1}(\epsilon_t^{(2)})$ tends to structurally deviate from $\epsilon_t^{(1)}$ to cater to different semantic concepts. Such structural misalignment between multi-view noise estimations yields a compromised structure in the fused noise $\hat{\epsilon}_t$, which degrades semantic faithfulness of the resulting z_{t-1} to both y_1 and y_2, and can lead to unnatural structural artifacts in the final results. Hence, we propose an optimization strategy to tackle such structural divergence issue. Since structural information is imperceptible in the noise space, we apply Tweedie's formula so as to align structure based on the predicted clean latents:

$$z_{t\to 0}^{(1)} = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t^{(1)}}{\sqrt{\bar{\alpha}_t}}, \;\; z_{t\to 0}^{(2)} = \frac{g(z_t) - \sqrt{1-\bar{\alpha}_t}\epsilon_t^{(2)}}{\sqrt{\bar{\alpha}_t}}. \tag{4}$$

To enforce structural consistency between $z_{t\to 0}^{(1)}$ and $g^{-1}(z_{t\to 0}^{(2)})$, we apply latent-space DCT transformation to constrain their low-frequency consistency, considering that DCT low-frequency spectral components correspond to global structure and layout. Let $dct_{th}^{(h)}(\cdot)$ denote the function of applying 1D DCT transformation over the h dimension and extracting the frequency components lower than the percentile threshold th, $dct_{th}^{(w)}(\cdot)$ be the correspondingly similar function applied over the w dimension (with details of these two functions illustrated and clarified in the **Appendix**). We design View-Cycle Low-Frequency Consistency (VLC) loss which enforces low-frequency consistency between the predicted clean latents $z_{t\to 0}^{(1)}$ and $g^{-1}(z_{t\to 0}^{(2)})$ for structural alignment:

$$L_{VLC} = \left|\left| dct_{th}^{(h)}(z_{t\to 0}^{(1)}) - dct_{th}^{(h)}(g^{-1}(z_{t\to 0}^{(2)})) \right|\right|_2^2 + \left|\left| dct_{th}^{(w)}(z_{t\to 0}^{(1)}) - dct_{th}^{(w)}(g^{-1}(z_{t\to 0}^{(2)})) \right|\right|_2^2. \tag{5}$$

To maximally disentangle structural alignment with semantic guidance, we minimize L_{VLC} w.r.t. the null-text embedding $y_\emptyset$. That is, y_1 and y_2 define the multi-view semantic concepts of the generated result while $y_\emptyset$ is employed for per-step structural alignment optimization, which is formulated as:

$$y_\emptyset := y_\emptyset - \lambda_1 \nabla_{y_\emptyset} L_{VLC}, \tag{6}$$

where λ_1 is the initial learning rate. By enforcing the two views to share a common low-frequency structure, The distinguishing features between the two concepts (e.g., dog vs. cat) are pushed into high-frequency details. After DCT alignment optimization, the resulting two noise predictions exhibit less structural conflict, leading to less semantic compromise and more harmonious content fusion.

3.3 Cross-Modal Semantic Enhancement

Fusing the multi-view noise predictions $\epsilon_t^{(1)}$ and $\epsilon_t^{(2)}$ into $\hat{\epsilon}_t$ integrates semantic concepts of each view, while also diluting them. To compensate for such semantic dilution issue, we propose cross-modal semantic enhancement optimization which tunes the input latent of ϵ_θ at each denoising step to maximize cross-attention activations associated to key concepts of P_1 and P_2. Cross-attention maps capture correlation intensity between textual semantics and spatial locations, which are computed as:

$$M_l^{(1)} = Softmax\left(\frac{Q_l^{(1)}(K_l^{(1)})^T}{\sqrt{d}}\right), \quad M_l^{(2)} = Softmax\left(\frac{Q_l^{(2)}(K_l^{(2)})^T}{\sqrt{d}}\right), \tag{7}$$

where l is layer index, $M_l^{(1)}$ and $M_l^{(2)}$ denote attention maps at the l-th cross-attention layer inside ϵ_θ given z_t and $g(z_t)$ as input, respectively, $Q_l^{(1)}$ and $Q_l^{(2)}$ are queries of the l-th cross-attention layer derived from z_t and $g(z_t)$, respectively, $K_l^{(1)}$ and $K_l^{(2)}$ are keys of the l-th cross-attention layer originated from y_1 and y_2, respectively. Based on all $\{M_l^{(1)}\}$ and $\{M_l^{(2)}\}$, attention maps

Algorithm 1 Complete algorithm of S2CO-Anagram

Input: A transformation $g(\cdot)$, source view text prompt P_1, transformed view text prompt P_2.
Output: Generated anagram image pair $(x^{(1)}, x^{(2)})$.

Randomly initialize a noise latent $z_T \sim \mathcal{N}(0, I)$.
for $t = T$ to 1 **do**
  Initialize a new null-text embedding $y_\emptyset$ encoded from the null-text prompt $P_\emptyset$="".
  for $k = 1$ to K **do**
    Input z_t and $g(z_t)$ to ϵ_θ, obtaining ($\epsilon_t^{(1)}$, $\epsilon_t^{(2)}$) and computing ($z_{t\to0}^{(1)}$, $z_{t\to0}^{(2)}$), ($\tilde{M}^{(1)}$, $\tilde{M}^{(2)}$).
    Update $y_\emptyset$ with a step of null-text structure alignment optimization using Eq. 6.
    Update z_t with a step of cross-modal semantic enhancement optimization using Eq. 8.
  Recompute noise predictions ($\epsilon_t^{(1)}$, $\epsilon_t^{(2)}$) with the optimized $y_\emptyset$ and z_t using Eq. 1.
  Obtain the fused noise $\hat{\epsilon}_t$ via Eq. 9, based on which compute z_{t-1} via Eq. 2.
Obtain the clean latent z_0 and generate the final anagram image pair ($x^{(1)}$, $x^{(2)}$) using Eq. 3.

corresponding to conceptually most crucial words in P_1 and P_2 (marked in red in Fig. 2) are extracted, normalized, and summed to yield the final $\tilde{M}^{(1)}$ and $\tilde{M}^{(2)}$ (with details described in the **Appendix**) which capture correlation intensity of each spatial location to the key concept of P_1 and P_2, respectively. To compensate for semantic dilution issue, we enhance semantic expressiveness by maximizing pixel activation of $\tilde{M}^{(1)}$ and $\tilde{M}^{(2)}$ w.r.t. the input latent z_t, which is formulated as:

$$z_t := z_t + \lambda_2 \nabla_{z_t} \left(\sum\nolimits_{(i,j)\in\Omega^{(1)}} \tilde{M}_{i,j}^{(1)} + \sum\nolimits_{(i,j)\in\Omega^{(2)}} \tilde{M}_{i,j}^{(2)} \right), \Omega^{(k)} = \left\{ (i,j) | \tilde{M}_{i,j}^{(k)} \geq \tau_p^{(k)} \right\}, \quad (8)$$

where i, j denote spatial indices, λ_2 is the initial learning rate, $\tau_p^{(k)}$ refers to the p-th percentile of the values in $\tilde{M}^{(k)}$. To avoid indiscriminate semantic enhancement that leads to semantic bleeding to all spatial locations, we only back-propagate through locations above the p-th percentile of each attention map, thereby reinforcing semantic expressiveness only in the most concept-relevant regions.

3.4 Attention-Guided Noise Fusion

Observing that direct noise fusion via averaging tends to produce blurring and semantically ambiguous results, which we attribute to the lack of spatial awareness of salient semantic concept, as average noise fusion applies equal-strength weighting to all locations. Therefore, we propose to leverage the concept saliency spatial indication provided by the extracted attention maps $\tilde{M}^{(1)}$ and $\tilde{M}^{(2)}$ for spatially adaptive noise fusion, which is formulated as:

$$\hat{\epsilon}_t = \epsilon_t^{(1)} \cdot \frac{\tilde{M}^{(1)}}{\tilde{M}^{(1)} + g^{-1}(\tilde{M}^{(2)})} + g^{-1}(\epsilon_t^{(2)}) \cdot \frac{g^{-1}(\tilde{M}^{(2)})}{\tilde{M}^{(1)} + g^{-1}(\tilde{M}^{(2)})}. \quad (9)$$

3.5 Structure-Semantic Co-optimization

Integrating the above three key components, we propose a structure-semantic co-optimized LDM for high-quality and fast visual anagram synthesis. At each denoising step, null-text structural alignment and semantic enhancement are jointly carried out prior to the attention-guided noise fusion. The complete algorithm of S2CO-Anagram is shown in Alg. 1. During per-step optimization, we set λ_1=0.1 in Eq. 6, λ_2=0.5 in Eq. 8, and set the DCT low-pass percentile threshold th=5 in Eq. 5. We adopt the few-step LDM SDXL-Turbo as the T2I backbone to maximize inference speed, with the total number of denoising steps set to T=4 and the CFG guidance strength to ω=1.5. As shown in Algorithm 1, a K-step inner loop is applied at each denoising step for structure-semantic co-optimization, with K=5 in our implementation. The percentile parameter in Eq. 8 is set p=60. Though the per-step optimization introduces additional computational overhead, the few-step denoising of SDXL-Turbo still keeps the overall time cost remarkably low, achieving considerably faster inference speed than related approaches. Our method demonstrates that high-quality visual anagram synthesis is also attainable using latent-based T2I diffusion models, even the distilled few-step ones.



Figure 3: Visual anagram synthesis results of our S2CO-Anagram using different transformation functions $g(\cdot)$ with keywords in each text prompt highlighted in red. **Better viewed with zoom in**.

4 Experimental Results

4.1 Qualitative Evaluations

Qualitative results and comparisons. As showcased in Fig. 3, our method produces high-quality anagram images under different transformations such as vertical flipping, diagonal flipping, 90° rotation, 180° rotation. The generated results are semantically faithful to both P_1 and P_2 in their corresponding views. The visual concepts of the two views are fused in a harmonious and tactful manner, allowing each view to retain its structural integrity and semantic clarity without sacrificing those of the other. In Fig. 4, we qualitatively compare our method with related SOTA approaches in synthesizing visual anagrams. Among the compared methods, Diffusion Illusions [5] exhibits relatively weak semantic fidelity, where key concepts from text prompts are not sufficiently and clearly expressed, which we attribute to the instability, high-variance gradient estimates, and mode collapse inherent in the score distillation optimization. SyncTweedies [45] generates results that are plagued by non-trivial noise and artifacts. Its fusion of concepts appears overly rigid, compromising visual harmony across both views. This suggests that multi-view fusion in feature space yields less smooth results and is more prone to structural artifacts compared to noise-space fusion. Visual

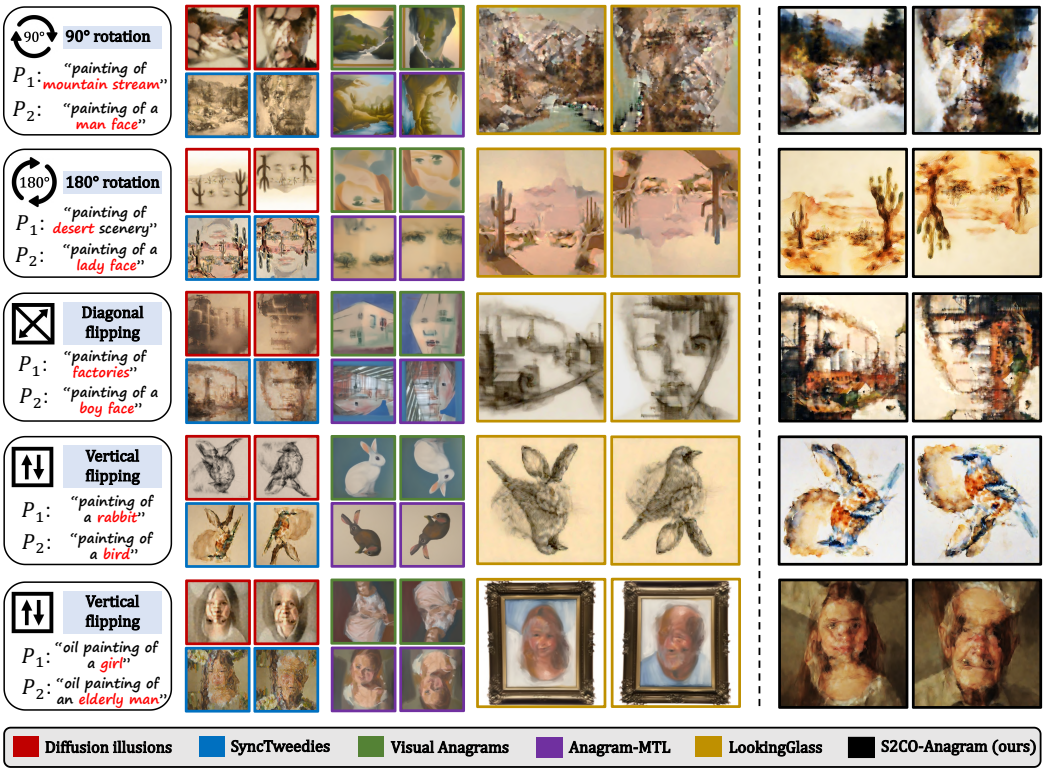


Figure 4: Qualitative method comparison. Results are presented at their actual resolution: thumbnails correspond to 256×256 outputs, and larger panels correspond to 512×512 results.

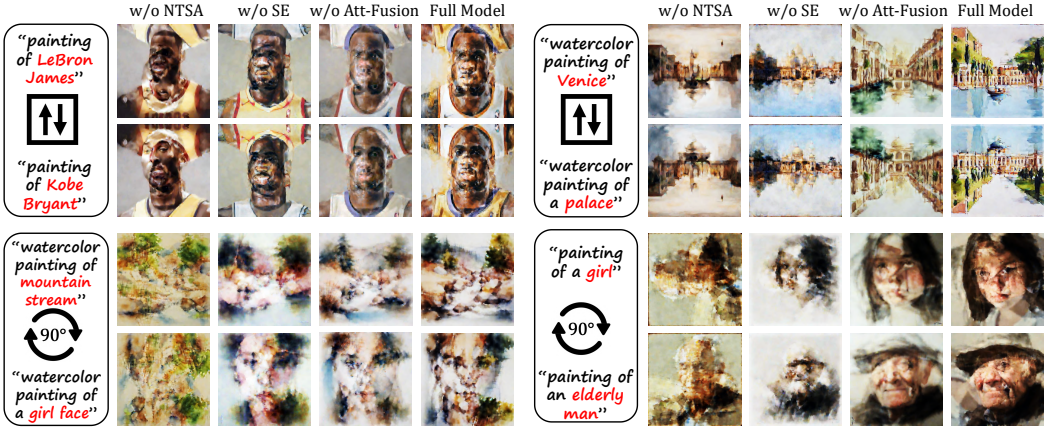


Figure 5: Qualitative ablation studies on key components of our method including null-text structure alignment (NTSA), semantic enhancement (SE), and attention-guided noise fusion (Att-Fusion).

Anagrams [31] and Anagram-MTL [78] produce results with limited semantic clarity, where key concepts from text prompts are not sufficiently discernible. Unlike the aforementioned methods based on image-space T2I model that are constrained to low-resolution outputs (256×256), LookingGlass [7] synthesizes higher-resolution anagram images by leveraging latent-based model. Nevertheless, its dependence on image-space warping operations introduces a lack of smoothness and naturalness into the synthesized results. In contrast, our S2CO-Anagram qualitatively outperforms all the above approaches in image resolution, conceptual fidelity, semantic clarity, and visual harmony.

Qualitative ablation study. To assess contribution of each core component in our method, a qualitative ablation study is performed and displayed in Fig. 5. Results w/o null-text structure



Figure 6: Diverse results with fixed text prompts (P_1, P_2) is enabled by varying the noise latent z_T.

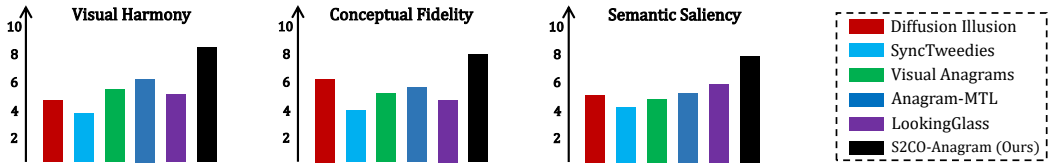


Figure 7: A user study comparing different methods for generating anagram images.

alignment (NTSA) exhibit severe structural artifacts, where the two views fail to fuse harmoniously. This is because the structural divergence along the multi-view denoising process pushes the fused latent toward a compromised structure which undermines semantic fidelity to both P_1 and P_2. By narrowing structural deviation via NTSA, the synchronous denoising converges toward a less conflicting direction, producing results that are visually more harmonious and conceptually more faithful. Compared with results w/o semantic enhancement (SE) optimization, the full model produces outputs that more clearly express the target semantics, demonstrating effectiveness of SE in promoting semantic fidelity. Results w/o attention-guided noise fusion (Att-Fusion) are relatively blurred and less detailed than those of the full model, demonstrating advantage of Att-Fusion over average-based noise fusion in preserving salient and fine-grained semantic features.

Generation diversity. As demonstrated in Fig. 6, our method enables random synthesis of diverse anagram images under fixed text prompts (P_1, P_2) by varying the initial Gaussian noise latent z_t. When z_t remains fixed, the deterministic nature of our sampling process yields a unique result.

4.2 Quantitative Analyses

Quantitative evaluations. We quantitatively evaluate methods on CIFAR10 2-views [78], a dataset with text prompt pairs using 10 classes of objects from CIFAR-10 [47]. For each pair of text prompts, we generate 10 images with different seeds and report the average results. We use CLIP [66] to measure visual-text semantic consistency (text fidelity) in the following three dimensions: (1) **Worst Alignment Score** ($\mathcal{A}_{min}$): the worst image-text CLIP score across two views; (2) **Concealment Score** ($\mathcal{C}$): measuring how well the generated anagram image from one view conceals textual concepts from the other view; (3) **Average Alignment Score** ($\mathcal{A}_{avg}$): the average image-text CLIP score across two views. The mathematical definitions of these metrics are defined and clarified in the **Appendix**. Besides, we measure visual quality of the generated results by evaluating the average **Aesthetic Score** ($\mathcal{AS}$) via the pretrained LAION Aesthetics Predictor V2 [70]. Finally, we evaluate model inference time which are averaged across 5 samples on a single RTX PRO 6000 GPU. Among the compared methods, our S2CO-Anagram delivers superior image resolution with significantly lower inference time, and consistently achieves the best performance in text fidelity metrics ($\mathcal{A}_{min}$, C, $\mathcal{A}_{avg}$) and visual aesthetic metric $\mathcal{AS}$, indicating that our method generates the most visually appealing anagram images with the clearest semantic discrimination and the highest computational efficiency.

Table 1: Quantitative evaluations across different methods in visual anagram synthesis.

Method	$\mathcal{A}_{min}\uparrow$	$\mathcal{C}\uparrow$	$\mathcal{A}_{avg}\uparrow$	$\mathcal{AS}\uparrow$	Resolution ↑	Inference Time ↓
Diffusion Illusions	0.2547	0.6867	0.2676	4.46	256×256	1027.3s
SyncTweedies	0.2568	0.6896	0.2692	4.28	256×256	5.9s
Visual Anagrams	0.2583	0.6744	0.2717	5.37	256×256	4.0s
Anagram-MTL	0.2711	0.6913	0.2816	5.65	256×256	16.2s
LookingGlass	0.2574	0.6848	0.2703	4.96	512×512	74.8s
S2CO-Anagram	**0.2726**	**0.6918**	**0.2824**	**6.14**	**512×512**	**2.6s**

Table 2: Quantitative ablation study on the contributions of the key model components.

NTSA	**SE**	**Att-Fusion**	$\mathcal{A}_{min}\uparrow$	$\mathcal{C}\uparrow$	$\mathcal{A}_{avg}\uparrow$	$\mathcal{AS}\uparrow$
×	×	×	0.2463	0.6672	0.2658	4.64
✓	×	×	0.2703	0.6894	0.2805	5.58
×	✓	×	0.2577	0.6745	0.2738	5.12
×	×	✓	0.2482	0.6690	0.2677	4.83
×	✓	✓	0.2603	0.6786	0.2764	5.29
✓	×	✓	0.2710	0.6903	0.2813	5.84
✓	✓	×	0.2719	0.6914	0.2819	6.06
✓	✓	✓	**0.2726**	**0.6918**	**0.2824**	**6.14**

Table 3: Ablation study on the DCT low-pass filtering percentile threshold th.

DCT_{th}	$\mathcal{A}_{min}\uparrow$	$\mathcal{C}\uparrow$	$\mathcal{A}_{avg}\uparrow$
DCT_0	0.2603	0.6786	0.2764
DCT_1	0.2655	0.6839	0.2785
DCT_2	0.2684	0.6887	0.2797
DCT_3	0.2710	0.6906	0.2811
DCT_4	0.2722	0.6916	0.2821
DCT_5	**0.2726**	**0.6918**	**0.2824**
DCT_{10}	0.2699	0.6895	0.2803
DCT_{20}	0.2616	0.6793	0.2772

Quantitative ablation study. Quantitative ablation study on core components of our method is reported in Tab. 2. Results show that all key contributions of our method synergetically improve the performance, which is basically in line with the qualitative results shown in Fig. 5. Among the proposed contributions, NTSA plays the most crucial role in promoting semantic fidelity of the generated anagram images. Ablation study on the DCT low-pass filtering percentile threshold th in NTSA is reported in Tab. 3. The absence of NTSA (th=0) yields the worst metrics, increasing th leads to strengthened structure alignment, thereby improving performance by alleviating structural artifacts. However, increasing th beyond an optimal point imposes excessive constraint on cross-view structure consistency, which is detrimental to the precise semantic expressiveness of the target concepts, resulting in degraded performance. We set th=5 based on empirical parameter tuning.

User study. We conduct a user study for subjective evaluation of the compared methods. Over 40 participants were recruited to rate the generated anagram images of our model and related methods on a 1–10 scale, with evaluation details provided in the Appendix. Fig. 7 presents the average user ratings for all compared methods. Our S2CO-Anagram surpasses other approaches by a large margin across all the evaluated dimensions, demonstrating significant advantages of our method for synthesizing visually harmonious, aesthetic appealing, and semantically faithful visual anagrams.

5 Conclusion

This paper presents S2CO-Anagram, a latent diffusion model (LDM) for fast and high-quality visual anagram synthesis. Tailored for the structural divergence and semantic dilution issues of the multi-view parallel denoising, we propose a structure-semantic co-optimization (S2CO) framework comprising null-text structure alignment, semantic enhancement, and attention-guided noise fusion. By integrating S2CO with pretrained few-step LDM, our method is able to generate higher-resolution anagram images with noticeably superior visual harmony, conceptual fidelity, semantic clarity, and aesthetic quality than related approaches, all while achieving significantly faster inference speed. However, our method still suffers from two limitations: it struggles to generate realistic-style anagram images despite producing visually appealing artistic visual anagrams, and it fails to yield satisfying results for more complex transformation functions such as jigsaw rearrangement. We will explore solutions to these two limitations in future work.

6 Acknowledgements

This work was supported by the National Natural Science Foundation of China [Grant No. 6260076518], and was also supported by the Major Project of the National Social Science Fund of China (Art Studies) [Grant No. 26ZD12].

References

[1] Stability AI. Adversarial diffusion distillation. `https://stability.ai/research/adversarial-diffusion-distillation`, 2026.

[2] Tilemachos Aravanis, Panagiotis Filntisis, Petros Maragos, and George Retsinas. Only-style: Stylistic consistency in image generation without content leakage. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6941–6951, 2025.

[3] Gantugs Atarsaikhan, Brian Kenji Iwana, and Seiichi Uchida. Contained neural style transfer for decorated logo generation. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 317–322. IEEE, 2018.

[4] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021.

[5] Ryan Burgert, Xiang Li, Abe Leite, Kanchana Ranasinghe, and Michael Ryoo. Diffusion illusions: Hiding images in plain sight. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[6] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22560–22570, 2023.

[7] Pascal Chang, Sergio Sancho, Jingwei Tang, Markus Gross, and Vinicius Azevedo. Lookingglass: Generative anamorphoses via laplacian pyramid warping. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24–33, 2025.

[8] Haibo Chen, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, Dongming Lu, et al. Artistic style transfer with internal-external learning and contrastive learning. *Advances in Neural Information Processing Systems*, 34:26561–26573, 2021.

[9] Ming-Te Chi, Chih-Yuan Yao, Eugene Zhang, and Tong-Yee Lee. Optical illusion shape texturing using repeated asymmetric patterns. *The Visual Computer*, 30(6):809–819, 2014.

[10] Wonwoong Cho, Yan-Ying Chen, Matthew Klenk, David I Inouye, and Yanxia Zhang. Attadapter: A robust and precise domain-specific multi-attributes t2i diffusion adapter via conditional variational autoencoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15626–15635, 2025.

[11] Hung-Kuo Chu, Wei-Hsin Hsu, Niloy J Mitra, Daniel Cohen-Or, Tien-Tsin Wong, and Tong-Yee Lee. Camouflage images. *ACM Trans. Graph.*, 29(4):51–1, 2010.

[12] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022.

[13] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8795–8805, 2024.

[14] Ciprian Corneanu, Raghudeep Gadde, and Aleix M Martinez. Latentpaint: Image inpainting in latent space with diffusion models. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 4334–4343, 2024.

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[16] Werner Ehm. A variational approach to geometric-optical illusions modeling. *Proceedings of Fechner Day*, 27(1):41–46, 2011.

[17] Masud An Nur Islam Fahim, Nazmus Saqib, and Jani Boutellier. Stam: Zero-shot style transfer using diffusion model via attention modulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6333–6343, 2025.

[18] Zhiyuan Fang, Rengan Xie, Xuancheng Jin, Qi Ye, Wei Chen, Wenting Zheng, Rui Wang, and Yuchi Huo. A 3 gs: Arbitrary artistic style into arbitrary 3d gaussian splatting. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17751–17760. IEEE, 2025.

[19] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9935–9946, 2023.

[20] William T Freeman, Edward H Adelson, and David J Heeger. Motion without movement. *ACM Siggraph Computer Graphics*, 25(4):27–30, 1991.

[21] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[22] Xiang Gao and Yunpeng Jia. Fbsdiff++: Improved frequency band substitution of diffusion features for efficient and highly controllable text-driven image-to-image translation. *arXiv preprint arXiv:2601.19115*, 2026.

[23] Xiang Gao and Jiaying Liu. Fbsdiff: Plug-and-play frequency band substitution of diffusion features for highly controllable text-driven image translation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4101–4109, 2024.

[24] Xiang Gao and Yuqi Zhang. Sragan: Saliency regularized and attended generative adversarial network for chinese ink-wash painting style transfer. *Pattern Recognition*, 162:111344, 2025.

[25] Xiang Gao, Yingjie Tian, and Zhiquan Qi. Rpd-gan: Learning to draw realistic paintings with generative adversarial network. *IEEE Transactions on Image Processing*, 29:8706–8720, 2020.

[26] Xiang Gao, Yuqi Zhang, and Yingjie Tian. Learning to incorporate texture saliency adaptive attention to image cartoonization. In *International Conference on Machine Learning*, pages 7183–7207. PMLR, 2022.

[27] Xiang Gao, Zhengbo Xu, Junhan Zhao, and Jiaying Liu. Frequency-controlled diffusion model for versatile text-guided image-to-image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1824–1832, 2024.

[28] Xiang Gao, Shuai Yang, and Jiaying Liu. Ptdiffusion: Free lunch for generating optical illusion hidden pictures with phase-transferred diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18240–18249, 2025.

[29] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[30] Daniel Geng, Inbum Park, and Andrew Owens. Factorized diffusion: Perceptual illusions by noise decomposition. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024.

[31] Daniel Geng, Inbum Park, and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24154–24163, 2024.

[32] Alex Gomez-Villa, Adrián Martín, Javier Vazquez-Corral, Marcelo Bertalmío, and Jesús Malo. On the synthesis of visual illusions using deep generative models. *Journal of Vision*, 22(8):2–2, 2022.

[33] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[34] Rui Guo, Jasmine Collins, Oscar de Lima, and Andrew Owens. Ganmouflage: 3d object nondetection with texture fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4702–4712, 2023.

[35] 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*, pages 4775–4785, 2024.

[36] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

[37] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[38] 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*, pages 1501–1510, 2017.

[39] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[40] Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Hao Kang, and Xin Lu. Infiniteyou: Flexible photo recrafting while preserving your identity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10898–10907, 2025.

[41] Shuhui Jiang, Jun Li, and Yun Fu. Deep learning for fashion style generation. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4538–4550, 2021.

[42] Yuxin Jiang, Liming Jiang, Shuai Yang, and Chen Change Loy. Scenimefy: Learning to craft anime scene via semi-supervised image-to-image translation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7357–7367, 2023.

[43] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[44] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[45] Jaihoon Kim, Juil Koo, Kyeongmin Yeo, and Minhyuk Sung. Synctweedies: A general generative framework based on synchronized diffusions. *arXiv preprint arXiv:2403.14370*, 2024.

[46] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10051–10060, 2019.

[47] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009.

[48] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017.

[49] Zhexin Liang, Zhaochen Li, Shangchen Zhou, Chongyi Li, and Chen Change Loy. Control color: Multimodal diffusion-based interactive image colorization: Z. liang et al. *International Journal of Computer Vision*, 133(11):7897–7923, 2025.

[50] Jia-Wei Liao, Winston Wang, Tzu-Sian Wang, Li-Xuan Peng, Cheng-Fu Chou, and Jun-Cheng Chen. Diffusion-based aesthetic qr code generation via scanning-robust perceptual guidance. *arXiv preprint arXiv:2403.15878*, 2024.

[51] Jia-Wei Liao, Winston Wang, Tzu-Sian Wang, Li-Xuan Peng, Ju-Hsuan Weng, Cheng-Fu Chou, and Jun-Cheng Chen. Diffqrcoder: Diffusion-based aesthetic qr code generation with scanning robustness guided iterative refinement. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5916–5925. IEEE, 2025.

[52] Tianwei Lin, Zhuoqi Ma, Fu Li, Dongliang He, Xin Li, Errui Ding, Nannan Wang, Jie Li, and Xinbo Gao. Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5141–5150, 2021.

[53] Wang Lin, Jingyuan Chen, Jiaxin Shi, Yichen Zhu, Chen Liang, Junzhong Miao, Tao Jin, Zhou Zhao, Fei Wu, Shuicheng Yan, et al. Non-confusing generation of customized concepts in diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 29935–29948, 2024.

[54] Haipeng Liu, Yang Wang, Biao Qian, Meng Wang, and Yong Rui. Structure matters: Tackling the semantic discrepancy in diffusion models for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8038–8047, 2024.

[55] Kaiyue Liu, Lei Wu, Mingzhe Yu, Xiaole Liu, Yajie Xu, and Xiangxu Meng. Arttypo: Multi-level controlled artistic typography with iterative feedback. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2025.

[56] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6649–6658, 2021.

[57] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

[58] Yiqun Mei, Jiacong Xu, and Vishal M Patel. Regs: Reference-based controllable scene stylization with gaussian splatting. *Advances in Neural Information Processing Systems*, 37: 4035–4061, 2024.

[59] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023.

[60] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4296–4304, 2024.

[61] Trong-Tung Nguyen, Quang Nguyen, Khoi Nguyen, Anh Tran, and Cuong Pham. Swiftedit: Lightning fast text-guided image editing via one-step diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21492–21501, 2025.

[62] Aude Oliva, Antonio Torralba, and Philippe G Schyns. Hybrid images. *ACM Transactions on Graphics (TOG)*, 25(3):527–532, 2006.

[63] Andrew Owens, Connelly Barnes, Alex Flint, Hanumant Singh, and William Freeman. Camouflaging an object from many viewpoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2782–2789, 2014.

[64] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5880–5888, 2019.

[65] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[67] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[68] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.

[69] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024.

[70] Christoph Schuhmann. Improved aesthetic predictor. GitHub repository. URL `https://github.com/christophschuhmann/improved-aesthetic-predictor`.

[71] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*.

[72] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[73] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1921–1930, 2023.

[74] Oguzhan Ulucan, Diclehan Ulucan, and Marc Ebner. A computational model for color assimilation illusions and color constancy. In *Proceedings of the Asian Conference on Computer Vision*, pages 630–647, 2024.

[75] Zhendong Wang, Jianmin Bao, Shuyang Gu, Dong Chen, Wengang Zhou, and Houqiang Li. Designdiffusion: High-quality text-to-design image generation with diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20906–20915, 2025.

[76] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13095–13105, 2023.

[77] Yangchen Xie, Xinyuan Chen, Li Sun, and Yue Lu. Dg-font: Deformable generative networks for unsupervised font generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5130–5140, 2021.

[78] Zhiyuan Xu, Yinhe Chen, Huan-ang Gao, Weiyan Zhao, Guiyu Zhang, and Hao Zhao. Diffusion-based visual anagram as multi-task learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 919–928. IEEE, 2025.

[79] Shuai Yang, Zhangyang Wang, Zhaowen Wang, Ning Xu, Jiaying Liu, and Zongming Guo. Controllable artistic text style transfer via shape-matching gan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4442–4451, 2019.

[80] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10743–10752, 2019.

[81] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23174–23184, 2023.

[82] Nir Zabari, Aharon Azulay, Alexey Gorkor, Tavi Halperin, and Ohad Fried. Diffusing colors: Image colorization with text guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023.

[83] Dingxi Zhang, Yu-Jie Yuan, Zhuoxun Chen, Fang-Lue Zhang, Zhenliang He, Shiguang Shan, and Lin Gao. Stylizedgs: Controllable stylization for 3d gaussian splatting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[84] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[85] Qing Zhang, Gelin Yin, Yongwei Nie, and Wei-Shi Zheng. Deep camouflage images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12845–12852, 2020.

[86] Yanbing Zhang, Mengping Yang, Qin Zhou, and Zhe Wang. Attention calibration for disentangled text-to-image personalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4764–4774, 2024.

[87] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–8, 2022.

[88] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10146–10156, 2023.

[89] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. Enabling instructional image editing with in-context generation in large scale diffusion transformer. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[90] Yang Zhou, Xu Gao, Zichong Chen, and Hui Huang. Attention distillation: A unified approach to visual characteristics transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18270–18280, 2025.

[91] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[92] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1219–1229, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's scope and contributions, including the problem addressed, proposed method, and superiority of our approach. These claims are consistent with the subsequent sections (proposed method, experimental results, and conclusion).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations of our work in Appendix F.1.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not involve theoretical proof, but include experimental proof of the structural divergence assumption in Appendix D.3.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: This paper has included algorithm details and implementation details in the main text and the Appendix. Our code will be publically available.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Code and data are not provided in this submission, but will be publically available and accessible soon.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: This paper provides comprehensive implementation details, including the dataset, hyperparameters, algorithm procedure, etc.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: This paper does not report error bars, owing to the substantial computational overhead and time cost involved.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our manuscirpt provides detailed information on the computer resources (in subsection 3.5) and time of execution (in subsection 4.2).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper does not involve human subjects, uses publicly available datasets, and presents foundational research with no direct path to harmful applications.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This paper has discussed both potential positive societal impacts and negative societal impacts of our work in Appendix F.2.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We believe that this paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All models and data used in this paper are properly cited. The code implemented for this work is original and written from scratch.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: New assets (code and model) are not yet released with this submission. They will be open-sourced later, at which point full documentation will be provided in the repository.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

A Preliminary Backgrounds

A.1 Diffusion Model Background

Since the advent of the Denoising Diffusion Probabilistic Model (DDPM) [37], diffusion models have soon dominated the field of generative AI, owing to their advantages in training stability and sampling diversity compared to GANs. Grounded in the theory of stochastic differential equations, a diffusion model learns to iteratively denoise a noise-corrupted input signal (e.g., an image or video clip), ultimately generating clean data that follow the target distribution. Conceptually, the model comprises a forward diffusion process and a reverse denoising process. The forward process gradually adds noise to the data over a series of steps, transforming the data into a standard Gaussian distribution. The reverse process learns to invert the forward process by iteratively removing noise, starting from pure noise and progressively reconstructing the original data. The model is trained to predict the noise added at each step of the forward process. Through this denoising objective, the model can generate new data samples by initializing from random noise and applying the reverse process.

Given the original data distribution $q(x_0)$, the forward diffusion process applies a T-step Markov chain to gradually add noise to the original data x_0 according to the conditional distribution $q(x_t|x_{t-1})$, which is defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathcal{I}), \tag{10}$$

where α_t follows a predefined schedule, $\alpha_t \in (0,1)$, $\alpha_t > \alpha_{t+1}$. Using the notation $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, we can derive the marginal distribution $q(x_t|x_0)$ that can be used to directly obtain x_t from x_0 in a single step for arbitrary time step t:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathcal{I}), \tag{11}$$

where $\sqrt{\bar{\alpha}_T} \approx 0$. With the forward diffusion process, the source data x_0 is transformed into x_T that follows an isotropic Gaussian distribution.

The reverse denoising process learns to conversely convert a Gaussian noise x_T to the manifold of the original data distribution $q(x_0)$ by gradually estimating and sampling from the posterior distribution $p(x_{t-1}|x_t)$. Since the posterior distribution $p(x_{t-1}|x_t)$ is mathematically intractable, we can derive the conditional posterior distribution $p(x_{t-1}|x_t, x_0)$ with the Bayes formula and some algebraic manipulation:

$$p(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t\mathcal{I}), \tag{12}$$

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha}_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t, \tag{13}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t, \tag{14}$$

where $\beta_t = 1 - \alpha_t$. However, the conditional posterior distribution $p(x_{t-1}|x_t, x_0)$ cannot be directly used for sampling since x_0 is unavailable at inference time (x_0 is the target of the sampling process). Therefore, DDPM tries to estimate the unknown x_0 given the x_t at each time step. Considering the reparameterization form of Eq. 11:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \tag{15}$$

where ϵ_t denotes the randomly sampled Gaussian noise that maps x_0 to x_t in a single step according to Eq. 11. Given Eq. 15, we can represent x_0 using x_t and ϵ_t:

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t). \tag{16}$$

However, the Gaussian noise ϵ_t sampled in the forward diffusion process is also unknown for the reverse denoising process where only x_t is available. Consequently, DDPM builds a noise estimation network ϵ_θ that predicts the sampled Gaussian noise ϵ_t in Eq. 16 with x_t and time step t as input, which is realized by training ϵ_θ with the following noise regression loss:

$$L = \|\epsilon_t - \epsilon_\theta(x_t, t)\|_2, \tag{17}$$

where $t \sim \text{Uniform}(\{1, ..., T\})$, $\epsilon_t \sim \mathcal{N}(0, \mathcal{I})$, x_t is computed via Eq. 15. After model training, $y_\theta(x_t)$, the estimation of x_0 given x_t, can be obtained simply by replacing ϵ_t in Eq. 16 with the predicted noise $\epsilon_\theta(x_t, t)$:

$$y_\theta(x_t) = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t, t)). \tag{18}$$

Replacing the unknown x_0 in Eq. 12 with its predicted estimation $y_\theta(x_t)$ given by Eq. 18, we can sample x_{t-1} based on x_t from the approximate posterior distribution $\mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, y_\theta(x_t)), \tilde{\beta}_t\mathcal{I})$, and thus sample the ultimate x_0 step by step from the initial Gaussian noise x_T.

A.2 Conditional Diffusion Model

Taking the image generation task as an example, conditional diffusion model tackles conditional image synthesis by introducing additional condition c to the model to guide image generation (denoising) process. In this paradigm, the condition signal c together with x_t and time step t are taken as input to the noise estimation network ϵ_θ, such that ϵ_θ is trained to conditionally predict the added Gaussian noise in the forward diffusion process, as supervised by the randomly sampled ϵ_t in Eq. 15. The training loss given by Eq. 17 is correspondingly updated as:

$$L = \|\epsilon_t - \epsilon_\theta(x_t, t, c)\|_2, \tag{19}$$

where $t \sim \text{Uniform}(\{1, ..., T\})$, $\epsilon_t \sim \mathcal{N}(0, \mathcal{I})$, x_t is computed via Eq. 15. After model training, the reverse sampling process is applied to generate new images from random Gaussian noise x_T, based on the step-by-step denoising according to the conditional posterior distribution given by Eq. 12, in which the unknown x_0 is approximated by the linear combination of x_t and the conditional noise estimation, *i.e.*, the $y_\theta(x_t)$ (approximate x_0 estimated by x_t) in Eq. 18 is updated as:

$$y_\theta(x_t, c) = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t, t, c)). \tag{20}$$

For text-to-image (T2I) diffusion model, the condition signal c is the textual embedding encoded from the input text prompt.

A.3 Latent Diffusion Model

Latent diffusion model (LDM) [67] compresses images from high-dimensional pixel space into low-dimensional feature space via pretrained autoencoder, and builds diffusion model based on latent feature space such that computational overhead for both training and inference can be dramatically lowered. The training of LDM is similar to Eq. 19 except that we use notation z to denote latent features:

$$L = \|\epsilon_t - \epsilon_\theta(z_t, t, c)\|_2, \tag{21}$$

where $\epsilon_t \sim \mathcal{N}(0, \mathcal{I})$, $z_t = \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_t$, $z_0 = E(x_0)$, x_0 is the noise-free clean image, E is the pretrained image encoder. The reverse denoising process from $z_T \sim \mathcal{N}(0, \mathcal{I})$ to z_0 is the same as $x_T \sim \mathcal{N}(0, \mathcal{I})$ to x_0 in DDPM. After reverse denoising process, the denoised clean features z_0 is decoded by the pretrained decoder D to yield the finally generated image x_0, *i.e.*, $x_0 = D(z_0)$. In LDM framework, the condition signal c could be the extracted image features that are concatenated with z_t as the input of ϵ_θ for image-to-image translation applications, and also could be the encoded textual features that are interacted with z_t via cross-attention layers inside ϵ_θ for text-to-image synthesis.

A.4 Adversarially Distilled Diffusion Model

Adversarial diffusion distillation (ADD) [69] reduces the inference steps of diffusion model required for high-quality image generation from over 50 steps to just 1–4 steps, enabling near real-time synthesis. The core components of ADD consist of score distillation from a pretrained teacher diffusion model and adversarial training with a discriminator network. The SDXL-Turbo, which we use in our proposed method, is a typical application of ADD that distills knowledge from the pretrained SDXL model. ADD integrates two complementary loss functions:

$$L = L_{adv}(\hat{x}_\theta(x_t, t), \phi) + \lambda L_{distill}(\hat{x}_\theta(x_t, t), \psi), \tag{22}$$

where θ denotes parameters of the ADD student model (initialized from pretrained SDXL), ψ denotes fixed parameters of the pretrained and frozen teacher diffusion model, ϕ denotes parameters of the discriminator network, λ is the balancing hyperparameter. The $\hat{x}_\theta(x_t, t)$ is the predicted clean image (or clean latent representation) produced by the ADD student model, parameterized by θ, given the noisy input x_t at a specific time step t:

$$\hat{x}_\theta(x_t, t) = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}, \tag{23}$$

where ϵ_θ is the noise estimation network of the ADD student model. The adversarial loss ensures that the generated images lie directly on the manifold of real images, mitigating the blurriness and artifacts common in distilled models. In adversarial training, the student diffusion model acts as the generator, the discriminator comprises a frozen pretrained feature backbone F (i.e., DINOv2 ViT-L) and a learnable lightweight discriminator heads $D_{\phi,k}$ attached to different layers k of the feature network F. Optional conditions such as text embeddings c_{text} and image embeddings c_{img} can be injected into both the generator and the discriminator.

Using the Hinge Loss objective, the generator adversarial loss is formulated as:

$$L^G_{adv}(\hat{x}_\theta(x_t, t), \phi) = -\mathbb{E}_{t,\epsilon,x_0}[\sum_k D_{\phi,k}(F_k(\hat{x}_\theta(x_t, t)))]. \tag{24}$$

The corresponding discriminator loss is formulated as:

$$\begin{aligned} L^D_{adv}(\hat{x}_\theta(x_t, t), \phi) =& \mathbb{E}_{x_0}\left[\sum_k \max(0, 1 - D_{\phi,k}(F_k(x_0))) + \gamma R1(\phi)\right] \\ &+ \mathbb{E}_{\hat{x}_\theta}\left[\sum_k \max(0, 1 + D_{\phi,k}(F_k(\hat{x}_\theta)))\right], \end{aligned} \tag{25}$$

where the first term encourages positive outputs for real images x_0, the second term encourages negative outputs for generated images $\hat{x}_0$. $R1(\phi)$ denotes the $R1$ gradient penalty, which stabilizes training by penalizing the gradient norm of discriminator outputs with respect to real data:

$$R1(\phi) = \frac{1}{2}\mathbb{E}_{x_0}[||\nabla_{x_0} D_\phi(x_0)||^2]. \tag{26}$$

The distillation loss transfers knowledge from the teacher diffusion model to the student model. Given the image $\hat{x}_\theta$ generated by the ADD student model, ADD firstly perturbs the generated $\hat{x}_\theta$ through teacher's forward diffusion process, yielding a noisy image at a randomly sampled time step t:

$$\hat{x}_{\theta,t} = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \;\; \epsilon_t \sim \mathcal{N}(0, \mathcal{I}). \tag{27}$$

Then, the teacher model performs denoising prediction on the perturbed sample $\hat{x}_{\theta,t}$, and compute distance between student output and teacher reconstruction, which is formulated as:

$$L_{distill}(\hat{x}_\theta(x_t, t), \psi) = \mathbb{E}_{t,\epsilon_t}[c(t) \cdot d(\hat{x}_\theta, \hat{x}_\psi(sg(\hat{x}_{\theta,t}); t))], \tag{28}$$

where $sg(\cdot)$ denotes the stop-gradient operation (teacher gradients do not backpropagate to student), $d(\cdot, \cdot)$ is a distance metric such as MSE, $c(t)$ is a time-dependent weighting coefficient. The teacher's reconstruction is defined as:

$$\hat{x}_\psi(\hat{x}_{\theta,t}, t) = \frac{\hat{x}_{\theta,t} - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}_\psi(\hat{x}_{\theta,t}, t)}{\sqrt{\bar{\alpha}_t}}, \tag{29}$$

where $\hat{\epsilon}_\psi(\hat{x}_{\theta,t}, t)$ is the teacher's predicted noise.

A.5 Classifier-free Guidance

Classifier-free guidance (CFG) [36] is a technique used in conditional generative models, particularly diffusion probabilistic models, to control the trade-off between sample diversity and fidelity to the conditioning information (e.g., a text prompt or a class label). Unlike classifier guidance, which requires a separately trained classifier to steer the generation process, CFG achieves this control without any auxiliary model, relying solely on the original generative model's own estimates of

conditional and unconditional score functions. In CFG, the predicted noise ϵ_t at time step t interpolates between the text-conditioned and unconditional noise estimations:

$$\epsilon_t = \omega\epsilon_\theta(x_t, t, y) + (1-\omega)\epsilon_\theta(x_t, t, \emptyset), \tag{30}$$

where $\emptyset$ denotes the null text embedding which is encoded from the empty string "", ω is the guidance strength parameter controlling the intensity of CFG. For ω=0, the generation is purely unconditional, ignoring the condition entirely. For ω=1, the standard conditional model is recovered. For ω>1, the condition is artificially amplified, forcing the sample to adhere more strictly to the provided context.

A.6 DCT Background

The Discrete Cosine Transform (DCT) is a fundamental technique in signal processing and data compression. It belongs to a family of unitary transforms that express a finite sequence of data points as a sum of cosine functions oscillating at different frequencies. Unlike the Discrete Fourier Transform (DFT), the DCT operates purely on real numbers and avoids the discontinuities at block boundaries by implicitly assuming symmetric extension of the signal. In this method, the DCT is employed to extract low-frequency components, offering the advantage of quantitatively controlling the extent of low-frequency signal extraction by adjusting the filtering threshold. We use the DCT-II form where the one-dimensional DCT (1D-DCT) is formulated as:

$$S[k] = \sqrt{\frac{2}{N}} \cdot c[k] \cdot \sum_{n=0}^{N-1} x[n] \cdot \cos(\frac{\pi k(2n+1)}{2N}), \tag{31}$$

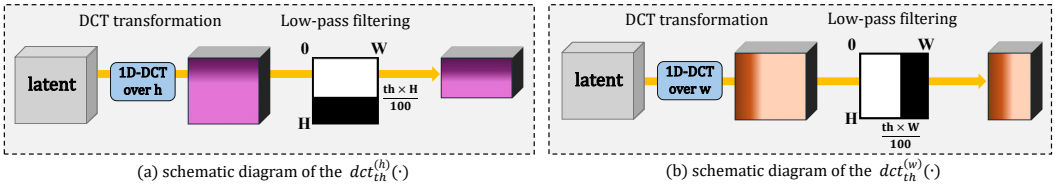
$$c[k] = \begin{cases} \frac{1}{\sqrt{2}} & k = 0 \\ 1 & k \neq 0 \end{cases} \tag{32}$$

where N denotes the sequence length, $k = 0, 1, \ldots, N-1$. x refers to the input sequence, S is the corresponding DCT spectrum.

B More Implementation Details

B.1 DCT Filtering in Null-text Structure Alignment

For the DCT filtering operations used in the VLC loss in our method, the input sequence x in Eq. 31 is either $z^{(1)}_{t\to 0}$, the predicted clean latent in the source view, or $g^{-1}(z^{(2)}_{t\to 0})$, the inverse transformation of the predicted clean latent in the transformed view. Let H, W respectively denote the height and width of the latent feature, then the sequence length N in Eq. 31 is either H or W depending on which spatial dimension to apply 1D-DCT on. In our method, $dct^{(h)}_{th}(\cdot)$ refers to the operation of applying 1D-DCT transformation over the h dimension of the latent feature followed by the low-pass filtering which removes the frequency components higher than the percentile threshold th in the h dimension. Similarly, $dct^{(w)}_{th}(\cdot)$ denotes the same DCT filtering operation with the 1D-DCT transformation applied over the w dimension of the latent feature.



(a) schematic diagram of the $dct^{(h)}_{th}(\cdot)$ (b) schematic diagram of the $dct^{(w)}_{th}(\cdot)$

Figure 8: Schematic diagram of $dct^{(h)}_{th}(\cdot)$ and $dct^{(w)}_{th}(\cdot)$ operations applied on the latent feature.

Fig. 8 illustrates the implementation details of the percentile-based low-pass DCT filtering operations, denoted as $dct^{(h)}_{th}(\cdot)$ and $dct^{(w)}_{th}(\cdot)$, which are employed in our proposed VLC loss. For $dct^{(h)}_{th}(\cdot)$, the 1D-DCT is firstly applied along the h dimension of the input latent feature. Subsequently, frequency

components exceeding $\frac{th \times H}{100}$ in the h dimension are filtered out, thereby retaining only low-frequency spectral components along the h dimension. Similarly, $dct_{th}^{(w)}(\cdot)$ applies the 1D-DCT along the w dimension of the input latent feature, then removes frequency components higher than $\frac{th \times W}{100}$ in the w dimension, and outputs the resulting low-frequency spectral components.

B.2 Attention Maps Extraction in Semantic Enhancement

In cross-modal semantic enhancement optimization, cross-attention maps are extracted and utilized to guide the optimization of input latent feature to enhance semantic expressiveness of the generated results. The computation of the cross-attention maps is reformulated below:

$$M_l^{(1)} = Softmax\left(Q_l^{(1)}(K_l^{(1)})^T/\sqrt{d}\right), \quad M_l^{(2)} = Softmax\left(Q_l^{(2)}(K_l^{(2)})^T/\sqrt{d}\right), \tag{33}$$

where l denotes layer index, $M_l^{(1)}$ and $M_l^{(2)}$ denote attention maps at the l-th cross-attention layer inside the denoising network ϵ_θ, given z_t and $g(z_t)$ as input, respectively, $Q_l^{(1)}$ and $Q_l^{(2)}$ are queries of the l-th cross-attention layer derived from z_t and $g(z_t)$, respectively, $K_l^{(1)}$ and $K_l^{(2)}$ are keys of the l-th cross-attention layer originated from the textual embedding y_1 and y_2, respectively, d denotes the feature dimension of the query and key at the current layer. $M_l^{(1)}$ comprises multiple attention maps with each attention map associated with a word embedding in y_1, and similarly for $M_l^{(2)}$ with y_2. We sum over the selected attention maps corresponding to the conceptually most crucial words in P_1 and P_2 (marked in red in Fig. 2), yielding $\tilde{M}_l^{(1)}$ and $\tilde{M}_l^{(2)}$. Let L denotes the total number of cross-attention layers inside ϵ_θ. By upscaling all the extracted $\{\tilde{M}_l^{(1)}\}_{l=1}^L$ to the same spatial size of latent features and summing them together, we obtain $\tilde{M}^{(1)}$. Analogously, $\tilde{M}^{(2)}$ is obtained by upscaling all the extracted $\{\tilde{M}_l^{(2)}\}_{l=1}^L$ to the latent feature resolution and summing them together. In $\tilde{M}^{(1)}$, pixel intensity reflects conceptual correlation strength of each spatial location to the key concept of P_1, and similarly for $\tilde{M}^{(2)}$ to P_2. The resulting attention maps $\tilde{M}^{(1)}$ and $\tilde{M}^{(2)}$ are utilized for cross-modal semantic enhancement optimization as formulated by Eq. 8.

C Evaluation Metrics and Criteria

C.1 Evaluation Metrics

This subsection provides a detailed explanation of the quantitative evaluation metrics used in the main text to assess model performance in generating visual anagram images, along with their corresponding mathematical definitions. Let P_1 and P_2 denote the two multi-view text prompts, $x^{(1)}$ and $x^{(2)}$ denote the two views of the generated anagram image that are respectively aligned with P_1 and P_2 in semantic concept. The evaluation metrics are defined as follows:

Worst Alignment Score ($\mathcal{A}_{min}$): the worst alignment (CLIP image-text cosine similarity) between (P_1, $x^{(1)}$) and (P_2, $x^{(2)}$), which is mathematically defined as:

$$\mathcal{A}_{min} = min[CLIP(P_1, x^{(1)}), CLIP(P_2, x^{(2)})], \tag{34}$$

$$CLIP(P, x) = \frac{E_{text}(P) \cdot E_{img}(x)}{||E_{text}(P)||\ ||E_{img}(x)||}, \tag{35}$$

where E_{text} and E_{img} are CLIP text encoder and CLIP image encoder, respectively.

Average Alignment Score ($\mathcal{A}_{avg}$): the average image-text semantic alignment score across the two views, which is mathematically defined as:

$$\mathcal{A}_{avg} = \frac{CLIP(P_1, x^{(1)}) + CLIP(P_2, x^{(2)})}{2}. \tag{36}$$

Concealment Score ($\mathcal{C}$): measuring how well the generated anagram image from one view conceals textual concepts from the other view. This metric evaluates model's performance in suppressing semantic ambiguity, i.e., the ability of the model in presenting semantic alignment of a view to its paired text prompt while mitigating semantic leakage from the other text prompt. The mathematical definition is provided as follows:

$$u_{11}, u_{12} = Softmax\left[CLIP(P_1, x^{(1)}), CLIP(P_1, x^{(2)})\right], \tag{37}$$

$$u_{21}, u_{22} = Softmax\left[CLIP(P_2, x^{(1)}), CLIP(P_2, x^{(2)})\right], \tag{38}$$

$$v_{11}, v_{21} = Softmax\left[CLIP(P_1, x^{(1)}), CLIP(P_2, x^{(1)})\right], \tag{39}$$

$$v_{12}, v_{22} = Softmax\left[CLIP(P_1, x^{(2)}), CLIP(P_2, x^{(2)})\right], \tag{40}$$

$$C = (u_{11} + u_{22} + v_{11} + v_{22})/4. \tag{41}$$

C.2 Evaluation Criteria in User Study

The user study shown in Fig. 7 is designed to subjectively evaluate methods in generating anagram images. Over 40 participants were recruited to evaluate the output visual anagrams of different methods using a 1–10 rating scale across the following three dimensions:

1. **Visual harmony**: the degree to which the multi-view contents are harmoniously blended together without noticeable structural artifacts and unnatural visual elements;
2. **Conceptual fidelity**: the faithfulness of each view to its corresponding text prompt in semantic concept, along with the ability to conceal concept from the other prompt;
3. **Semantic saliency**: the degree of clarity and distinctiveness of the semantic content in each view, where semantically blurry or ambiguous results are considered less preferable.

D Auxiliary Experiments

D.1 Ablation Study on Denoising Steps

Table 4: Ablation study on the denoising steps T on the CIFAR10 2-views dataset.

denoising steps T	$\mathcal{A}_{min} \uparrow$	$\mathcal{C} \uparrow$	$\mathcal{A}_{avg} \uparrow$	$\mathcal{AS} \uparrow$
1	0.2486	0.6702	0.2618	3.55
2	0.2579	0.6887	0.2704	4.73
3	0.2708	0.6915	0.2818	6.05
4	**0.2726**	**0.6918**	**0.2824**	**6.14**
5	0.2675	0.6876	0.2800	5.86
6	0.2603	0.6851	0.2715	5.44

Table 4 presents ablation study on the total number of the denoising steps T in our method. Results show that the textual fidelity and the aesthetic score of the generated anagram images improve with the increase of T, but drop rapidly when $T > 4$. This is due to that our model is built upon SDXL-Turbo, which achieves optimal generation quality within $1 \leq T \leq 4$. Extending T to $T > 4$ degrades visual quality of the generated images and thus leads to declined evaluation metrics. Within $1 \leq T \leq 4$, smaller value of T results in worse evaluation results, suggesting that the multi-view synchronous denoising is not well-suited for very few sampling steps (e.g., $T = 1$ or $T = 2$). This conclusion aligns with our observation that our method with $T = 1$ or $T = 2$ produces overly blurred anagram images which struggle to sufficiently reflect the target semantic concepts.

D.2 Ablation Study on Inner Loop Steps

We also perform ablation study on the inner loop steps K in our structure-semantic co-optimization (S2CO) framework. Results presented in Tab. 5 demonstrate that increasing the inner loop steps K (the number of S2CO optimization times at each denoising step) yields monotonic performance gains in multi-view textual fidelity, although the improvements become marginal for $K > 5$. Results also show that the overall inference time scales approximately linearly with the increase of K. Through a balanced consideration of the quantitative metrics and the time cost, we select $K = 5$ as the optimal trade-off between model performance and computational efficiency.

Table 5: Ablation study on the inner loop steps K on the CIFAR10 2-views dataset.

inner loop step K	$\mathcal{A}_{min} \uparrow$	$\mathcal{C} \uparrow$	$\mathcal{A}_{avg} \uparrow$	inference time
0	0.2482	0.6690	0.2677	0.4s
1	0.2544	0.6736	0.2706	0.9s
2	0.2617	0.6821	0.2739	1.4s
3	0.2668	0.6874	0.2773	1.8s
4	0.2707	0.6905	0.2809	2.2s
5	0.2726	0.6918	0.2824	2.6s
6	0.2733	0.6923	0.2827	3.1s
7	0.2738	0.6925	0.2829	3.6s
8	0.2740	0.6926	0.2829	4.1s

D.3 Quantitative Verification of Structural Divergence

We design the following experiment to quantitatively verify the structural divergence issue in the multi-view synchronous denoising process and how the null-text structure alignment (NTSA) mitigates this issue. Based on a standard T-step multi-view parallel denoising trajectory composed of $z_T \rightarrow \cdots z_t \rightarrow \cdots z_1 \rightarrow z_0$ (no null-text structure alignment and semantic enhancement optimizations involved), we apply attention-guided noise fusion at the first $T-1$ steps. For the last denoising step $z_1 \rightarrow z_0$, we do not perform multi-view noise fusion and instead directly predict the two image views $x_0^{(1)}$ and $x_0^{(2)}$ according to their respective noise estimations:

$$\epsilon_1^{(1)} = \omega\epsilon_\theta(z_1, t=1, y_1) + (1-\omega)\epsilon_\theta(z_1, t=1, y_\emptyset), \tag{42}$$

$$\epsilon_1^{(2)} = \omega\epsilon_\theta(g(z_1), t=1, y_2) + (1-\omega)\epsilon_\theta(g(z_1), t=1, y_\emptyset), \tag{43}$$

$$z_0^{(1)} = (z_1 - \sqrt{1-\bar{\alpha}_1}\epsilon_1^{(1)})/\sqrt{\bar{\alpha}_1}, \;\; z_0^{(2)} = (g(z_1) - \sqrt{1-\bar{\alpha}_1}\epsilon_1^{(2)})/\sqrt{\bar{\alpha}_1}, \tag{44}$$

$$x_0^{(1)} = D(z_0^{(1)}), \quad x_0^{(2)} = D(z_0^{(2)}), \tag{45}$$

where D is VAE decoder, y_1 and y_2 are textual embeddings encoded from the multi-view text prompts P_1 and P_2, respectively. Structural divergence occurs between $x_0^{(1)}$ and $g^{-1}(x_0^{(2)})$ due to different semantic guidance from y_1 and y_2. We use the **DINO-ViT Self-Similarity Distance** [12] as the metric to measure the structure distance between $x_0^{(1)}$ and $g^{-1}(x_0^{(2)})$, which we denote as **Structure Distance w/o NTSA**. The larger this metric is, the more structural divergence between the generated two images when transformed back to the original view.

To quantitatively verify the effectiveness of NTSA in narrowing structural divergence, we apply NTSA at the final denoising step $z_1 \rightarrow z_0$, with the generated two images recomputed from the optimized null-text embedding $\tilde{y}_\emptyset$, where $\tilde{y}_\emptyset$ is optimized with K steps under our proposed VLC loss. The generated two images with the last-step NTSA are formulated as:

$$\tilde{\epsilon}_1^{(1)} = \omega\epsilon_\theta(z_1, t=1, y_1) + (1-\omega)\epsilon_\theta(z_1, t=1, \tilde{y}_\emptyset), \tag{46}$$

$$\tilde{\epsilon}_1^{(2)} = \omega\epsilon_\theta(g(z_1), t=1, y_2) + (1-\omega)\epsilon_\theta(g(z_1), t=1, \tilde{y}_\emptyset), \tag{47}$$

$$\tilde{z}_0^{(1)} = (z_1 - \sqrt{1-\bar{\alpha}_1}\tilde{\epsilon}_1^{(1)})/\sqrt{\bar{\alpha}_1}, \;\; \tilde{z}_0^{(2)} = (g(z_1) - \sqrt{1-\bar{\alpha}_1}\tilde{\epsilon}_1^{(2)})/\sqrt{\bar{\alpha}_1}, \tag{48}$$

$$\tilde{x}_0^{(1)} = D(\tilde{z}_0^{(1)}), \quad \tilde{x}_0^{(2)} = D(\tilde{z}_0^{(2)}), \tag{49}$$

Similarly, we use DINO-ViT Self-Similarity Distance to measure the structure distance between $\tilde{x}_0^{(1)}$ and $g^{-1}(\tilde{x}_0^{(2)})$, which we denote as **Structure Distance w/ NTSA**.

We randomly sample 10 pairs of (P_1,P_2) in CIFAR10 2-views dataset, and evaluate the average Structure Distance w/o and w/ NTSA for different denoising steps T based on the above-mentioned experiment setting. Results evaluated with the vertical flipping transformation and the 90° rotation transformation are respectively reported in Tab. 6 and Tab. 7. The non-zero values of the Structure Distance metric indicate that structural divergence indeed exists along the denoising process, and that fewer denoising steps T is more prone to structural divergence issue. By comparing the Structure Distance w/o and w/ NTSA under different denoising steps, we see that NTSA consistently narrows the Structure Distance between the generated two images in the original view, which quantitatively verify the effectiveness of NTSA in mitigating structural divergence.

Table 6: Comparison of Structure Distance w/o and w/ NTSA evaluated at different denoising steps in the transformation of vertical flipping. Results are averaged over 10 samples in CIFAR10 2-views.

Metric evaluated in **vertical flipping transformation**	$T=1$	$T=2$	$T=3$	$T=4$
Structure Distance w/o NTSA	0.074	0.066	0.056	0.049
Structure Distance w/ NTSA	0.046	0.040	0.035	0.032

Table 7: Comparison of Structure Distance w/o and w/ NTSA evaluated at different denoising steps in the transformation of 90° rotation. Results are averaged over 10 samples in CIFAR10 2-views.

Metric evaluated in **90° rotation transformation**	$T=1$	$T=2$	$T=3$	$T=4$
Structure Distance w/o NTSA	0.077	0.068	0.057	0.050
Structure Distance w/ NTSA	0.049	0.042	0.036	0.034

D.4 Quantitative Investigation of the Percentile Parameter in Semantic Enhancement

In semantic enhancement optimization, the extracted attention maps $\tilde{M}^{(1)}$ and $\tilde{M}^{(2)}$ are leveraged to optimize the input latent feature at each denoising step, which we reformulate below:

$$z_t := z_t + \lambda_2 \nabla_{z_t} \left(\sum_{(i,j)\in\Omega^{(1)}} \tilde{M}^{(1)}_{i,j} + \sum_{(i,j)\in\Omega^{(2)}} \tilde{M}^{(2)}_{i,j} \right), \ \Omega^{(k)} = \left\{ (i,j) | \tilde{M}^{(k)}_{i,j} \geq \tau^{(k)}_p \right\}, \quad (50)$$

where i, j denote spatial indices, $\tau^{(k)}_p$ refers to the p-th percentile of the values in $\tilde{M}^{(k)}$. We only back-propagate through locations above the p-th percentile of each attention map to reinforce semantic expressiveness only in the most concept-relevant regions. We conduct ablation study over the percentile parameter p on a randomly selected subset of CIFAR10 2-views dataset which contains 20 (P_1, P_2) samples. Aesthetic Score ($\mathcal{AS}$) is evaluated over different values of p, with the average results reported in Tab. 8. Setting $p = 100$ is equivalent to the absence of semantic enhancement, which produces results with the worst aesthetic quality. Setting $p = 0$ corresponds to the maximum semantic enhancement, which over-amplifies attention activations and causes semantic information to diffuse into conceptually irrelevant image regions, thereby degrading the aesthetic score. Through hyper-parameter tuning, we quantitatively find that $p = 60$ yields the best aesthetic quality of the generated anagram images.

Table 8: Ablation study over the percentile parameter p in semantic enhancement optimization.

Metric	$p=0$	$p=20$	$p=40$	$p=60$	$p=80$	$p=100$
Aesthetic Score ($\mathcal{AS}$)	5.96	6.03	6.09	6.12	5.98	5.87

D.5 Evaluation on Other Form of Structure Alignment

Our focus is not latent-space DCT filtering itself, but rather aligning cross-view structure by optimizing null-text embedding to disentangle structure from semantics——a principle that generalizes beyond DCT; other frequency-domain filtering, e.g., wavelet, could also be applicable (Tab. 9):

D.6 Evaluation on an LLM-Augmented Prompt Dataset

To evaluate methods under more challenging semantic conditions, we constructed an additional multi-view text prompt dataset using large language model. We defined four types of paired text prompts templates: ("painting of A", "painting of B"), ("oil painting of A", "oil painting of B"), ("watercolor painting of A", "watercolor painting of B"), and ("artistic drawing of A", "artistic drawing of B"). ChatGPT was then employed to randomly complete the entity nouns A and B in the templates, resulting in a text prompt dataset consisting of 120 samples of (P_1, P_2) pairs, which we term the LLM-Augmented dataset.

Table 9: Comparison between different frequency-domain transformations for structure alignment.

Frequency-domain filter for NTSA	$\mathcal{A}_{min} \uparrow$	$\mathcal{C} \uparrow$	$\mathcal{A}_{avg} \uparrow$	$\mathcal{AS} \uparrow$	Inference Time $\downarrow$
DCT filter	0.2719	0.6913	0.2821	6.16	2.2s
Wavelet filter	0.2726	0.6918	0.2824	6.14	2.6s

Table 10: Quantitative evaluations across different methods based on the LLM-Augmented dataset.

Method	$\mathcal{A}_{min} \uparrow$	$\mathcal{C} \uparrow$	$\mathcal{A}_{avg} \uparrow$	$\mathcal{AS} \uparrow$	Resolution $\uparrow$	Inference Time $\downarrow$
Diffusion Illusions	0.2515	0.6845	0.2633	4.15	256×256	1027.3s
SyncTweedies	0.2538	0.6860	0.2664	4.09	256×256	5.9s
Visual Anagrams	0.2573	0.6778	0.2702	5.14	256×256	4.0s
Anagram-MTL	0.2661	0.6882	0.2775	5.36	256×256	16.2s
LookingGlass	0.2538	0.6833	0.2684	4.83	512×512	74.8s
S2CO-Anagram	**0.2707**	**0.6894**	**0.2796**	**5.92**	**512×512**	**2.6s**

Quantitative evaluations based on the curated LLM-Augmented dataset are displayed in Tab. 10. Results show that our S2CO-Anagram still achieves top performance in both text fidelity (quantified by $\mathcal{A}_{min}$, $\mathcal{C}$, and $\mathcal{A}_{avg}$) and visual quality (quantified by $\mathcal{AS}$), which is basically consistent with the conclusion obtained on the CIFAR10 2-views dataset as shown in Tab. 1.

E More Visual Results

To complement visual results of our method, we present additional anagram images generated by our S2CO-Anagram using various transformation functions, as shown in Fig. 9 through Fig. 17. Results fully demonstrate the generative diversity of our method, which produces diverse high-quality samples by randomly initializing the noise latent z_T at each runtime. In Fig. 18 and Fig. 19, we display more visual comparison of our method with Visual Anagrams [31] and Anagram-MTL [78], our results present better image resolution, aesthetic quality, visual harmony, conceptual fidelity, and semantic distinctiveness than the compared approaches.

F Limitations and Discussions

F.1 Limitations and Future Work

This paper proposes S2CO-Anagram for high-quality and fast visual anagram synthesis. Our method achieves superior aesthetic quality, text fidelity, semantic clarity, and image resolution than related SOTA approaches with substantially faster inference speed. However, our method still suffer from the following two limitations: (1) our method is capable of generating visually appealing anagram images with an artistic style, however, its performance in generating realistic-style anagram images is relatively poor; (2) we empirically observe that our method fails to generate satisfying results for more complex transformation functions such as jigsaw rearrangement. In future work, we will explore strategies to resolve these issues.

F.2 Social Impact Discussions

The proposed S2CO-Anagram framework for fast visual anagram synthesis offers several beneficial societal implications. First, it democratizes illusionary digital art creation by providing artists, designers, and content creators with an efficient tool that generates high-quality visual anagrams in 2 seconds, significantly lowering the barrier to creative expression. Second, educational applications may arise, enabling interactive learning tools that illustrate principles of visual perception, Gestalt psychology, and human creativity. Third, the framework could support therapeutic or cognitive training applications, such as visual rehabilitation exercises or interventions that engage pattern recognition and mental rotation skills. Fourth, this technology could inspire new forms of advertising, branding, and visual communication, in which dual-interpretation imagery captures audience attention and conveys messages in an engaging manner.

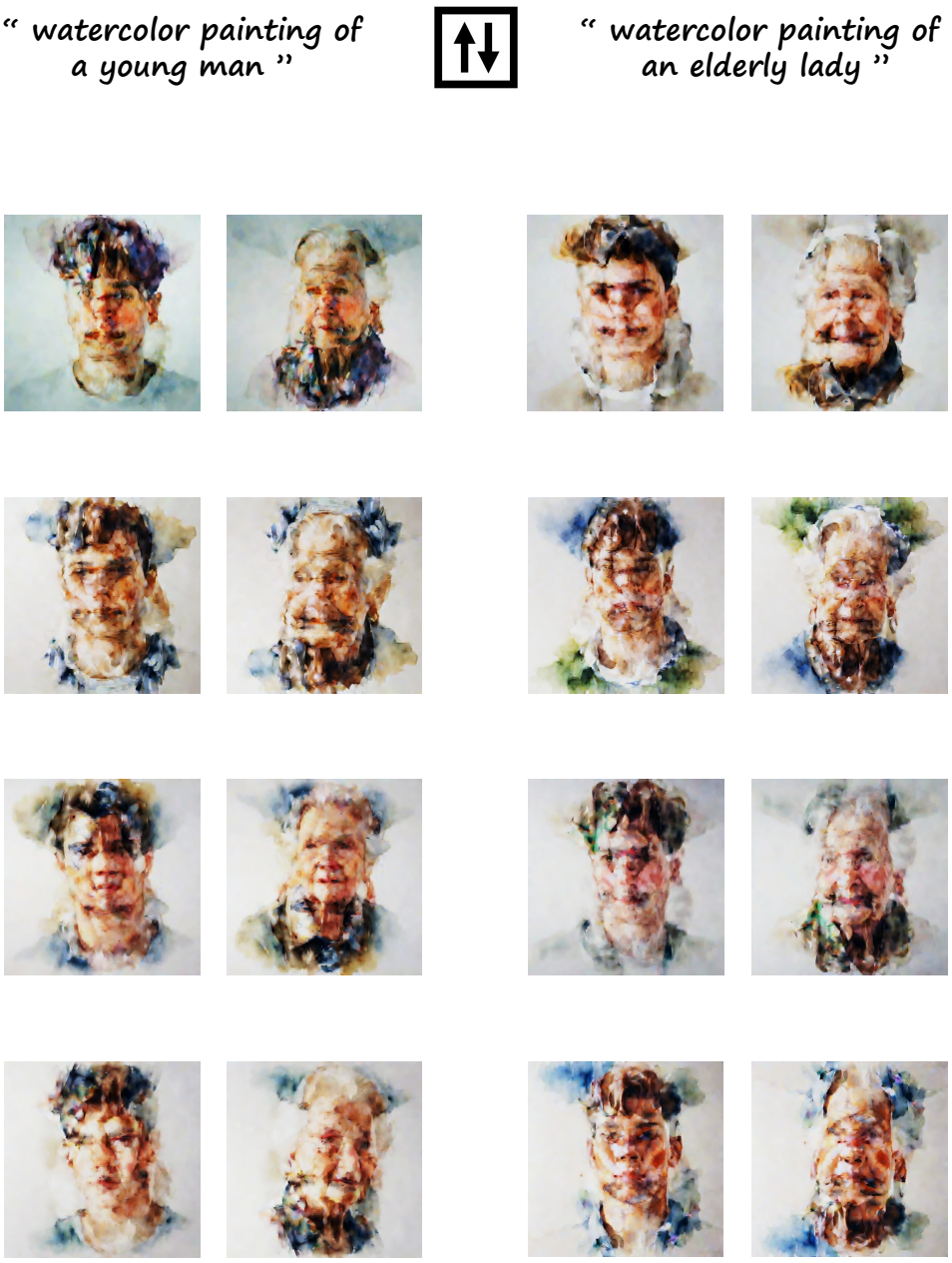


Figure 9: More diverse results-case 1.

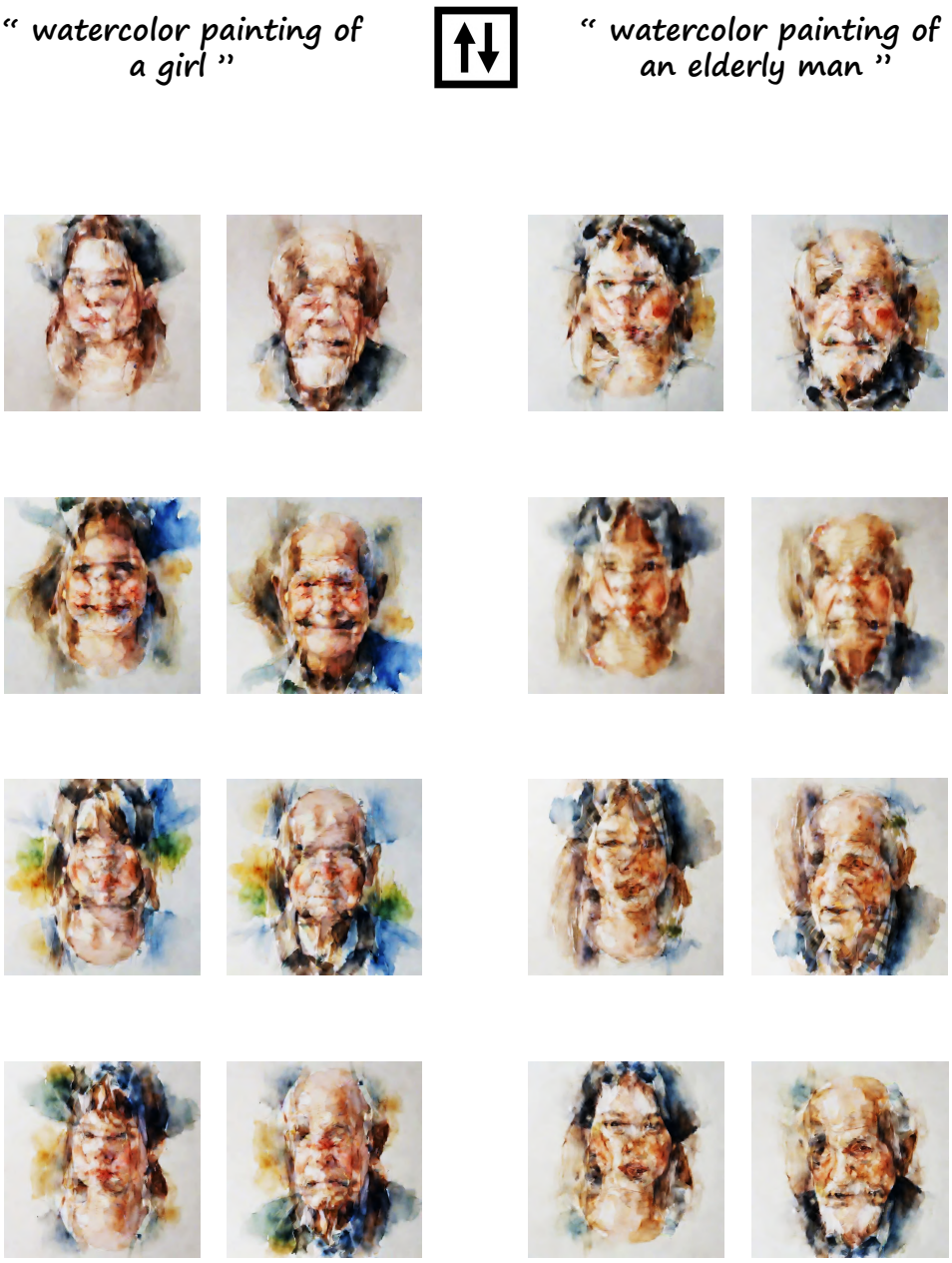


Figure 10: More diverse results-case 2.



Figure 11: More diversified results-case 3.

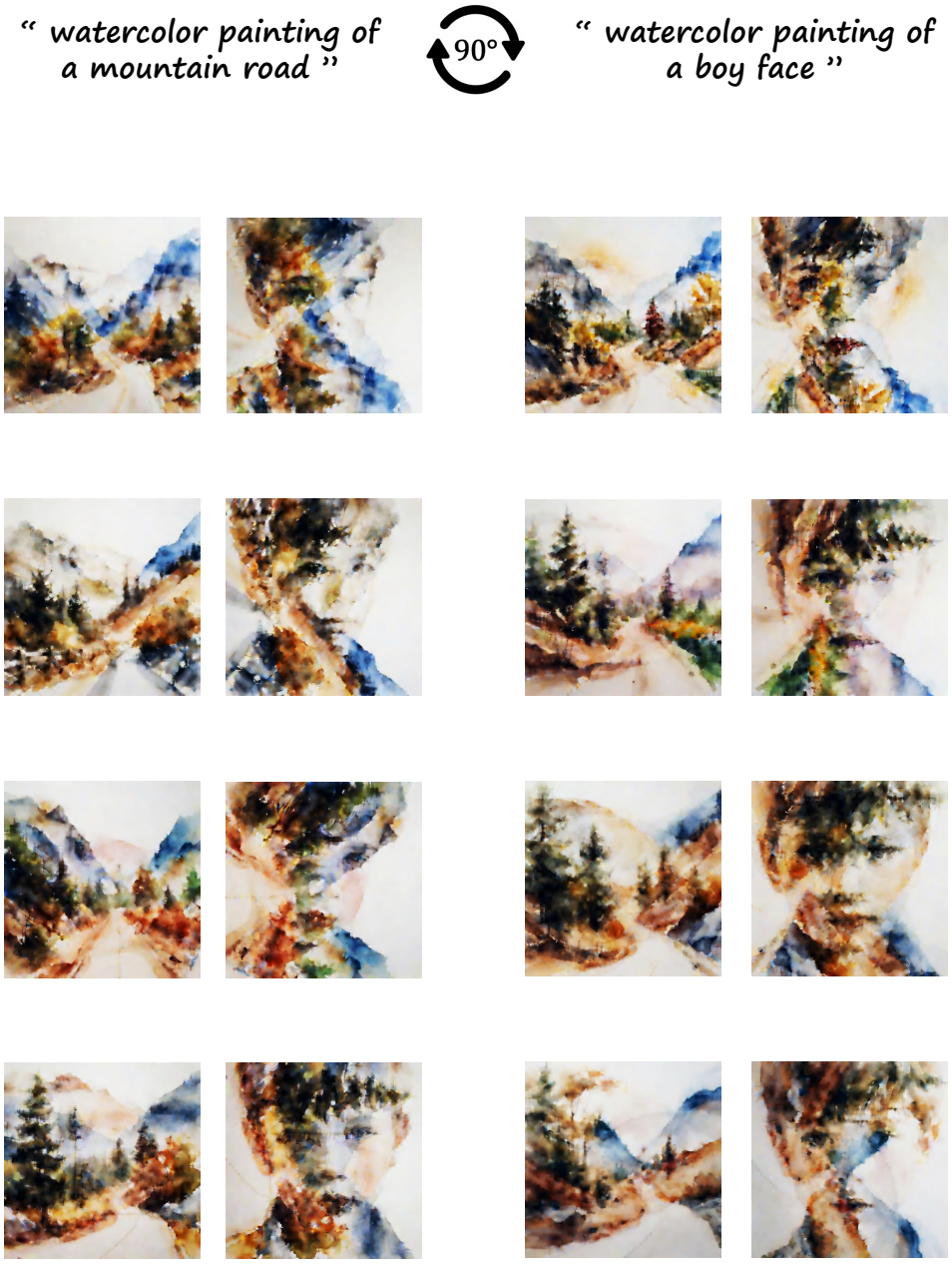


Figure 12: More diverse results-case 4.

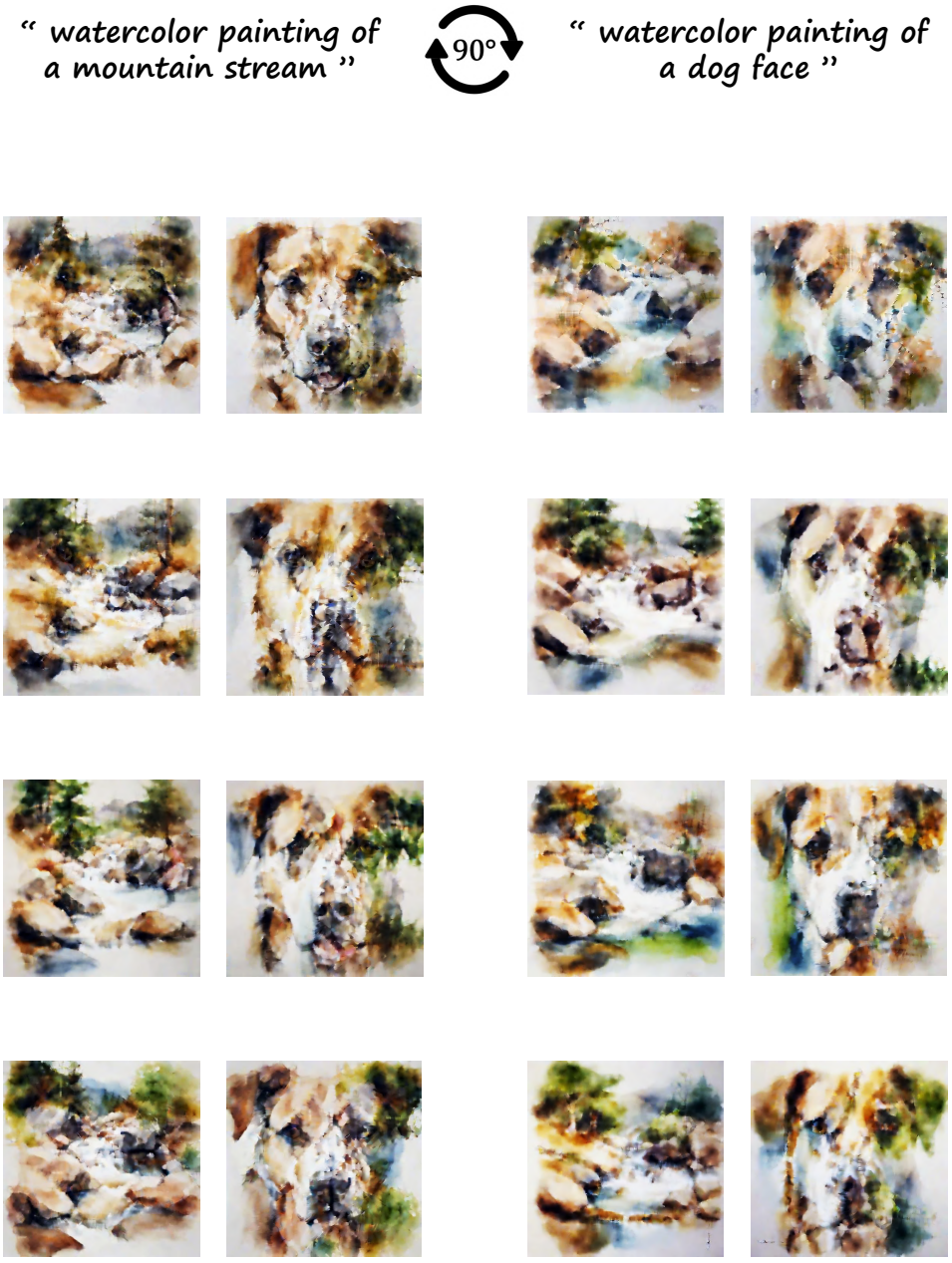


Figure 13: More diversified results-case 5.

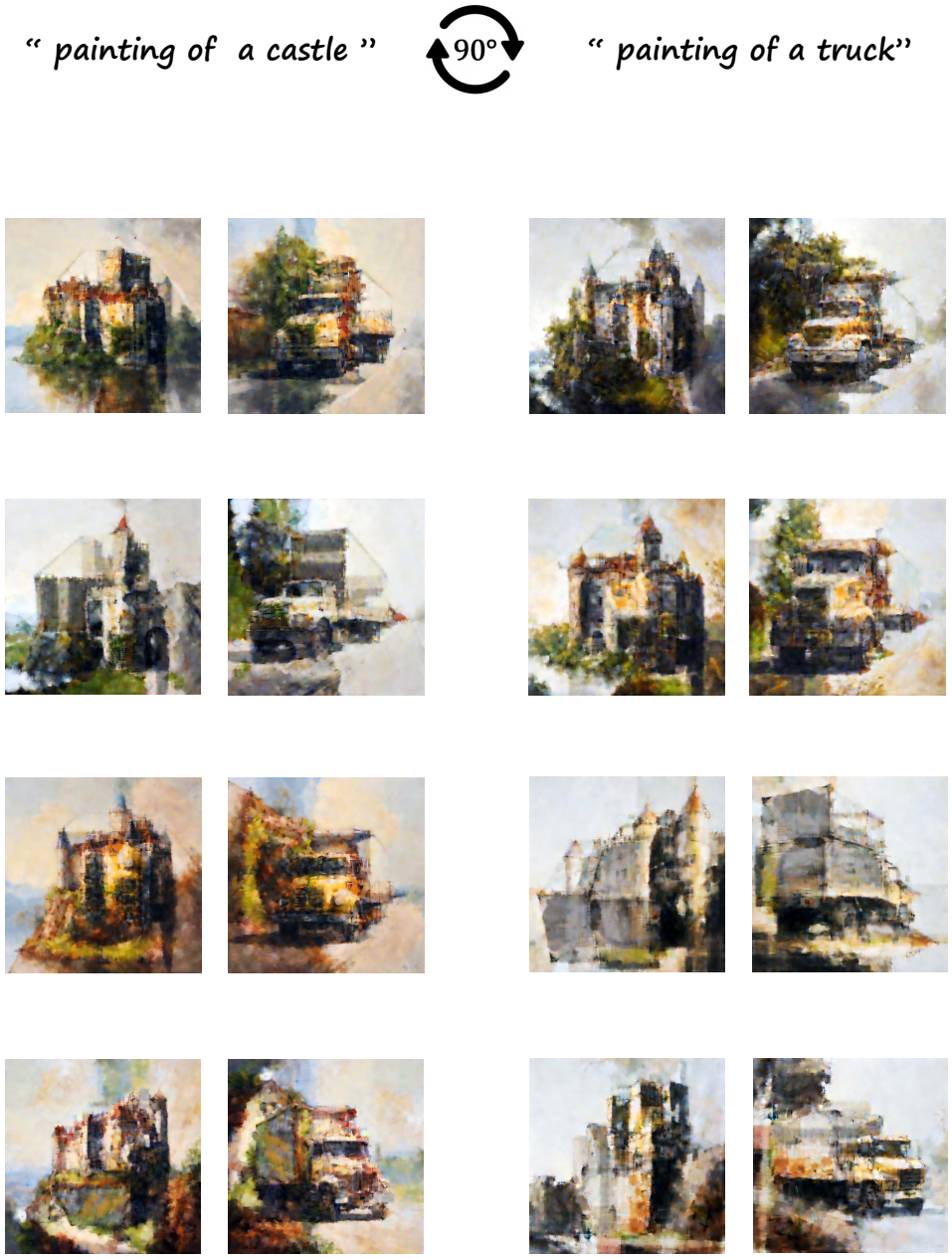


Figure 14: More diversified results-case 6.

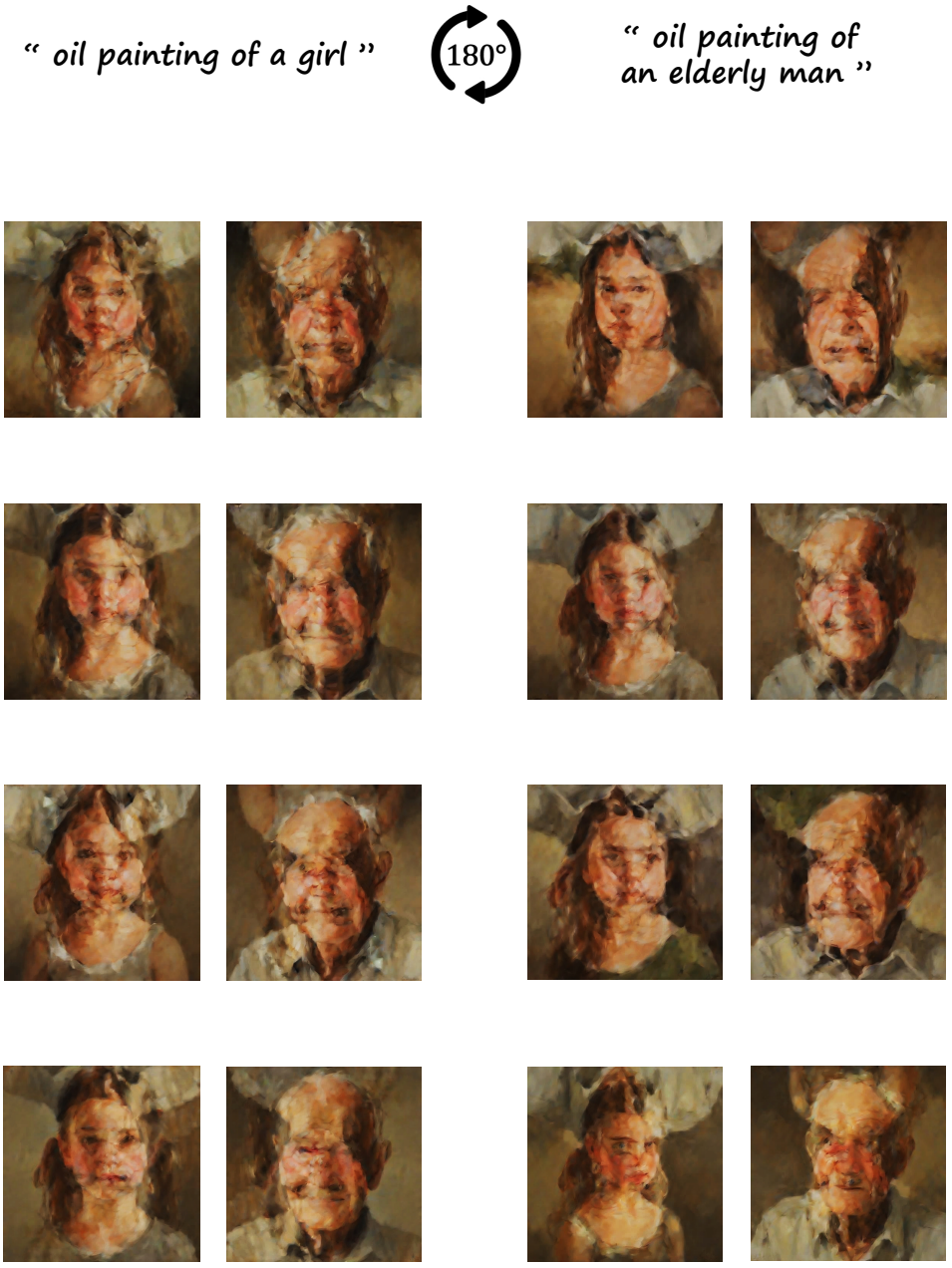


Figure 15: More diversified results-case 7.

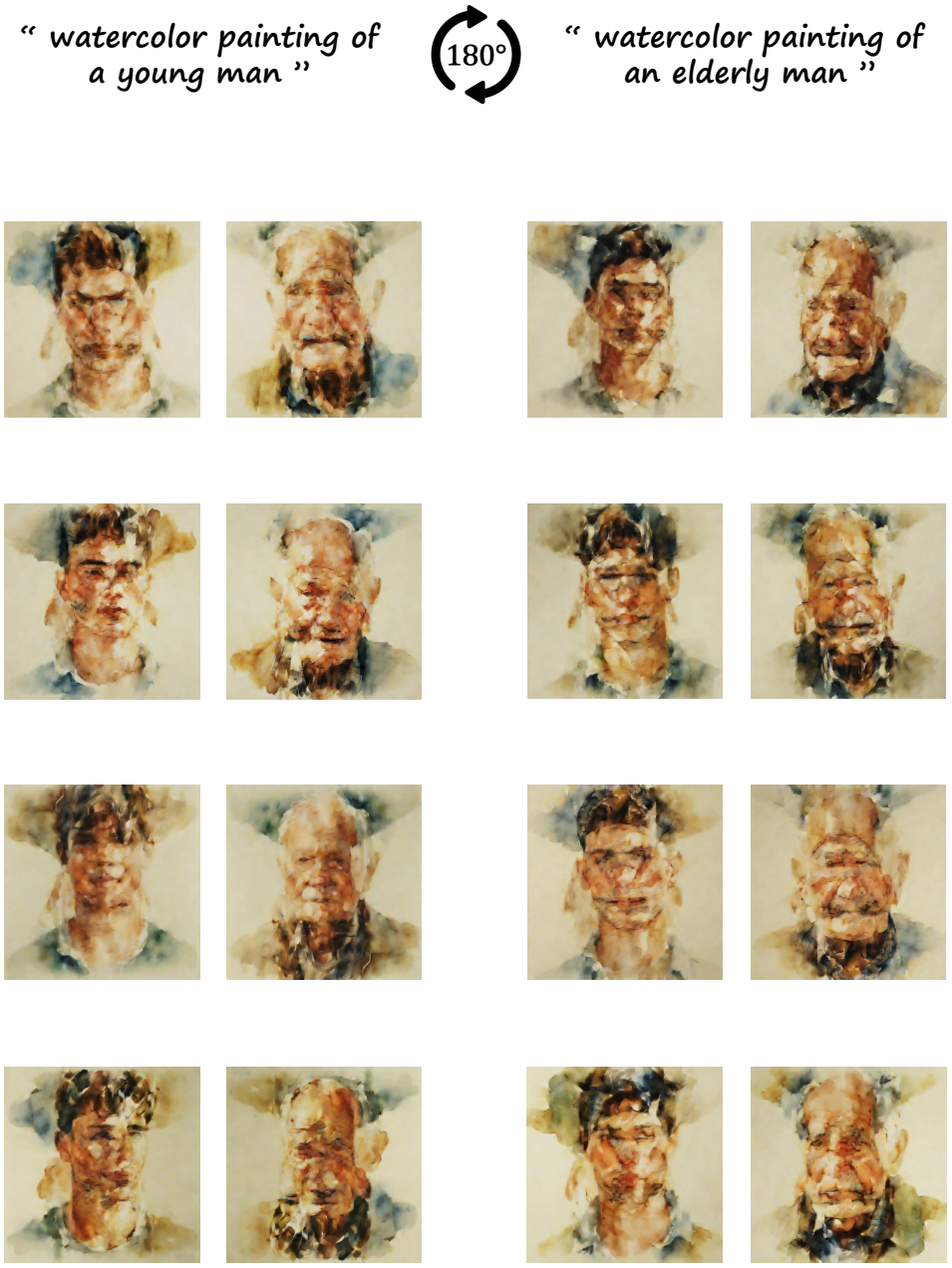


Figure 16: More diversified results-case 8.

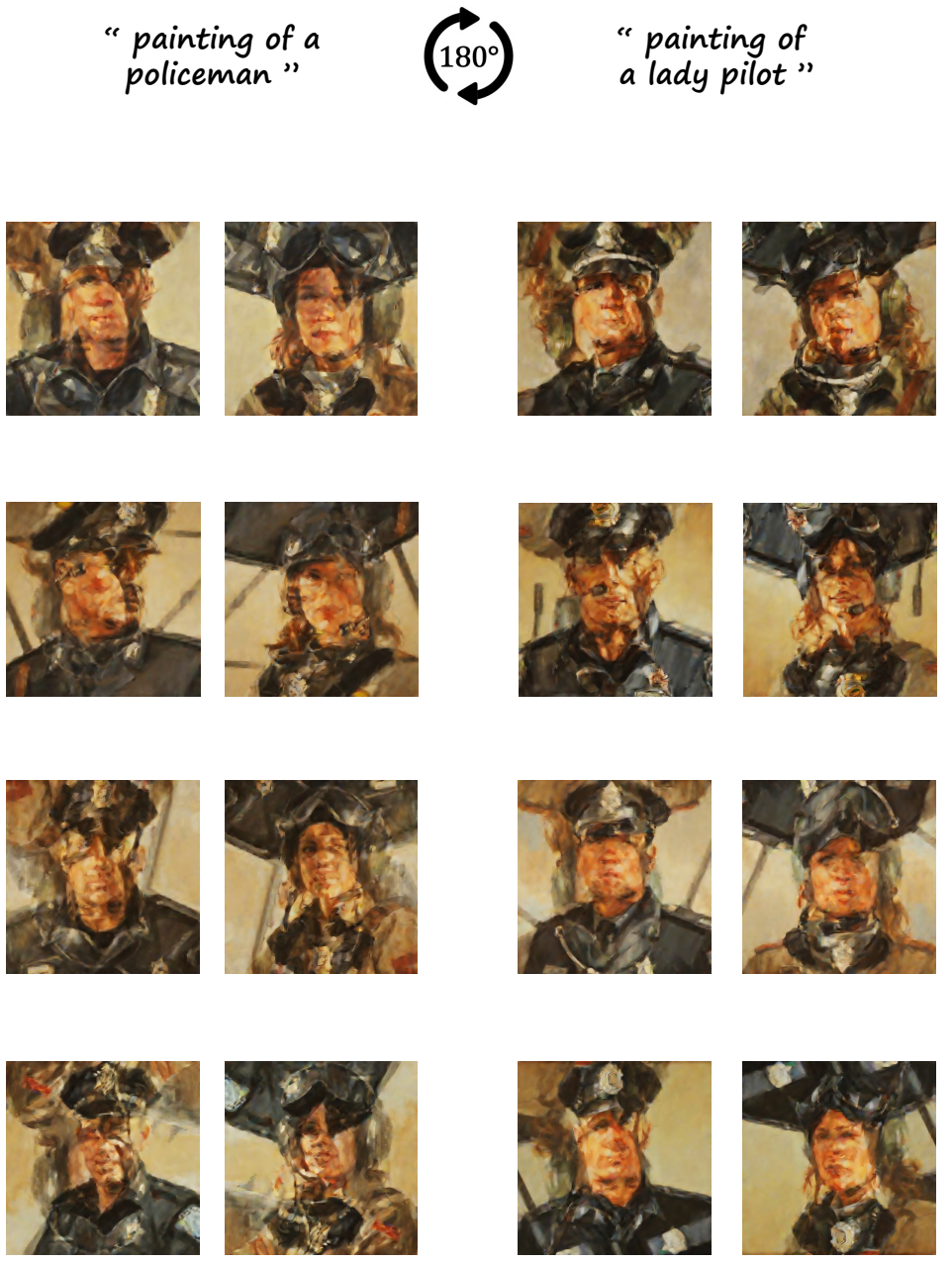


Figure 17: More diversified results-case 9.

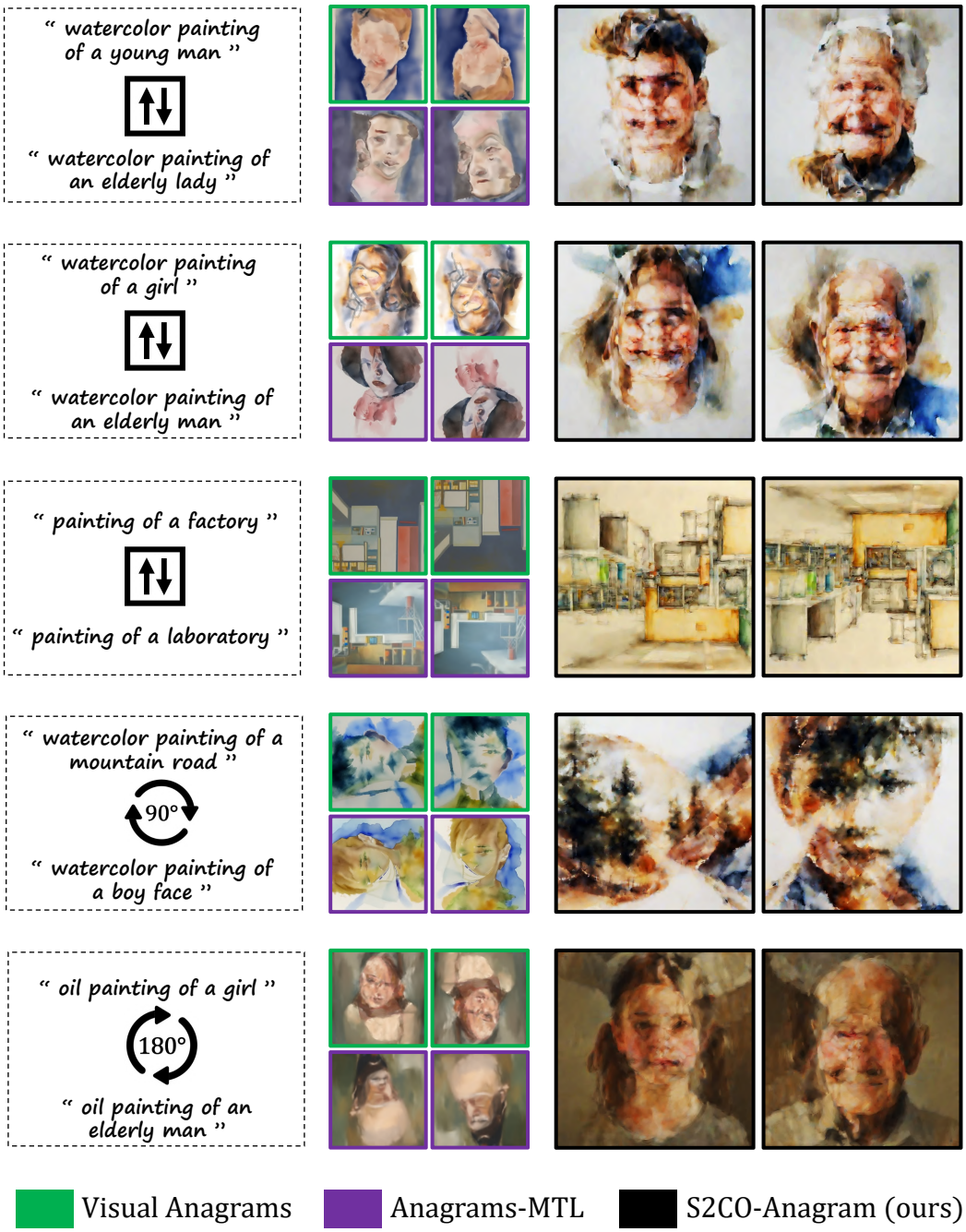


Figure 18: More qualitative method comparison-case 1. Small images are of 256×256 resolution, large images are of 512×512 resolution.

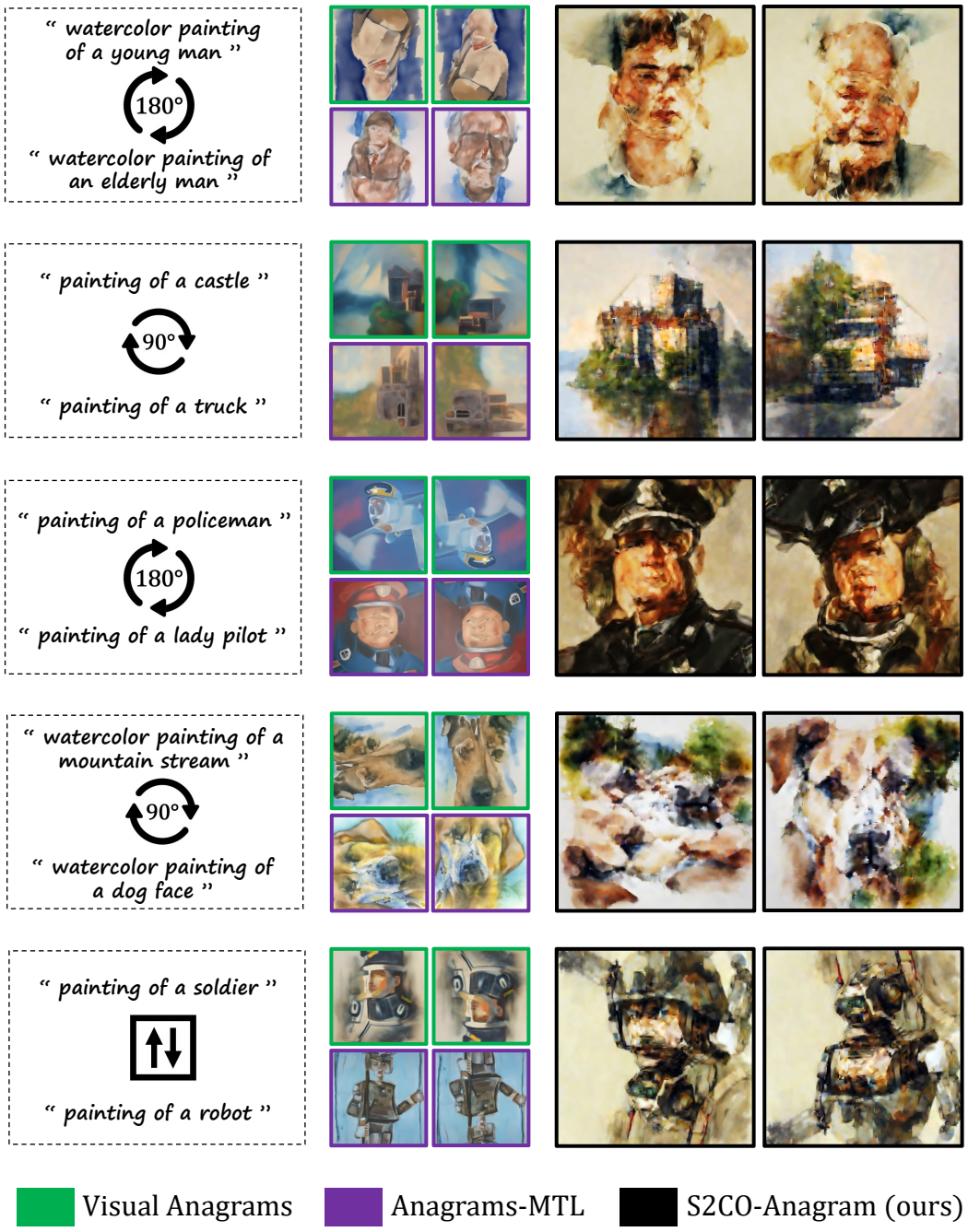


Figure 19: More qualitative method comparison-case 2. Small images are of 256×256 resolution, large images are of 512×512 resolution.