CLeaR: A Unified Framework for Resolving the Leakage–Degradation Dilemma in Style Transfer

Teng Zhou
Zhejiang University
tengzhou@zju.edu.cn

Yunhao Chen*
Fudan University
yhchen24@m.fudan.edu.cn

Abstract

Style transfer aims to render target content in the style of a reference image, but existing methods often suffer from content leakage, where objects, layouts, or semantics from the style reference appear in the generated output. Although prior data-driven and training-free methods can reduce leakage, they often face a leakage-degradation dilemma: stronger content suppression may weaken style fidelity, while richer style preservation may reintroduce unwanted reference content. We identify this dilemma across the full style-transfer pipeline, including feature separation, feature-space grounding, and diffusion generation. To address these issues, we propose CLeaR, a training-free framework for content-leakage-resistant style transfer. CLeaR first uses Orthogonal Subspace Projection to define content-reduced style targets in each vision foundation model (VFM) feature space. It then performs Ensemble Inversion, which optimizes a shared pixel-space style anchor satisfying style constraints across multiple VFMs. Finally, Energy-Guided Calibration maintains style alignment during diffusion sampling by steering the denoising trajectory toward the ensemble-defined style manifold. We further provide a theoretical analysis showing that the style-anchor estimation error decreases with the number of VFMs. Experiments on StyleBench demonstrate that CLeaR improves style alignment, reduces content leakage, and achieves better LLM-as-Judge evaluation compared with existing methods. The code is available at https://github.com/0606zt/CLeaR.

1 Introduction

Given a style reference image and a target content condition, style transfer (ST) aims to render the target content in the reference style. Since early neural style transfer, this task has been closely tied to content-style (C-S) separation [9]. With diffusion models becoming the dominant backbone for image generation [25, 22], ST is increasingly formulated as conditional generation [44, 53, 4, 12, 42]. However, incomplete C-S disentanglement can cause content leakage, where objects, layouts, or semantics from the style reference appear in the output [56, 27, 42], reducing quality and user control.

To mitigate content leakage, existing methods commonly use visual representations from pretrained or learned encoders to define a style condition that is later injected into a diffusion generator. Data-driven methods train style-aware encoders, disentanglement modules, adapters, or projection layers on paired or curated stylization data [8, 43, 54, 35, 17, 18, 23, 47]. Although effective in specific settings, they require additional training and are often constrained by the training distribution. Training-free methods instead manipulate pretrained VFM or diffusion features directly, such as subtracting content-related text features from image features [42], masking feature dimensions, shuffling spatial tokens, modifying attention keys and values, or guiding the denoising trajectory with reference-based objectives [56, 45, 28, 4, 12, 27]. Despite these differences, both paradigms typically define style within a particular feature space and rely on the diffusion model to generate the final image from the resulting edited or guided condition.

*Equal Contribution, Corresponding Author

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

Although these methods can reduce content leakage in some cases, they often lack stable control over the boundary between content removal and style preservation. As a result, existing methods tend to fail on different sides of the same dilemma [27]: methods that preserve rich style cues may also retain unwanted reference content, whereas methods that suppress reference content more strongly may remove important stylistic details. We refer to this as the leakage-degradation dilemma.

This dilemma is not caused by a single isolated step. It arises throughout the style-transfer pipeline, from separating C-S features, to grounding the style signal in a representation space, to injecting that signal during diffusion generation.

The first source is the *feature separation stage*, where content-related and style-related components are separated within a reference representation. Existing operations such as subtraction, masking, patch manipulation, or attention editing can reduce content signals, but are mostly heuristic and may either leave residual content or discard style information correlated with content. To address this issue, we introduce Orthogonal Subspace Projection (OSP), which defines the style target geometrically in each VFM feature space. Given a content proxy, it removes the component of the reference feature aligned with the content direction and retains the orthogonal residual as the style feature.

The second source is the *feature-space grounding stage*, where the extracted style signal is grounded in a particular representation space. Most existing methods rely on a single VFM feature space or a single learned representation, which is fragile because different VFMs encode style and content differently [24, 14, 51, 32]. Directly aligning multiple VFM spaces is impractical because their dimensions, metrics, and semantic structures differ, while learning cross-space projectors would add training cost and risk domain overfitting. Instead, we make use of a simpler and more general observation: image pixel space provides a shared domain associated with all VFM feature spaces. This allows heterogeneous style constraints to be combined by optimizing an image itself, rather than by aligning feature spaces directly. Based on this idea, we introduce Ensemble Inversion (EI), which optimizes a learnable image tensor into a shared style anchor whose features satisfy the model-specific style constraints across all VFM branches.

The third source is the *diffusion generation stage*, where the extracted style condition is injected into the generative process. Standard condition injectors such as IP-Adapter [49] are trained on ordinary image-text pairs rather than on content-reduced style anchors, so the denoising trajectory may drift from the intended style or reintroduce leaked content. To address this issue, we introduce Energy-Guided Calibration (EGC), a test-time guidance mechanism inspired by prior diffusion guidance methods [50, 1]. At selected denoising steps, it computes an ensemble feature-alignment energy between the current decoded image and the target style representation, and uses its gradient to calibrate the latent update toward the VFM-defined style manifold.

In summary, we introduce CLeaR (Content Leakage Resistant), a training-free framework for C-S disentangled style transfer. CLeaR defines content-reduced style targets through Orthogonal Subspace Projection, reconciles heterogeneous VFM targets through Ensemble Inversion, and maintains style alignment through Energy-Guided Calibration. Together, these components reduce content leakage at both the representation-extraction stage and the generation stage.

Our contributions are summarized as follows:

- We propose CLeaR, a training-free framework that extracts a shared pixel-space style anchor via Orthogonal Subspace Projection and Ensemble Inversion, and maintains style alignment through Energy-Guided Calibration.
- We provide a theoretical analysis showing that the style-anchor estimation error decreases with the number of VFMs under limited cross-model dependence.
- Experiments on StyleBench [8] show improvements in style alignment, content leakage metrics, and LLM-as-Judge evaluation.

2 Related Work

Style Transfer Gatys et al. [9] formalized style transfer by optimizing deep feature statistics. With diffusion models [25, 22] and conditioning adapters [49], style transfer has increasingly become a conditional generation task, where the central challenge is still C-S separation [44].

Existing methods can be broadly grouped into data-driven and training-free approaches. Data-driven methods learn style representations from paired or curated stylization data [18, 8, 43, 23, 54, 35, 17], or train projection modules for C-S composition [47]. These methods can be effective, but require additional training and are often constrained by the training distribution. Training-free methods instead manipulate pretrained VFM or diffusion features directly, including subtracting textual concepts from image embeddings [42], masking feature dimensions, shuffling spatial patches [56, 45, 19], modifying attention keys/values [28, 4, 12, 11], or adjusting the reverse diffusion trajectory through inversion or test-time optimization [53, 12, 27, 55]. Though effective, these methods often face a leakage-degradation dilemma across the style-transfer pipeline: feature operations may leave residual content or remove useful style cues, single-space representations are fragile across VFMs, and style conditions drift during diffusion sampling. Based on this analysis, CLeaR combines Orthogonal Subspace Projection, Ensemble Inversion, and Energy-Guided Calibration to form a unified solution for content-leakage-resistant style transfer.

Model Inversion Model inversion (MI) techniques originally emerged to analyze privacy risks by reconstructing training data from machine learning models [7, 6]. Early MI approaches primarily target classification networks, using gradient-based optimization and generative priors to reconstruct the original inputs from deep representations [52, 13, 38]. While traditional methods are generally tailored for classifiers, recent advancements [3] have extended inversion paradigms to extract high-dimensional visual information from complex, large-scale VFMs successfully.

These inversion techniques provide a natural bridge between feature spaces and the pixel domain. And in our C-S separation process, the image pixel space serves as a universal grounding domain that corresponds to all VFM feature spaces. Leveraging this insight, we employ model inversion to combine heterogeneous representations.

3 Methodology

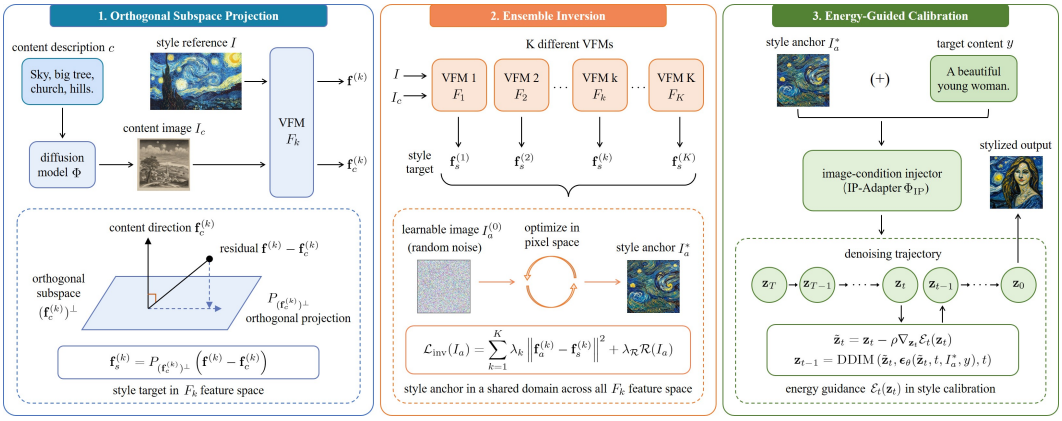


Figure 1: Overview of the CLeaR framework. The three components target the three failure sources in the entire style transfer process: Orthogonal Subspace Projection (feature separation), Ensemble Inversion (feature-space grounding), and Energy-Guided Calibration (diffusion generation).

CLeaR contains three components that correspond to the three failure sources discussed in Sec. 1. Orthogonal Subspace Projection defines a content-reduced style target in each VFM feature space. Ensemble Inversion combines these model-specific targets by optimizing a shared pixel-space style anchor. Energy-Guided Calibration then keeps the generated image aligned with the target style during diffusion sampling. The overall pipeline is summarized in Fig. 1 and Appendix A.

3.1 Orthogonal Subspace Projection

Given a style reference image I and its content description c [16], our goal is to remove content-related information from the reference feature while retaining style information. Let $F : \mathcal{I} \to \mathcal{H}$ be a pretrained VFM, and denote the reference feature as $\mathbf{f} = F(I)$. To obtain a modality-matched

content feature, we generate a content proxy image $I_c = \Phi(c)$ using a text-to-image diffusion model Φ, and compute $\mathbf{f}_c = F(I_c)$ [3, 10, 15].

We define the raw residual as

$$\mathbf{d} = \mathbf{f} - \mathbf{f}_c. \tag{1}$$

Since $\mathbf{d}$ may still contain content-aligned information, we remove the component of $\mathbf{d}$ along the content direction $\mathbf{f}_c$. We seek a style feature $\mathbf{f}_s = \mathbf{d} + \boldsymbol{\delta}$ that is orthogonal to $\mathbf{f}_c$, while keeping the adjustment $\boldsymbol{\delta}$ as small as possible:

$$\min_{\boldsymbol{\delta}} \|\boldsymbol{\delta}\|^2 \quad \text{s.t.} \quad \langle \mathbf{d} + \boldsymbol{\delta}, \mathbf{f}_c \rangle = 0. \tag{2}$$

Solving them gives

$$\mathbf{f}_s = \mathbf{d} - \frac{\langle \mathbf{d}, \mathbf{f}_c \rangle}{\|\mathbf{f}_c\|^2} \mathbf{f}_c = P_{\mathbf{f}_c^\perp}(\mathbf{d}), \tag{3}$$

where $P_{\mathbf{f}_c^\perp}$ denotes projection onto the orthogonal complement of $\mathbf{f}_c$. Thus, $\langle \mathbf{f}_s, \mathbf{f}_c \rangle = 0$, and $\mathbf{f}_s$ serves as a content-reduced style target in the VFM feature space.

3.2 Ensemble Inversion

The projection in Sec. 3.1 defines a style target within one VFM feature space. However, different VFMs capture different aspects of style, and their feature spaces cannot be directly aligned because they may have different dimensions and geometries. Instead, we make use of a simpler observation: image pixel space provides a shared domain associated with all VFM feature spaces.

Let $\{F_k\}_{k=1}^K$ be a set of pretrained VFMs. For each model, we compute

$$\mathbf{f}^{(k)} = F_k(I), \qquad \mathbf{f}_c^{(k)} = F_k(I_c), \tag{4}$$

and obtain the model-specific style target

$$\mathbf{f}_s^{(k)} = P_{(\mathbf{f}_c^{(k)})^\perp} \left(\mathbf{f}^{(k)} - \mathbf{f}_c^{(k)} \right). \tag{5}$$

We then optimize a single image I_a whose features match these style targets across all VFM branches. Specifically, let

$$\mathbf{f}_a^{(k)} = F_k(I_a). \tag{6}$$

The ensemble inversion objective is

$$\mathcal{L}_{\text{inv}}(I_a) = \sum_{k=1}^{K} \lambda_k \left\| \mathbf{f}_a^{(k)} - \mathbf{f}_s^{(k)} \right\|^2 + \lambda_{\mathcal{R}} \mathcal{R}(I_a), \tag{7}$$

where λ_k controls the contribution of each VFM and $\mathcal{R}(\cdot)$ is a regularizer for spatial smoothness.

Starting from a randomly initialized image tensor $I_a^{(0)}$, we update the anchor by gradient descent:

$$I_a^{(t+1)} = I_a^{(t)} - \eta \nabla_{I_a} \mathcal{L}_{\text{inv}} \left(I_a^{(t)} \right), \tag{8}$$

where η is the learning rate. After optimization, the resulting image I_a^* is used as the style anchor. This anchor provides a shared pixel-space representation whose VFM features match the content-reduced style targets across the ensemble.

3.3 Energy-Guided Calibration

The style anchor I_a^* is used as the image condition for diffusion generation. However, standard condition injectors such as IP-Adapter are trained on ordinary image-text pairs and may drift from the intended style during denoising. To reduce this drift, we introduce a test-time Energy-Guided Calibration step, following the idea of gradient-based diffusion guidance [50, 1].

At denoising step t, we first estimate the clean latent from the current noisy latent:

$$\hat{\mathbf{z}}_0(\mathbf{z}_t) = \frac{\mathbf{z}_t - \sigma_t \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, I_a^*, y)}{\alpha_t}, \tag{9}$$

where α_t and σ_t are determined by the noise schedule. Using the same VFM ensemble as in Ensemble Inversion, we define the style-alignment energy as

$$\mathcal{E}_t(\mathbf{z}_t) = \sum_{k=1}^{K} \lambda_k \left\| F_k(D(\hat{\mathbf{z}}_0(\mathbf{z}_t)),) - \mathbf{f}_s^{(k)} \right\|^2 . \tag{10}$$

where D is the VAE decoder. A lower value of $\mathcal{E}_t$ indicates stronger alignment with the target style representation.

We then use the energy gradient to calibrate the current latent:

$$\tilde{\mathbf{z}}_t = \mathbf{z}_t - \rho \nabla_{\mathbf{z}_t} \mathcal{E}_t(\mathbf{z}_t), \tag{11}$$

where ρ controls the calibration strength. The calibrated latent is passed to DDIM [37] update:

$$\mathbf{z}_{t-1} = \mathrm{DDIM}\left(\tilde{\mathbf{z}}_t, \boldsymbol{\epsilon}_\theta(\tilde{\mathbf{z}}_t, t, I_a^*, y), t\right). \tag{12}$$

This calibration requires no retraining and improves style alignment by steering the sampling trajectory toward the ensemble-defined style target.

3.4 Theoretical Analysis

We analyze how the accuracy of ensemble inversion depends on the number of feature extractors. Let $x^\star \in \mathbb{R}^d$ denote the ideal content-free style anchor, and let $F_1, \ldots, F_K$ be K feature extractors. For each model k, let the target style feature be s_k. We consider the estimator

$$\hat{x}_K = \arg\min_x \left\{ \frac{1}{K} \sum_{k=1}^{K} \|F_k(x) - s_k\|^2 + \lambda \|x - x^\star\|^2 \right\}, \tag{13}$$

where $\lambda > 0$ is a regularization coefficient.

Let

$$J_k := \nabla F_k(x^\star), \qquad H_K := \frac{1}{K} \sum_{k=1}^{K} J_k^\top J_k. \tag{14}$$

We write the model-specific target error as

$$s_k = F_k(x^\star) + \varepsilon_k, \qquad \mu_k := \mathbb{E}[\varepsilon_k], \qquad \tilde{\varepsilon}_k := \varepsilon_k - \mu_k. \tag{15}$$

Assumption 1 (Diversity and local identifiability)**.** *The centered residuals have bounded second moments and limited average cross-model dependence:*

$$\|\mathbb{E}[\tilde{\varepsilon}_k \tilde{\varepsilon}_k^\top]\|_{\mathrm{op}} \le \sigma^2, \qquad \|\mathbb{E}[\tilde{\varepsilon}_k \tilde{\varepsilon}_j^\top]\|_{\mathrm{op}} \le \rho_K \sigma^2 \quad (k \neq j), \tag{16}$$

where $\rho_K \in [0, 1]$ is an ensemble dependence coefficient. In addition, the ensemble is locally identifiable:

$$H_K \succeq m_K I \qquad \text{for some } m_K > 0. \tag{17}$$

Assumption 1 is motivated by two standard ideas. First, heterogeneous vision encoders can provide complementary feature views rather than identical errors, making cross-model dependence, rather than exact equality, the key quantity to control. Second, local recovery in inverse problems is typically tied to a nondegenerate Jacobian or sensitivity matrix. These perspectives are consistent with recent evidence on complementary encoder biases and ensemble diversity, as well as standard identifiability analyses in inverse problems [14, 46, 26, 5].

Theorem 1 (Dependence-controlled scaling with ensemble size)**.** *Under Assumption 1 and the usual local regularity conditions deferred to Appendix B, the expected inversion error satisfies*

$$\mathbb{E}\|\hat{x}_K - x^\star\|^2 \lesssim \frac{d\, \bar{M}_K\, \sigma^2}{(m_K + \lambda)^2} \left(\rho_K + \frac{1 - \rho_K}{K} \right), \tag{18}$$

where

$$\bar{M}_K := \frac{1}{K} \sum_{k=1}^{K} \|J_k\|_{\mathrm{op}}^2. \tag{19}$$

In particular, the gain from enlarging the ensemble is governed by the cross-model dependence coefficient ρ_K. When the centered residuals are weakly correlated across models, that is, when ρ_K is small, the dominant term decreases approximately as $1/K$.

Empirical connection The bound in Eq. 18 predicts that the benefit of increasing K depends on the complementarity of the VFM residuals. This prediction is consistent with the VFM composition ablation in Fig. 5: individual VFMs emphasize different aspects of C-S disentanglement, while increasing the number of VFMs improves both SA and CL. This empirical trend supports the interpretation that Ensemble Inversion benefits from complementary VFM constraints.

4 Experiments

4.1 Settings

Evaluation Dataset We evaluate our method on StyleBench [8], a recent dataset containing 40 content images and 490 style images across 73 distinct styles. For each style, we select one representative image (I) and generate its content description (c) using Qwen3 [48]. For content, we use the textual prompts (y) corresponding to the 40 content images as generation conditions. This yields 2,920 style-content pairs for our experiments.

Implementation Details For each style-content pair, we generate 10 images over 25 random seeds. We use five VFMs in CLeaR: CLIP [24, 30], CSD-CLIP [36], DINO [33], VGG [34], and Inception [39]. In Ensemble Inversion, all VFMs are equally weighted with $\lambda_k = 0.2$, and the regularizer $\mathcal{R}(\cdot)$ is total variation (TV) [29] with $\lambda_{\mathcal{R}} = 0.05$. We optimize for 300 iterations with learning rate $\eta = 0.01$. For Energy-Guided Calibration, guidance is applied only in the refinement stage [50], i.e., $t : T/10 \to 0$, for 5 iterations with correction strength $\rho = 0.01$. Further details are in Sec. 4.3.

Evaluation Metrics We evaluate seven metrics across four aspects:

- **Style Alignment**: We measure style alignment using CSD-CLIP cosine similarity and VGG Gram-matrix style loss [9] between the generated images and the style references.
- **Content Alignment**: We measure content alignment using CLIP cosine similarity between the generated images and the target content prompts.
- **Content Leakage**: We measure content leakage using DINO similarity and KID [2] between the generated images and the content images. We also use Qwen3 to rate leaked reference-content elements on a 1–5 scale.
- **Aesthetic Quality**: We measure aesthetic quality using the CLIP-Aesthetic score [31].

Baselines We compare CLeaR with eight recent methods: Attention Distillation [55], CSGO [47], DEADiff [23], InstantStyle [42], MaskST [56], RB-Modulation [27], StyleAligned [12], and StyleShot [8]. Since style transfer methods vary in input modalities, and our framework uses a style image and a content text prompt, we restrict our comparison to baselines that adopt the same modality combination to ensure a fair evaluation.

4.2 Comparisons

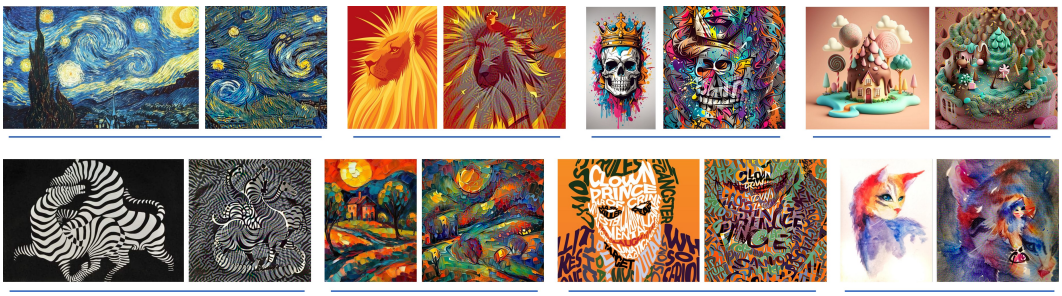


Figure 2: Visualization of style anchor images extracted by Ensemble Inversion. For each example, the left image is the reference and the right image is the style anchor. The anchor successfully achieves C-S disentanglement, retaining stylistic elements (e.g., color, texture, brushstroke) while suppressing content semantics.

Visualization of Style Anchors Fig. 2 visualizes the key intermediate output of our method, the style anchor image extracted via Ensemble Inversion. For each example, the left image is the original style reference I, and the right image is the corresponding style anchor I_a^*. The style anchor largely removes the content objects of the reference while faithfully preserving style-related representations. For instance, in the first example (row 1, col 1), the anchor removes the large tree, hills, and church steeple, retaining only the starry brushstrokes. In the seventh example (row 2, col 3), the clown figure is removed, leaving only the typographic style. These indicate that our method effectively disentangles content elements from stylistic attributes. More examples are provided in Appendix D.

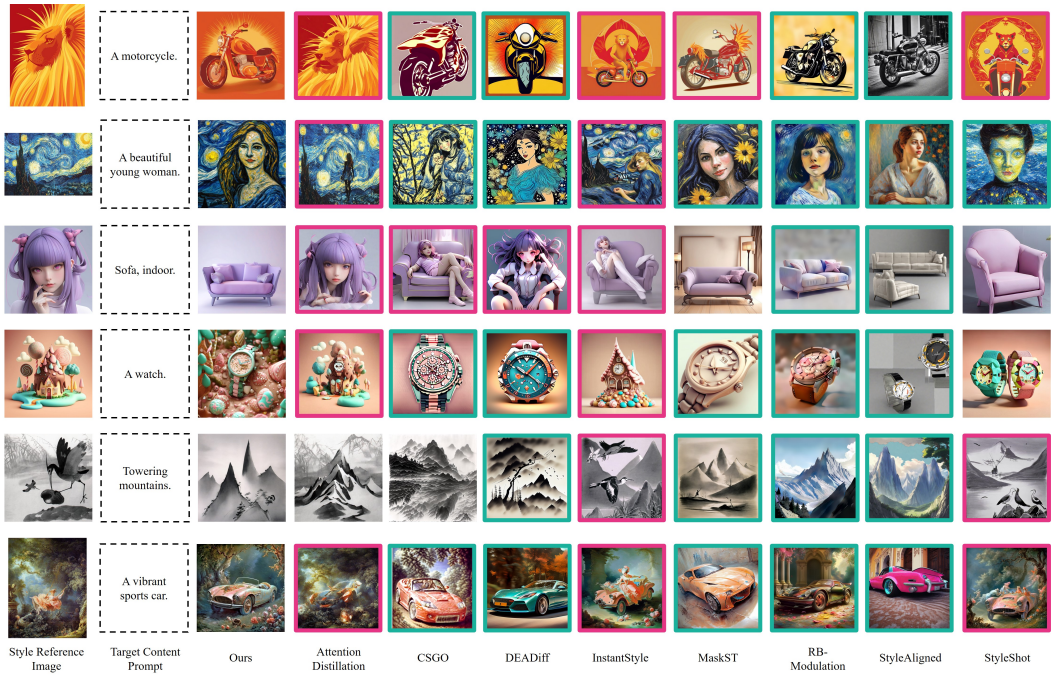


Figure 3: Qualitative comparison results. Existing methods suffer from either content leakage or style degradation, while our method achieves robust stylization without either problem.

Qualitative Results Fig. 3 shows the qualitative comparisons, with magenta boxes marking content leakage and cyan boxes marking style degradation. Existing methods often fail on one side of the leakage-degradation dilemma: data-driven methods such as StyleShot, DEADiff, and CSGO show limited style alignment on out-of-distribution references; feature-manipulation methods such as InstantStyle and MaskST either retain residual content or suppress style details; and trajectory-level methods such as RB-Modulation and StyleAligned can depend on unavailable style-name priors. Attention Distillation preserves texture but often introduces reference content. In contrast, CLeaR achieves stronger style transfer with less visible content leakage. More results are in Appendix C.

Quantitative Results Tab. 1 reports the quantitative comparison. CLeaR achieves the best performance on most metrics, with clear gains in the two key aspects of C-S disentanglement: style alignment and content leakage suppression. Compared with the strongest baseline, CLeaR improves CSD by 8.3%, Style Loss by 64.9%, DINO by 60.7%, KID by 51.1%, and AI-Scoring by 8.05%, while maintaining competitive content alignment with the second-best CLIP score. Its slightly lower aesthetic score may result from the inversion process, which prioritizes disentanglement and structural preservation over aesthetic smoothing. We further validate these conclusions in Appendix E using held-out VFMs and AI evaluators, which show consistent trends and confirm that the reported gains are not biased by the closed-loop evaluation setup.

4.3 Ablation Studies

Our ablations examine three key factors in CLeaR: the content description of the style reference in OSP, the VFM composition in EI, and the guidance stage in EGC. For clarity, we report two

Table 1: Quantitative comparison of CLeaR and other baselines, with the **best** and second-best scores marked accordingly. CLeaR outperforms others across most aspects, especially in style alignment, content leakage, and content alignment.

	Style Alignment		Content Alignment	Content Leakage			Aesthetic Quality
	CSD↑	Style Loss↓	CLIP↑	DINO↓	KID↑	AI Scoring↑	CLIP-Aesthetic↑
Attention Distillation	0.630	0.106	0.179	0.322	0.112	1.84	5.861
CSGO	0.438	0.302	0.215	0.168	0.237	3.65	5.912
DEADiff	0.369	0.259	0.225	0.209	0.233	3.64	6.349
InstantStyle	0.492	0.119	0.221	0.263	0.092	2.75	6.507
MaskST	0.416	0.139	0.225	0.197	0.103	3.45	6.605
RB-Modulation	0.443	0.177	0.234	0.213	0.111	3.82	**6.677**
StyleAligned	0.206	0.252	**0.241**	0.177	0.120	3.85	6.573
StyleShot	0.505	0.097	0.202	0.172	0.174	3.72	6.164
Ours	**0.682**	**0.034**	0.238	**0.066**	**0.358**	**4.16**	6.445

aggregate metrics: style alignment (SA↑) and content leakage suppression (CL↑). We first normalize each metric to $[0, 1]$ across all compared settings. SA is the average of normalized CSD and $1-$ normalized Style Loss, since higher CSD and lower Style Loss indicate better style alignment. CL is the average of $1-$ normalized DINO, normalized KID, and normalized AI score, since lower DINO similarity, higher KID, and higher AI score indicate less leakage.

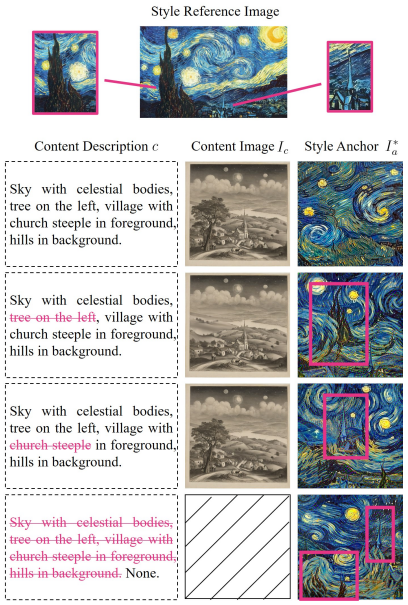


Figure 4: Style anchor extraction with varying content descriptions c. For a given reference, different c yield different content images I_c and style anchors I_a^*. Removing an object from c preserves that object in I_a^*, showing flexible disentanglement control.

Content Description of Style Reference The orthogonal projection in Eq. 3 requires a content description c of the style reference image I to generate a content image $I_c = \Phi(c)$. This description determines which content elements are treated as "content" to be subtracted from the style feature. To examine the flexibility of our C-S disentanglement, we manually vary c by removing specific objects (e.g., "tree", "church steeple") and observe the resulting style anchor I_a^* extracted via Ensemble Inversion.

As shown in Fig. 4, when c contains all content elements (row 1), I_c includes the full object set, and the extracted anchor retains only the starry style brushstrokes. Removing "tree" from c (row 2) causes the tree to reappear in I_a^*, while the church steeple is suppressed; removing "church steeple" (row 3) yields the opposite pattern. When no content is explicitly described (row 4), the inversion lacks guidance and retains many content artifacts. These results demonstrate that our framework allows fine-grained control over which content elements are disentangled, simply by editing the natural-language description c, without retraining any component.

VFM Selection To assess the effect of VFM composition, we evaluate Ensemble Inversion with $k = 1, 3, 5$ models selected from the five VFMs used in our main experiments.

As shown in Fig. 5, CLIP, DINO, and Inception mainly help suppress content leakage, while VGG and CSD-CLIP better preserve fine-grained style patterns. Increasing the number of VFMs consistently improves both SA and CL, showing that complementary VFM features lead to stronger C-S disentanglement.

Calibration Stage We apply Energy-Guided Calibration only in the final $T/10$ timesteps by default for efficiency. To study the effect of guidance timing, we divide sampling into three intervals: the chaotic stage $(t : T \to 4T/5)$, where the image is mostly noise; the semantic stage $(t : 4T/5 \to T/2)$,

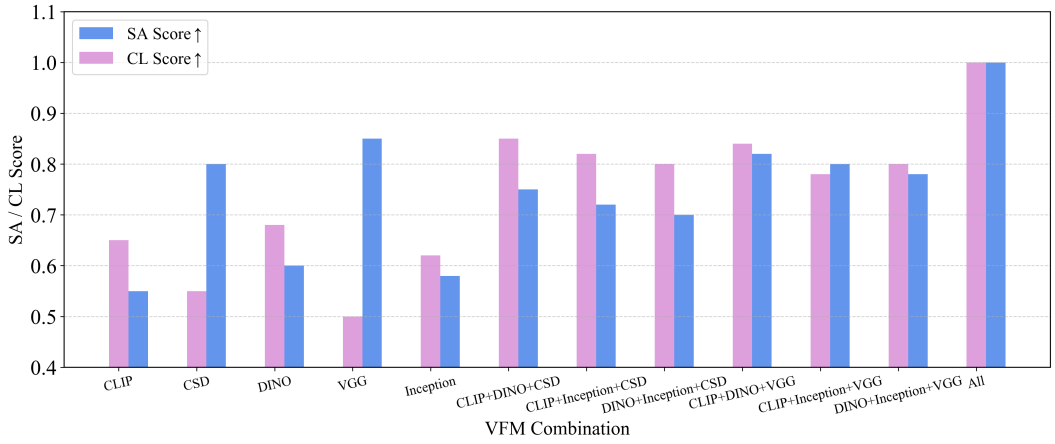


Figure 5: Ablation of VFM compositions. CLIP, DINO, and Inception excel at suppressing content leakage, while VGG and CSD-CLIP better preserve fine-grained style patterns. Increasing the number of models improves both metrics, with the full ensemble of five achieving the best performance.

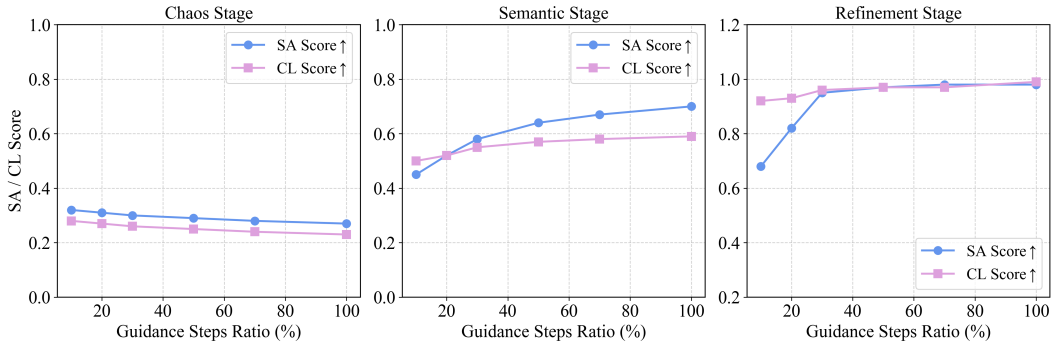


Figure 6: Effect of guidance stage in Energy-Guided Calibration. We compare applying guidance in the chaotic, semantic, and refinement stages, with the x-axis indicating the guided proportion of each stage. Guidance is ineffective in the chaotic stage, while in the semantic and refinement stages it improves SA with little change in CL. The best efficiency-performance trade-off is achieved by guiding the final 20% of the refinement stage, equivalent to the final $T/10$ steps.

where the global structure emerges; and the refinement stage $(t : T/2 \to 0)$, where local textures and details are formed. For each interval, we vary the guided timestep ratio from 10% to 100%.

As shown in Fig. 6, guidance is ineffective in the chaotic stage. In the semantic and refinement stages, SA improves as more guidance steps are used, while CL remains relatively stable, indicating that calibration mainly enhances style alignment without increasing content leakage. The best efficiency-performance trade-off is achieved by guiding 20% of the refinement stage, equivalent to the final $T/10$ timesteps, which we use as the default setting.

5 Conclusion

We presented CLeaR, a training-free framework for content-leakage-resistant style transfer. We identified that the leakage-degradation dilemma arises across the entire style transfer process: from separating C-S features, to grounding the style signal in a representation space, to injecting it during diffusion generation. CLeaR addresses these three sources with Orthogonal Subspace Projection for content-reduced style targets, Ensemble Inversion for multi-VFM style anchor extraction, and Energy-Guided Calibration for style-preserving diffusion sampling. Our theoretical analysis supports the benefit of using multiple VFMs, and experiments on StyleBench show improved style alignment and content leakage suppression. These results demonstrate the effectiveness of addressing C-S disentanglement jointly across feature separation, style representation, and generation.

Limitations Despite its effectiveness, CLeaR has several limitations. First, it introduces additional test-time computation via Ensemble Inversion and Energy-Guided Calibration. While the measured runtime is comparable to several training-free baselines, it is still slower than simpler feed-forward methods. Second, as analyzed in Appendix G, ambiguous content descriptions, degenerate anchors, and adapter misinterpretation of inversion noise can still lead to leakage and artifacts. These cases represent an important direction for future work.

Acknowledgement and Disclosure

This work was conducted without external funding or third-party support. The authors declare that they have no financial or non-financial competing interests related to this work.

References

[1] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 843–852, 2023.

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018.

[3] Yunhao Chen, Shujie Wang, Xin Wang, Ran He, Xingjun Ma, and Yu-Gang Jiang. Leakyclip: Extracting training data from clip. *arXiv preprint arXiv:2508.00756*, 2025.

[4] 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.

[5] Ariel Cintrón-Arias, HT Banks, Alex Capaldi, and Alun L Lloyd. A sensitivity matrix based methodology for inverse problem formulation. *arXiv preprint arXiv:2004.06831*, 2020.

[6] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

[7] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In *23rd USENIX security symposium (USENIX Security 14)*, pages 17–32, 2014.

[8] Junyao Gao, Yanan Sun, Yanchen Liu, Yinhao Tang, Yanhong Zeng, Ding Qi, Kai Chen, and Cairong Zhao. Styleshot: A snapshot on any style. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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

[10] Brian Gordon, Yonatan Bitton, Yonatan Shafir, Roopal Garg, Xi Chen, Dani Lischinski, Daniel Cohen-Or, and Idan Szpektor. Mismatch quest: Visual and textual feedback for image-text misalignment. In *European Conference on Computer Vision*, pages 310–328. Springer, 2024.

[11] Feihong He, Gang Li, Fuhui Sun, Mengyuan Zhang, Lingyu Si, Xiaoyan Wang, and Li Shen. Freestyle: Free lunch for text-guided style transfer using diffusion models. *Pattern Recognition*, page 113093, 2026.

[12] 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.

[13] Mostafa Kahla, Si Chen, Hoang Anh Just, and Ruoxi Jia. Label-only model inversion attacks via boundary repulsion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15025–15033. IEEE, 2022.

[14] Oguzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. In *European Conference on Computer Vision*, pages 113–132. Springer, 2024.

[15] Bumsoo Kim, Yeonsik Jo, Jinhyung Kim, and Seunghwan Kim. Misalign, contrast then distill: Rethinking misalignments in language-image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2563–2572, 2023.

[16] Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhuo Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, et al. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*, 2026.

[17] Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Yibo Wang, Xintao Wang, Yujiu Yang, and Ying Shan. Stylecrafter: Enhancing stylized text-to-video generation with style adapter. *arXiv preprint arXiv:2312.00330*, 2023.

[18] Pingchuan Ma, Xiaopei Yang, Yusong Li, Ming Gui, Felix Krause, Johannes Schusterbauer, and Björn Ommer. Scflow: Implicitly learning style and content disentanglement with flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14919–14929, 2025.

[19] Lilian Ngweta, Subha Maity, Alex Gittens, Yuekai Sun, and Mikhail Yurochkin. Simple disentanglement of style and content in visual representations. In *International Conference on Machine Learning*, pages 26063–26086. PMLR, 2023.

[20] Zach Nussbaum, Brandon Duderstadt, and Andriy Mulyar. Nomic embed vision: Expanding the latent space. *arXiv preprint arXiv:2406.18587*, 2024.

[21] OpenAI. Gpt-5.5 system card. `https://openai.com/index/gpt-5-5-system-card/`, 2026.

[22] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

[23] Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8693–8702, 2024.

[24] 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.

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn 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.

[26] Simon Rouchier. Solving inverse problems in building physics: An overview of guidelines for a careful and optimal use of data. *Energy and Buildings*, 166:178–195, 2018.

[27] Litu Rout, Yujia Chen, Nataniel Ruiz, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Rb-modulation: Training-free stylization using reference-based modulation. In *International Conference on Learning Representations*, volume 2025, pages 56870–56905, 2025.

[28] Aniket Roy, Shubhankar Borse, Shreya Kadambi, Debasmit Das, Shweta Mahajan, Risheek Garrepalli, Hyojin Park, Ankita Nayak, Rama Chellappa, Munawar Hayat, et al. Duolora: Cycle-consistent and rank-disentangled content-style personalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15395–15404, 2025.

[29] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[30] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *International Conference on Machine Learning*, pages 43685–43704. PMLR, 2024.

[31] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[32] Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in vlms. *Advances in Neural Information Processing Systems*, 38:17929–17960, 2026.

[33] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[35] Kihyuk Sohn, Lu Jiang, Jarred Barber, Kimin Lee, Nataniel Ruiz, Dilip Krishnan, Huiwen Chang, Yuanzhen Li, Irfan Essa, Michael Rubinstein, et al. Styledrop: Text-to-image synthesis of any style. *Advances in Neural Information Processing Systems*, 36:66860–66889, 2023.

[36] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024.

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[38] Lukas Struppek, Dominik Hintersdorf, Antonio De Almeida Correia, Antonia Adler, and Kristian Kersting. Plug & play attacks: Towards robust and flexible model inversion attacks. In *International Conference on Machine Learning*, pages 20522–20545, 2022.

[39] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[40] Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

[41] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[42] Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024.

[43] Ye Wang, Zili Yi, Yibo Zhang, Peng Zheng, Xuping Xie, Jiang Lin, Yilin Wang, and Rui Ma. Omnistyle2: Scalable and high quality artistic style transfer data generation via destylization. *arXiv preprint arXiv:2509.05970*, 2025.

[44] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7677–7689, 2023.

[45] Zhouxia Wang, Xintao Wang, Liangbin Xie, Zhongang Qi, Ying Shan, Wenping Wang, and Ping Luo. Styleadapter: A unified stylized image generation model. *International Journal of Computer Vision*, 133(4):1894–1911, 2025.

[46] Danny Wood, Tingting Mu, Andrew M Webb, Henry WJ Reeve, Mikel Lujan, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of machine learning research*, 24(359):1–49, 2023.

[47] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 38, pages 100506–100546, 2025.

[48] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[49] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[50] 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.

[51] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023.

[52] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 250–258. IEEE, 2020.

[53] 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.

[54] Zhanjie Zhang, Quanwei Zhang, Wei Xing, Guangyuan Li, Lei Zhao, Jiakai Sun, Zehua Lan, Junsheng Luan, Yiling Huang, and Huaizhong Lin. Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 7396–7404, 2024.

[55] 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.

[56] Lin Zhu, Xinbing Wang, Chenghu Zhou, Qinying Gu, and Nanyang Ye. Less is more: Masking elements in image condition features avoids content leakages in style transfer diffusion models. In *International Conference on Learning Representations*, volume 2025, pages 66094–66126, 2025.

A Overall Pipeline of CLeaR

We provide detailed pseudocode in Alg. 1 to facilitate a better understanding of our framework. The algorithm consists of three main stages: (1) computing orthogonal style anchors for multiple VFMs, (2) inverting these anchors into a pixel-space style image via ensemble optimization, and (3) generating the final output with energy-guided calibration.

In Stage 1, we generate a content image I_c from the description c and compute orthogonal style features $\mathbf{f}_s^{(k)}$ for each VFM. Stage 2 optimizes a pixel-space image I_a to match these features simultaneously, producing a style anchor I_a^*. Stage 3 uses the IP-Adapter Φ_{IP} to condition the diffusion model on I_a^* and the target content y (or I_y), while additionally applying an energy correction based on the same VFMs. The energy guidance is only applied during the semantic stage (e.g., intermediate timesteps) to balance efficiency and style fidelity. The final output I_{out} is the decoded latent $\mathbf{z}_0$. Hyperparameters such as λ_k, $\lambda_{\mathcal{R}}$, η, ρ are set empirically.

Algorithm 1 Overall Pipeline of CLeaR

Require: Style image I, content description c, target content prompt y (or target image I_y), pre-trained VFMs $\{F_k\}_{k=1}^K$, text-to-image model Φ, IP-Adapter Φ_{IP}, VAE decoder D, DDIM sampler with $\boldsymbol{\epsilon}_\theta$

Ensure: Output image I_{out}

— Stage 1: orthogonal style anchors —
$I_c \leftarrow \Phi(c)$ {generate content image}
for each $k = 1$ to K **do**
  $\mathbf{f}^{(k)} \leftarrow F_k(I)$, $\mathbf{f}_c^{(k)} \leftarrow F_k(I_c)$
  $\mathbf{f}_s^{(k)} \leftarrow \mathbf{f}^{(k)} - \mathbf{f}_c^{(k)}$
  $\mathbf{f}_s^{(k)} \leftarrow \mathbf{f}_s^{(k)} - \frac{\langle \mathbf{f}_s^{(k)}, \mathbf{f}_c^{(k)} \rangle}{\|\mathbf{f}_c^{(k)}\|^2} \mathbf{f}_c^{(k)}$ {orthogonal projection}
end for
— Stage 2: ensemble inversion to obtain style anchor image —
Initialize $I_a \sim \mathcal{N}(0, \sigma^2)$ {random tensor}
for $t = 1$ to T_{inv} **do**
  for each $k = 1$ to K **do**
    $\mathbf{f}_a^{(k)} \leftarrow F_k(I_a)$
  end for
  $\mathcal{L}_{\text{inv}} \leftarrow \sum_{k=1}^K \lambda_k \|\mathbf{f}_a^{(k)} - \mathbf{f}_s^{(k)}\|^2 + \lambda_{\mathcal{R}} \mathcal{R}(I_a)$
  $I_a \leftarrow I_a - \eta \nabla_{I_a} \mathcal{L}_{\text{inv}}$
end for
$I_a^* \leftarrow I_a$
— Stage 3: energy-guided calibration during generation —
Initialize $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$
for $t = T$ down to 1 **do**
  if $t \in \mathcal{T}_{\text{cal}}$ **then**
    $\boldsymbol{\epsilon}_t \leftarrow \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, I_a^*, y)$
    $\hat{\mathbf{z}}_0 \leftarrow (\mathbf{z}_t - \sigma_t \boldsymbol{\epsilon}_t)/\alpha_t$
    $\hat{\mathbf{x}}_0 \leftarrow D(\hat{\mathbf{z}}_0)$
    $\mathcal{E}_t \leftarrow \sum_{k=1}^K \lambda_k \left\| F_k(\hat{\mathbf{x}}_0) - \mathbf{f}_s^{(k)} \right\|^2$
    $\tilde{\mathbf{z}}_t \leftarrow \mathbf{z}_t - \rho \nabla_{\mathbf{z}_t} \mathcal{E}_t$
    $\mathbf{z}_{t-1} \leftarrow \text{DDIM}\big(\tilde{\mathbf{z}}_t, \boldsymbol{\epsilon}_\theta(\tilde{\mathbf{z}}_t, t, I_a^*, y), t\big)$
  else
    $\mathbf{z}_{t-1} \leftarrow \text{DDIM}\big(\mathbf{z}_t, \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, I_a^*, y), t\big)$
  end if
end for
$I_{\text{out}} \leftarrow D(\mathbf{z}_0)$
return I_{out}

B Proof of the Multi-Model Inversion Error Bound

We provide a local first-order analysis of ensemble inversion.

Setup Let

$$\hat{x}_K = \arg\min_x \left\{ \frac{1}{K} \sum_{k=1}^{K} \|F_k(x) - s_k\|^2 + \lambda \|x - x^\star\|^2 \right\}, \tag{20}$$

where $x^\star \in \mathbb{R}^d$ is the ideal content-free style anchor. Write

$$\hat{\Delta}_K := \hat{x}_K - x^\star. \tag{21}$$

Assumption 2 (Local first-order regime)**.** *Each F_k is differentiable in a neighborhood of $x^\star$, and for sufficiently small Δ,*

$$F_k(x^\star + \Delta) = F_k(x^\star) + J_k \Delta + r_k(\Delta), \qquad J_k := \nabla F_k(x^\star), \tag{22}$$

where the remainder $r_k(\Delta)$ is lower-order than the linear term as $\|\Delta\| \to 0$.

Assumption 3 (Ensemble-average bias cancellation)**.** *The ensemble-averaged image-space bias vanishes:*

$$\frac{1}{K} \sum_{k=1}^{K} J_k^\top \mu_k = 0, \qquad \mu_k := \mathbb{E}[\varepsilon_k]. \tag{23}$$

Assumption 4 (Bounded Jacobians)**.** *The local Jacobians are uniformly bounded:*

$$\|J_k\|_{\mathrm{op}}^2 \le M_k. \tag{24}$$

Proposition 1. *Under Assumption 1 and Appendix Assumptions 2-4,*

$$\mathbb{E}\|\hat{x}_K - x^\star\|^2 \lesssim \frac{d\, \bar{M}_K\, \sigma^2}{(m_K + \lambda)^2} \left(\rho_K + \frac{1 - \rho_K}{K} \right), \tag{25}$$

where

$$\bar{M}_K := \frac{1}{K} \sum_{k=1}^{K} M_k. \tag{26}$$

Proof. Under Assumption 2, the leading behavior is governed by the linearized objective

$$\tilde{\Delta}_K = \arg\min_\Delta \left\{ \frac{1}{K} \sum_{k=1}^{K} \|J_k \Delta - \varepsilon_k\|^2 + \lambda \|\Delta\|^2 \right\}. \tag{27}$$

Expanding the quadratic objective gives

$$\begin{aligned} \frac{1}{K} \sum_{k=1}^{K} \|J_k \Delta - \varepsilon_k\|^2 + \lambda \|\Delta\|^2 &= \frac{1}{K} \sum_{k=1}^{K} \left(\Delta^\top J_k^\top J_k \Delta - 2 \varepsilon_k^\top J_k \Delta + \varepsilon_k^\top \varepsilon_k \right) + \lambda \Delta^\top \Delta \\ &= \Delta^\top H_K \Delta - 2 b_K^\top \Delta + \lambda \Delta^\top \Delta + \text{constant}, \end{aligned} \tag{28}$$

where

$$b_K := \frac{1}{K} \sum_{k=1}^{K} J_k^\top \varepsilon_k. \tag{29}$$

Differentiating Eq. 28 with respect to Δ and setting the gradient to zero yields

$$(H_K + \lambda I) \tilde{\Delta}_K = b_K. \tag{30}$$

Hence

$$\tilde{\Delta}_K = (H_K + \lambda I)^{-1} b_K. \tag{31}$$

By Assumption 1,

$$H_K + \lambda I \succeq (m_K + \lambda) I, \tag{32}$$

so

$$\|(H_K + \lambda I)^{-1}\|_{\mathrm{op}} \le \frac{1}{m_K + \lambda}. \tag{33}$$

Therefore,

$$\mathbb{E}\|\tilde{\Delta}_K\|^2 \le \frac{1}{(m_K + \lambda)^2} \mathbb{E}\|b_K\|^2. \tag{34}$$

We now decompose b_K into bias and centered fluctuation terms:

$$b_K = \frac{1}{K} \sum_{k=1}^{K} J_k^\top \mu_k + \frac{1}{K} \sum_{k=1}^{K} J_k^\top \tilde{\varepsilon}_k =: \beta_K + u_K. \tag{35}$$

By Assumption 3,

$$\beta_K = 0. \tag{36}$$

Hence

$$b_K = u_K \qquad \text{and} \qquad \mathbb{E}\|b_K\|^2 = \mathbb{E}\|u_K\|^2. \tag{37}$$

Now

$$u_K = \frac{1}{K} \sum_{k=1}^{K} J_k^\top \tilde{\varepsilon}_k, \tag{38}$$

so

$$\mathbb{E}\|u_K\|^2 = \frac{1}{K^2} \sum_{k=1}^{K} \sum_{j=1}^{K} \mathbb{E}\left[\tilde{\varepsilon}_k^\top J_k J_j^\top \tilde{\varepsilon}_j\right]. \tag{39}$$

For the diagonal terms,

$$\begin{aligned} \mathbb{E}\left[\tilde{\varepsilon}_k^\top J_k J_k^\top \tilde{\varepsilon}_k\right] &= \mathrm{tr}\left(J_k J_k^\top \mathbb{E}[\tilde{\varepsilon}_k \tilde{\varepsilon}_k^\top]\right) \\ &\le d \, \|J_k\|_{\mathrm{op}}^2 \, \sigma^2 \\ &\le d \, M_k \, \sigma^2. \end{aligned} \tag{40}$$

For the off-diagonal terms $k \neq j$, Assumption 1 and Assumption 4 give

$$\begin{aligned} \left|\mathbb{E}\left[\tilde{\varepsilon}_k^\top J_k J_j^\top \tilde{\varepsilon}_j\right]\right| &= \left|\mathrm{tr}\left(J_k J_j^\top \mathbb{E}[\tilde{\varepsilon}_j \tilde{\varepsilon}_k^\top]\right)\right| \\ &\le d \, \|J_k\|_{\mathrm{op}} \|J_j\|_{\mathrm{op}} \, \rho_K \sigma^2 \\ &\le d \, (M_k M_j)^{1/2} \rho_K \sigma^2 \\ &\le \frac{d \, \rho_K \, \sigma^2}{2} (M_k + M_j). \end{aligned} \tag{41}$$

Substituting Eq. 40 and Eq. 41 into Eq. 39, we obtain

$$\begin{aligned} \mathbb{E}\|u_K\|^2 &\le \frac{1}{K^2} \left[\sum_{k=1}^{K} d \, M_k \, \sigma^2 + \sum_{k \neq j} \frac{d \, \rho_K \, \sigma^2}{2} (M_k + M_j) \right] \\ &= \frac{d \, \sigma^2}{K^2} \left[\sum_{k=1}^{K} M_k + \rho_K (K-1) \sum_{k=1}^{K} M_k \right] \\ &= d \, \bar{M}_K \, \sigma^2 \left(\frac{1}{K} + \rho_K \frac{K-1}{K} \right) \\ &= d \, \bar{M}_K \, \sigma^2 \left(\rho_K + \frac{1 - \rho_K}{K} \right). \end{aligned} \tag{42}$$

Combining Eq. 34, Eq. 37, and Eq. 42 yields

$$\mathbb{E}\|\tilde{\Delta}_K\|^2 \le \frac{d \, \bar{M}_K \, \sigma^2}{(m_K + \lambda)^2} \left(\rho_K + \frac{1 - \rho_K}{K} \right). \tag{43}$$

Under Assumption 2, the neglected remainder is lower-order in the local regime, so the same leading-order bound applies to the exact estimator $\hat{x}_K - x^\star$. □

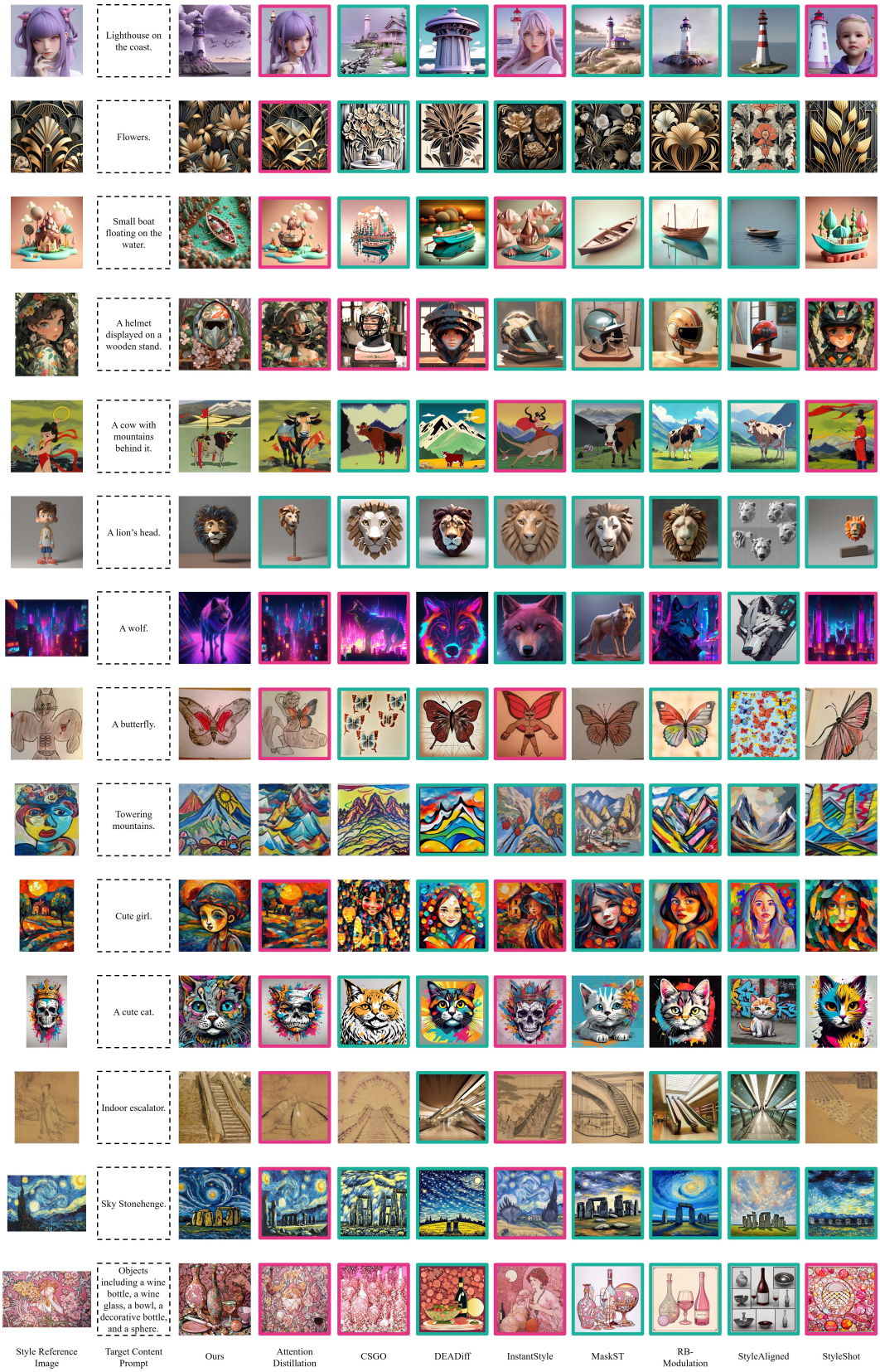


Figure A1: Extended qualitative comparisons, with boxes highlighting content leakage and style degradation. Our method maintains balanced stylization across all examples.

C More Qualitative Comparison Results

Fig. A1 provides additional qualitative comparisons. Consistent with the observations in Sec. 4.2, existing methods still suffer from either content leakage (magenta boxes) or style degradation (cyan boxes) across diverse examples, while our CLeaR achieves clean stylization without compromising either aspect.

D More Examples of Style Anchors

Fig. A2 presents additional style anchors extracted by our Ensemble Inversion. Across various artistic styles, the anchors consistently suppress content semantics (e.g., objects, faces, text) while preserving style attributes such as color, texture, and brushstrokes.

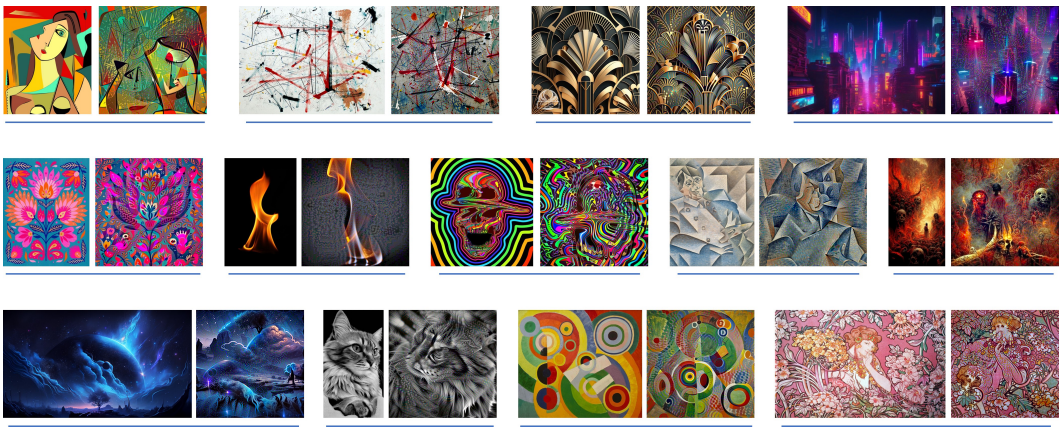


Figure A2: Additional style anchor examples. Each pair shows the original reference (left) and the extracted anchor (right).

E Comparisons on Held-Out VFMs and AI Evaluator

Table A1: Performance of held-out VFMs and independent AI evaluators, with the **best** and second-best scores marked accordingly.

	Style Alignment		Content Alignment		Content Leakage				Aesthetic Quality	
	SigLIP↑	Nomic↑	SigLIP↑	Nomic↑	SigLIP↓	Nomic↓	Kimi-K2.6↑	GPT-5.5↑	Kimi-K2.6↑	GPT-5.5↑
Attention Distillation	0.669	0.842	0.0867	0.0573	0.588	0.797	2.03	1.68	3.20	3.10
CSGO	0.592	0.800	0.117	0.0777	0.539	0.774	3.68	3.27	3.43	3.36
DEADiff	0.563	0.788	0.126	0.0812	0.552	0.761	3.71	3.24	3.62	3.58
InstantStyle	0.609	0.800	0.123	0.0826	0.529	0.760	2.91	2.47	3.89	3.86
MaskST	0.588	0.790	0.128	0.0833	**0.521**	0.753	3.56	3.08	4.03	4.01
RB-Modulation	0.584	0.789	0.136	0.0887	0.557	0.761	3.92	3.53	**4.48**	4.39
StyleAligned	0.544	0.772	**0.137**	**0.0893**	0.531	0.751	4.03	3.58	4.39	4.18
StyleShot	0.612	0.814	0.0997	0.0698	0.534	0.771	3.79	3.41	3.78	3.74
Ours	**0.754**	**0.876**	**0.137**	0.0803	0.527	**0.748**	**4.28**	**3.92**	4.20	**4.52**

To address the concern that CLeaR optimizes and evaluates on overlapping VFM features (CSD-CLIP, VGG, DINO) and that Qwen3 is used both for content description generation and AI-based evaluation, we conduct additional experiments with held-out VFMs and independent AI evaluators. We add SigLIP [41] and Nomic [20] as additional VFMs. Since our original selection (e.g., CLIP, DINO, CSD-CLIP) already covers the most standard models for style transfer and vision-language representation, few suitable alternatives remain. For AI evaluation, we include Kimi-K2.6 [40] and GPT-5.5 [21]. To ensure comprehensive coverage, we evaluate these new metrics across all four aspects.

As shown in Tab. A1, CLeaR consistently achieves the best or competitive performance across all held-out VFMs and independent AI evaluators. These results confirm that our improvements generalize beyond the models used during optimization and are not biased by the closed-loop evaluation setup.

F Ablation on Style Categories

To investigate performance variation across different styles, we compute SA and CL scores for all 73 style types in StyleBench.

As seen in Tab. A2, styles whose stylistic identity is carried by rendering rather than subject matter (e.g., Impressionism, Watercolor, Line Art, and Primitivism) achieve high scores on both SA and CL, indicating easier C-S disentanglement when stylistic signals are distributed across the entire image. In contrast, styles associated with religious, mythological, or historical subjects (e.g., Classicism, Rococo, and Baroque) exhibit lower CL scores even when SA remains high. Such themes often involve complex figures, narratives, or symbols that are hard to enumerate in c, making C-S disentanglement more challenging during Ensemble Inversion.

Table A2: Breakdown across the 73 categories on StyleBench. Rich-texture styles yield high SA and CL, while subject-centric styles show lower CL despite strong SA.

Style	SA↑	CL↑	Style	SA↑	CL↑	Style	SA↑	CL↑
3D Model	0.792	0.276	Expressionist	0.855	0.476	Origami	0.815	0.528
3D Model 01	0.729	0.457	Fantasy Art	0.975	0.453	Orphism	0.853	0.604
3D Model 02	0.764	0.511	Fauvism	0.926	0.339	Others	0.799	0.380
3D Model 03	0.703	0.431	Flat Vector	0.585	0.338	Photographic	0.741	0.311
3D Model 04	0.802	0.462	Folk Art	0.498	0.320	Pixel Art	0.674	0.420
3D Model 05	0.726	0.433	Gongbi	0.908	0.473	Pointilism	0.855	0.139
Abstract	0.741	0.496	Graffiti	0.640	0.279	Pop Art	0.671	0.479
Abstract 01	0.845	0.498	Hyperrealism	0.808	0.299	Post-Impressionism	0.972	0.466
Analog Film	0.635	0.382	Icon	0.476	0.605	Precisionism	0.893	0.491
Anime	0.837	0.394	Icon 01	0.517	0.551	Primitivism	0.883	0.661
Anime 01	0.543	0.459	Icon 02	0.551	0.432	Psychedelic	0.434	0.402
Anime 02	0.730	0.422	Impressionism	0.967	0.921	Realism	0.924	0.547
Anime 03	0.866	0.546	Ink and Wash Painting	0.864	0.246	Rococo	0.868	0.186
Anime 04	0.823	0.458	IsoMetric	0.708	0.442	Smoke&Light	0.820	0.569
Anime 05	0.907	0.358	Japonism	0.743	0.217	Statue	0.685	0.497
Anime 06	0.878	0.631	Line Art	0.803	0.675	Steampunk	0.665	0.264
Anime 07	0.817	0.389	Low Poly	0.780	0.546	Stick Figure	0.697	0.498
Art Deco	0.605	0.278	Luminism	0.992	0.529	Stickers	0.580	0.388
Baroque	0.901	0.266	Macabre	0.937	0.584	Surrealist	0.890	0.507
Children's Painting	0.755	0.554	MineCraft	0.864	0.459	Symbolism	0.892	0.441
Classicsm	0.882	0.0184	Monochrome	0.774	0.467	Tonalism	0.887	0.492
Constructivism	0.710	0.506	Neo-Figurative Art	0.922	0.541	Typography	0.200	0.440
Craft Clay	0.475	0.508	Neoclassicism	0.813	0.286	Watercolor	0.858	0.667
Cublism	0.930	0.369	Nouveau	0.920	0.395			
Cyberpunk	0.923	0.499	Op Art	0.672	0.538			

G Analysis on Failure Modes

We analyze three representative failure modes of CLeaR. All stem from limitations in current pretrained modules rather than the disentanglement formulation itself.

Incomplete Content Description In theory, an ideal c allows OSP to remove all content-related components. However, as discussed in Appendix F, certain styles are difficult to describe clearly, leading to residual content leakage.

Degenerate Style Anchor Fig. A3 shows cases where Ensemble Inversion produces a degenerate anchor that either contains almost no information (col 2) or replicates the style reference (cols 1, 3, 4). This failure mode is frequently observed in styles such as Line Art, Stick Figure, and Icon. Although orthogonal projection is designed to remove content from the style target, current VFM embeddings do not always provide a direction that separates style from content. When style is tightly coupled with a specific object, the inversion objective is satisfied either by an anchor without meaningful structure or by one that copies the reference.



Figure A3: Degenerate style anchors. Ensemble Inversion may produce an anchor with no structure or one that copies the reference, especially when content and style are tightly coupled.

Generation Artifacts Fig. A4 shows another failure mode that arises when the style anchor I_a^* is used as the image condition for diffusion generation, where the output contains noisy lines and blobs. As discussed in Sec. 3.3, the inverted anchor lies outside the adapter's training distribution, so the adapter may misinterpret noise in the anchor as semantic content, forming artifacts in the final result.



Figure A4: Artifacts in the generation stage. The adapter may misinterpret inversion noise in style anchor as semantic content, producing noisy lines and blobs.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: It is well illustrated

   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: Please refer to the experiment part where we state that our method may influence the aesthetic score.

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: [Yes]

   Justification: Each assumption is listed and well shown.

   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: Code is in the Supp.

   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: [Yes]

   Justification: In the Supp

   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: Setting is well illustrated

   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: [Yes]

   Justification: We average teh results of multiple seeds.

   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: In the code.

   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: Of course.

   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: [N/A]

    Justification: Technical Paper with no social impact

    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: No risk

    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: Well Attributed

    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: [N/A]

    Justification: [N/A]

    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: [N/A]

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: [N/A]

    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:[N/A]

    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.