Learning to Stylize by Learning to Destylize: A Scalable Paradigm for Supervised Style Transfer

Ye Wang[1], Zili Yi[2], Yibo Zhang[1,3], Peng Zheng[1,3], Xuping Xie[1], Jiang Lin[2], Yijun Li[4], Yilin Wang[4,*†], and Rui Ma[1,5,*]

[1] Jilin University
[2] Nanjing University
[3] Shanghai Innovation Institute
[4] Adobe
[5] Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, MOE, China

Fig. 1: Diverse stylization results of our method. Our framework can generate high-fidelity stylization results across a wide range of artistic styles at 1K resolution.

Abstract. This paper introduces a scalable paradigm for supervised style transfer by inverting the problem: instead of learning to stylize directly, we learn to destylize, reducing stylistic elements from artistic images to recover their natural counterparts and thereby producing authentic, pixel-aligned training pairs at scale. To realize this paradigm, we propose DeStylePipe, a progressive, multi-stage destylization framework that begins with global general destylization, advances to category-wise instruction adaptation, and ultimately deploys specialized model adaptation for complex styles that prompt engineering alone cannot handle. Tightly integrated into this pipeline, DestyleCoT-Filter employs Chain-of-Thought reasoning to assess content preservation and style removal at each stage, routing challenging samples forward while discarding persistently low-quality pairs. Built on this framework, we construct

* Corresponding authors.
† Project lead.

DeStyle-350K, a large-scale dataset aligning diverse artistic styles with their underlying content. We further introduce BCS-Bench, a benchmark featuring balanced content generality and style diversity for systematic evaluation. Extensive experiments demonstrate that models trained on DeStyle-350K achieve superior stylization quality, validating destylization as a reliable and scalable supervision paradigm for style transfer. Our project page: https://wangyephd.github.io/projects/DeStyle/index.html

Keywords: Style Transfer · Destylization

1 Introduction

Style transfer [6], which aims to render an image in a specific artistic style while preserving its underlying semantic content, has witnessed remarkable progress, evolving from early optimization methods [6,7,11] to highly expressive diffusion-based solutions [10,19,26,27,33]. However, it remains fundamentally ill-posed due to the absence of definitive "ground-truth" stylization for a given content-style pair. Most prior works attempt to circumvent this from a model-centric perspective [2, 4–7, 11, 16, 18, 19, 24, 28, 40, 41]. Yet, without explicit supervision, they often suffer from inaccurate style representation and uncontrollable optimization. OmniStyle [29] recently pioneered a data-centric approach by synthesizing a large-scale paired dataset. However, because its targets are generated by existing style transfer models, it inevitably provides pseudo-supervision, yielding unauthentic approximations that fail to guarantee consistent stylization. Representative results are shown in Fig. 1.

In this paper, we propose a fundamentally different path toward supervised style transfer: **learning to stylize by learning to destylize**. Instead of synthesizing stylized images from scratch, we reverse the process by automatically reducing stylistic elements to extract structure-aligned natural content from diverse artistic inputs. This paradigm establishes a reverse formulation: original, unaltered artistic images directly serve as authentic ground-truth targets, while the destylized images act solely as content inputs. By doing so, the supervision quality is strictly anchored to high-quality original art. Furthermore, any minor imperfections in the destylized content naturally serve as robust data augmentation, improving model generalization without compromising the target supervision. To scale this paradigm, we propose **DeStylePipe**, a progressive multi-stage framework that robustly extracts structure-aligned natural content through three stages: global general destylization, category-wise instruction adaptation, and specialized model adaptation. To ensure stringent data quality, we integrate **DestyleCoT-Filter**, an interpretable Chain-of-Thought evaluation mechanism. Crucially, by operating on destylized natural images rather than complex stylized outputs, this filter fundamentally aligns with the primary training distribution of MLLMs, ensuring robust and reliable quality control. Built on this framework, we construct **DeStyle-350K**, a large-scale dataset comprising 350K high-quality triplets ⟨ **de-stylized image**, **reference image**, **style**

image ⟩[6] across over 500 styles and 100 semantic classes. Finally, we present **BCS-Bench**, a benchmark with balanced content generality and stylistic diversity for systematic evaluation of style transfer methods. Our main contributions are summarized as follows:

- We introduce a novel paradigm that reframes style transfer as a **supervised** learning problem via **destylization**. This reverse formulation enables unaltered style image to serve as direct learning targets, providing authentic supervision signals and effectively addressing the long-standing "ground-truth" absence problem.
- We propose **DeStylePipe**, a progressive multi-stage destylization framework, tightly integrated with **DestyleCoT-Filter**, an interpretable evaluation mechanism that accurately assesses content preservation and style removal to guarantee stringent data quality.
- We construct **DeStyle-350K**, which, to the best of our knowledge, is the largest-scale high-quality triplet dataset offering authentic supervision for style transfer. Comprising 350K high-quality triplets, it effectively overcomes the unreliability inherent in prior pseudo-target datasets.
- We present **BCS-Bench**, a comprehensive benchmark featuring balanced content generality and stylistic diversity . Utilizing this benchmark, we validate the high efficacy of DeStyle-350K by fine-tuning multiple models (e.g., Flux.2-Klein-9B/4B, Qwen-Image-Edit). The consistently superior results across different architectures firmly demonstrate that our dataset provides reliable and highly effective supervision for style transfer.

2 Related Work

Style Transfer. Style transfer has evolved from early handcrafted filters [25,38] and optimization-based feature matching [6,7,11] to efficient feed-forward models [3,9,14,15,39]. Recently, diffusion-based methods [2,26,33,34] have shown immense promise through both tuning-based [30,41,42] and tuning-free [10,17,27] strategies. However, the persistent absence of definitive targets often forces these methods to rely on noisy learning signals, such as handcrafted statistical metrics or synthetic pseudo-supervision [29]. To address this fundamental limitation, we propose a destylization paradigm that extracts structure-aligned natural content directly from artistic images, thereby constructing grounded supervision pairs. Based on this, we introduce DeStyle-350K, providing authentic ground truth to reliably train and evaluate state-of-the-art style transfer models.

Datasets for Style Transfer. Early style-related datasets including WikiArt [21] and Style30K [13], provide abundant artistic exemplars but lack the aligned triplets necessary for supervised training. While recent efforts have introduced synthetic triplet datasets [29,33], they are inherently bottlenecked by the biases of the style transfer models used to create them. Furthermore, forcing MLLMs to evaluate these stylized domains often yields unreliable filtering, leading to noisy

[6] green: input; blue: "ground-truth"

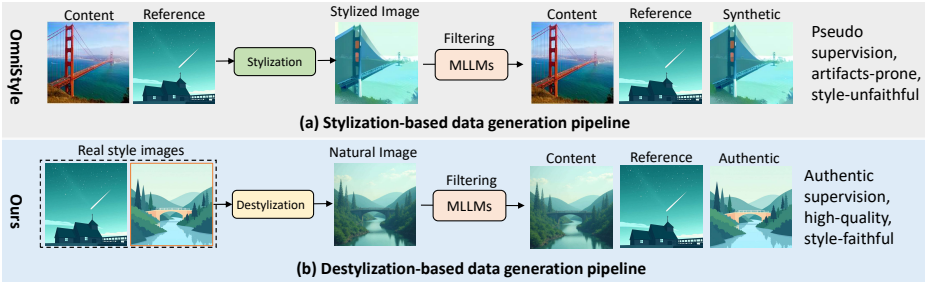

Fig. 2: Stylization-based vs. destylization-based data generation pipelines.

supervision and style drift. Our destylization-based pipeline circumvents this by reversing the stylization process to recover natural content, which aligns seamlessly with the primary training distribution of MLLMs. This approach enables robust, interpretable filtering, ultimately yielding a high-quality triplet dataset with accurate content alignment and authentic style supervision.

3 DeStyle-350K Construction

3.1 Motivation: From Stylization to Destylization

A fundamental challenge in artistic style transfer is the absence of ground-truth pairs consisting of natural images and corresponding high-quality artistic counterparts. As illustrated in Fig. 2(a), previous paradigms, such as OmniStyle [29], primarily rely on a stylization-based data pipeline. These methods utilize existing style transfer models to apply artistic textures to natural content images, creating "pseudo-triplets." However, this approach inherently suffers from a supervision bottleneck: the training signals are bounded by the limitations of the surrogate models, often inheriting artifacts, content leakage, and a significant distribution shift from real-world art. Consequently, models trained on such data struggle to capture the authentic essence of diverse artistic styles.

To break this limitation, we propose a paradigm shift: **learning to stylize by learning to destylize**, as shown in Fig. 2(b). Instead of synthesizing pseudo-artworks, we start from authentic, high-quality style images, including classical masterpieces and high-fidelity synthetic art, and aim to recover their underlying natural content. Formally, this "destylization" process can be defined as $I_{natural} = \mathcal{F}_{desty}(I_{style} \mid T)$, where $\mathcal{F}_{desty}$ denotes our progressive destylization pipeline and T represents the textual instructions that guide the model to reduce stylistic elements while preserving content structures. By systematically reversing the stylistic process, our framework provides **authentic style supervision** because the training target is the **unmodified style image**, be it a classical masterpiece or a high-fidelity synthetic artwork characterized by **integral artistic distribution**. Unlike stylization-based pipelines that rely on algorithmic approximations of a transfer process, our target images possess the coherent textures and global stylistic consistency of original art. Thus, the model learns from **primary stylistic distributions** rather than degraded, transformation-induced representations.

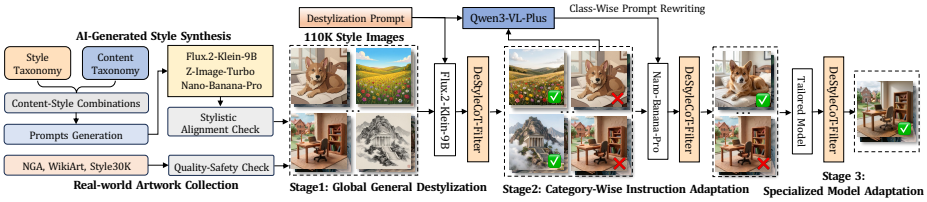

Fig. 3: Overview of DeStylePipe framework. Starting from multi-source artistic image collection, a three-stage progressive destylization pipeline is applied, with DestyleCoT-Filter after each stage to assess quality and route failed cases forward.

3.2 Multi-source Artistic Image Collection

Real-world Artwork Collection. We aggregate an extensive corpus of artistic images from diverse digital galleries (see Fig.3), including the National Gallery of Art (NGA), WikiArt [21], and Style30K [13]. To ensure quality and safety, we filter the collection for inappropriate content and unethical risks, followed by a resolution-based screening that discards samples below 1024×1024 pixels. This process yields a high-quality subset of 50K artworks spanning primary stylistic distributions across classical and modern human-crafted art.

AI-Generated Style Synthesis. To expand stylistic diversity, we develop an automated generation pipeline based on our hierarchical taxonomy (see Fig.3). First, we pair the 6 major artistic categories (abstract art, cultural regional, digital style, illustration and comic, material crafts, and traditional fine arts) with the 7 primary content categories (animals, architecture, human, interior, landscape, plant, and product and object) to create an exhaustive set of content-style combinations. Qwen3 [35] then expands these pairs into descriptive prompts to drive multiple T2I models, including Flux.2-Klein-9B [12], Z-Image-Turbo [23], and Nano-Banana-Pro [22], synthesizing 90K raw artistic images. Finally, to guarantee that each image strictly adheres to its intended style, we employ Qwen3-VL-Plus [1] for a rigorous stylistic alignment check. Only samples that faithfully represent the target style are retained, resulting in approximately 60K high-fidelity images for subsequent destylization. A detailed taxonomy including all sub-categories is provided in the Supplementary Material.

3.3 DeStylePipe: Progressive Multi-Stage Destylization Framework

Given the curated 110K artistic images, we introduce DeStylePipe, a progressive multi-stage framework designed to perform systematic destylization and recover the underlying natural content. The overall architecture is illustrated in Fig. 3 .

Stage 1: Global General Destylization. In the first stage, all 110K stylistic images undergo a foundational destylization. We leverage Flux.2-Klein-9B [12] to perform the initial inference, guided by a universal instruction: *"Transform this image into a style-free natural image."* This process efficiently handles samples with low stylistic complexity. Then, each output is evaluated by our DestyleCoT-Filter (Sec. 3.4) based on content preservation and stylistic removal. Samples that

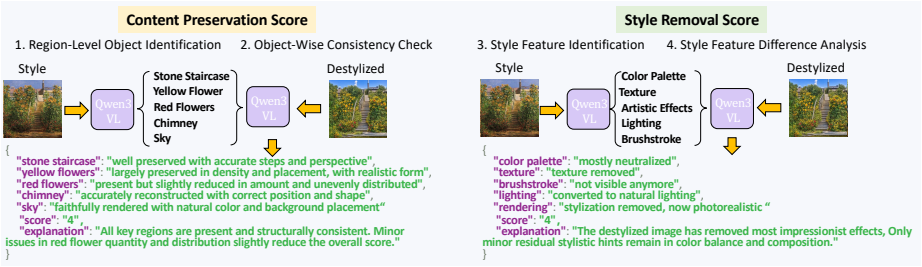

Fig. 4: Overview of DestyleCoT-Filter. A CoT-driven multi-dimensional filtering framework scoring across content preservation and style removal to ensure data quality.

satisfy these dual constraints are directly incorporated into the final dataset, while the remainder proceed to the subsequent stages for further processing.

Stage 2: Category-Wise Instruction Adaptation. For samples failing Stage 1, we implement a more granular destylization strategy. First, Qwen3-VL-Plus [1] is utilized to analyze these failed stylistic images, identifying their specific stylistic attributes to perform category-wise prompt rewriting. Then, we employ Nano-Banana-Pro [22] to execute the destylization with these tailored instructions. These customized prompts, coupled with the more advanced image editing model, enable more precise destylization in complex artistic domains. Finally, the outputs are re-evaluated by the DestyleCoT-Filter; samples satisfying the quality constraints are incorporated into the dataset, while the remaining failed cases proceed to Stage 3 for final processing.

Stage 3: Specialized Model Adaptation. While prompt engineering works for most style domains, certain complex styles that are difficult to describe with text—such as needle felted, clay, or origami—remain hard to process even with tailored instructions. To solve this, we implement a specialized model adaptation strategy for these challenging cases. We fine-tune a customized destylization model via LoRA to specifically learn these complex style removals. The training dataset is constructed by sampling natural images from the HQ-50K [36] collection as ground-truth, and then using Nano-Banana-Pro to synthesize their stylized versions using these persistent styles as references. This process yields 20K structure-aligned pairs for direct supervision. Finally, after a validation by the DestyleCoT-Filter, these qualified samples are integrated in the final dataset.

3.4 DestyleCoT-Filter

To ensure high data quality, we introduce DestyleCoT-Filter, a Chain-of-Thought-based mechanism designed to rigorously evaluate style-destylized image pairs. Unlike existing filters [29] that assess complex artistic domains, our approach evaluates the destylized outputs to better align with the pre-training data distribution of MLLMs. By shifting the assessment from abstract artistic styles to natural realism, we leverage the model's inherent understanding of natural scenes, leading to more stable and objective filtering results. As shown in Fig. 4, the pipeline evaluates two key dimensions through a structured reasoning process: content preservation and style removal.

Content Preservation. Directly prompting MLLMs for consistency scores often overlooks fine-grained semantic drift. To mitigate this, we employ Qwen3-VL-Plus with a structured CoT strategy that decomposes the evaluation into three sequential steps. First, the model performs region-level object identification to extract key semantic components from the source style image, such as "stone staircase" or "chimney." Next, it executes an object-wise consistency check by cross-referencing these objects within the destylized output to verify their presence, placement, and structural integrity. Finally, the model provides a content preservation scoring (0-5) based on these specific assessments. This step-by-step reasoning grounds the final judgment in explicit visual evidence, capturing even minor discrepancies in object distribution.

Style Removal. To quantify destylization degree, we adopt an attribute-based evaluation strategy consisting of two main steps. First, the model performs style feature identification to decompose the original image into distinct style dimensions, such as color palette, texture, brushstrokes, and lighting. Next, it executes a style feature difference analysis to evaluate the "residual style" in the destylized output. During this stage, the model checks if artistic elements have been neutralized or converted to natural standards, such as verifying if textures are removed or brushstrokes are no longer visible. Finally, the model assigns a style difference score (0-5) with a formal explanation. This structured reasoning ensures that the assessment captures specific stylistic remnants across all identified dimensions, leading to an objective measure of successful destylization.

Filtering Criterion. To ensure high data quality, we implement a dual-threshold selection strategy. A sample is admitted to the final dataset only if it achieves a score of ≥ 4 in both content preservation and style removal. This strict rule ensures that every pair is structurally consistent with the source and free of stylistic artifacts. By enforcing these constraints, we provide high-quality data for training robust stylization models.

Triplet Construction. Based on the high-quality samples filtered above, we construct triplets to facilitate stable training and prevent semantic leakage. Each triplet consists of a destylized image (content), a style reference (condition), and a style image (target). For each style target, we retrieve the Top-5 most similar style references from different semantic categories (e.g., matching an "animal" target with "non-animal" references). For AI-generated data, categories are derived from our content taxonomy, while for real-world art, they are identified by Qwen3-VL-Plus. If fewer than five candidates exist, we retain the maximum available. This cross-semantic matching ensures the model decouples style from category-specific priors, significantly enhancing training robustness. This rigorous triplet construction process yields the DeStyle-350K dataset.

Destylization Evaluation. As shown in the upper section of Table 1, we first apply the proposed DeStyleCoT-Filter to evaluate the entire DeStyle-350K dataset along two key dimensions. Across this large-scale dataset, DeStylePipe achieves an average style removal score of 4.57 and a content preservation score of 4.67 on a 5-point scale. These results indicate that our pipeline effectively removes artistic attributes while preserving structural consistency. Furthermore,

Table 1: Evaluation of destylization quality. All scores are averaged on a 0-5 scale (higher is better).

Evaluation Setting	Criterion	Scope	Avg. Score
DeStyleCoT-Filter	Style Removal	DeStyle-350K	4.57
	Content Preservation	DeStyle-350K	4.64
Expert User Study	Realism	1,000 Triplets	4.52
	Content Preservation	1,000 Triplets	4.71
	Style Removal	1,000 Triplets	4.48

as shown in the lower section of Table 1, we further conduct a comprehensive user study involving 10 experts with artistic backgrounds. The experts evaluate 1,000 randomly sampled triplets, each consisting of a de-stylized output, the reference content image, and the corresponding style image. For each triplet, scores ranging from 0 to 5 are assigned across three criteria: (1) *Realism*, (2) *Structural Preservation*, and (3) *Style Removal*. Although the absolute scores differ, the expert evaluation exhibits a similar performance trend, further corroborating the effectiveness of DeStylePipe.

3.5 Dataset Statistics

Overview. DeStyle-350K is a large-scale dataset consisting of 350K high-quality training triplets. This dataset comprises 150K samples derived from real-world artistic images and 200K samples from AI-generated stylistic images. As illustrated in Fig. 5, the dataset provides extensive coverage across 6 major artistic categories and 7 primary content categories. Specifically, DeStyle-350K encompasses over 500 fine-grained stylistic variations (e.g., *Ukiyo-e*, *Cyberpunk*, and *Needle Felted*) and 100 content classes, ensuring both stylistic diversity and content richness. More detailed information is provided in the Appendix.

Comparative Analysis. We compare DeStyle-350K with OmniStyle150K, the current largest open-source style transfer dataset, as summarized in Table 2. Our dataset provides a substantial expansion in scale (350K vs. 150K) and introduces a hybrid data source strategy by integrating real-world artistic images with synthetic ones. In terms of diversity, DeStyle-350K covers over 100 content categories and 30K style references, offering significantly finer granularity compared to the 20 categories and 1,000 style references in OmniStyle150K. Quantitative metrics for style consistency, content preservation, and visual quality are obtained by averaging Qwen3-VL-Plus scores across the entire dataset. These metrics are measured on a scale of 0-5 (higher is better). DeStyle-350K consistently achieves higher scores in all dimensions, particularly in style consistency score (4.54 vs. 3.78), indicating that target images synthesized by conventional style transfer models often suffer from style artifacts. Further details regarding the scoring prompts and evaluation protocols are provided in the Appendix.

4 Experiments

4.1 BCS-Bench and Experimental Setup

BCS-Bench. As summarized in Table 3, prior benchmarks primarily focus on style diversity while overlooking content variation: most use small, fixed con-

Fig. 5: Representative samples of DeStyle-350K. Our dataset spans six major style categories with over 500 fine-grained subcategories. Each triplet shows (left to right) the destylized image, reference image, and style image.

Table 2: Comparison between OmniStyle150K and our DeStyle-350K. Our dataset provides significantly larger scale, superior diversity, and enhanced quality.

Dataset	Size	Content Cat.	Style Ref.	Supervision	Style Cons. ↑	Cont. Pres. ↑	Vis. Qual. ↑
OmniStyle150K	150K	20	1K ref	Synthetic	3.78	4.46	4.24
DeStyle-350K	**350K**	**100+**	**30K ref**	**Real**	**4.54**	**4.67**	**4.71**

tent sets (e.g., 20 images) with limited semantic breadth and demographic balance, and some employ stylized or non-natural images, weakening their value as structural references. To address this gap, we introduce BCS-Bench (Fig. 6), a curated benchmark that balances content generality with stylistic diversity. It comprises 81 style images spanning 81 artistic styles (from 2D pixel art/sketch to 3D origami/clay art) paired with 60 content images across 7 major categories (animals, architecture, human, interior, landscape, plant, and product). Taking Human as an example (Fig. 6, upper-left), we stratify by age, skin tone, body shape, disability, and group composition (single vs. multi-person) for fair, diversity-aware evaluation. These pairings yield 4,860 unique combinations at 1024×1024, enabling comprehensive quantitative and qualitative evaluation.

Baselines. We evaluate our approach against two groups of methods: (1) Style Transfer Models: OmniStyle [29], USO [32], Qwen-Style [37], Attention Distillation (AD) [44], StyleID [2], CSGO [33], StyleShot [10]. (2) Image Editing Models: Closed- and open-source models including Nano-Banana-Pro, GPT-Image-1.5, Qwen-Image-Edit [31], FLUX.2-Klein9B [12], FLUX.2-Klein4B [12].

Evaluation Metrics. Our model is evaluated across three dimensions: content preservation (DINO similarity, CLIP text-image alignment), style similarity (CSD score [20], style loss), and VLM-based assessment. For the choice of VLM, we choose Qwen3-VL-Plus [1] to score four aspects from 0 to 5: Qwen-C (content preservation), Qwen-S (style similarity), Qwen-A (aesthetic quality), and Qwen-L (style semantic leakage, lower is better). Qwen-L measures how much semantic content from the style reference erroneously appears in the stylized output; lower indicates better style-content disentanglement.

Implementation Details. We train three baseline models on the DeStyle-350K dataset: Flux.2-Klein-9B, Flux.2-Klein-4B, and Qwen-Image-Edit-2511. For the Flux-based models, we perform full fine-tuning with a learning rate of 1e-5 and a

Table 3: Comparison of existing style transfer benchmarks and our proposed BCS-Bench. "N/A" denotes missing information.

Benchmark	Content Images	Content Categories	Content Ethics	Style Categories	Content-Style Pairs	Resolution
CAST [43]	N/A	N/A	N/A	N/A	50	N/A
AesPANet [8]	N/A	N/A	N/A	N/A	65	256×256
InST [42]	N/A	N/A	N/A	N/A	26	N/A
StyleID [2]	20	4	No	Only Oil paintings	800	512×512
StyleShot [10]	20	6	No	73	9,800	879×876
OmniStyle [29]	20	4	No	32	2,000	1024×1024
BCS-Bench (Ours)	**60**	**7**	Yes	**81**	**4,860**	**1024×1024**

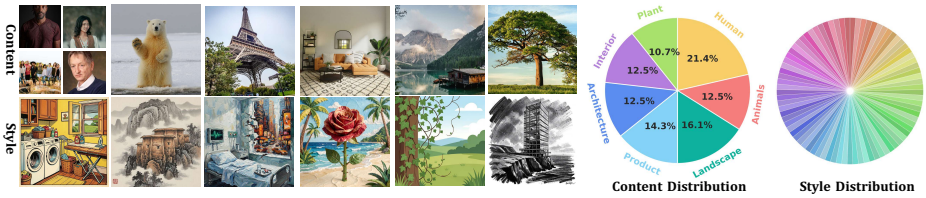

Fig. 6: Content and style examples (left) and distributions (right) of BCS-Bench. For clarity, 81 style categories information are omitted.

batch size of 32. For Qwen-Image-Edit-2511, we use LoRA fine-tuning (lr=1e-4, batch size 8) due to its large parameter count. All experiments are conducted on a cluster of 8×H20 GPUs.

4.2 Validation of Destylization Quality

Qualitative Destylization Results. As illustrated in Fig. 7, our DeStylePipe globally processes diverse artistic inputs and successfully translates them into highly realistic natural images. It effectively removes stylistic artifacts across various domains, including but not limited to dense oil paintings, flat 2D illustrations, and 3D papercrafts, yielding convincing photorealistic appearances. Beyond this authentic global destylization, our method excels in preserving intricate local structures. As highlighted by the red bounding boxes, DeStylePipe maintains near pixel-perfect alignment for complex local geometries, flawlessly retaining legible textual elements like the "FLAVOR OASIS" sign, highly structured objects such as the typewriter keyboard, and fine background semantics like the wall-mounted picture frame. This precise structural preservation, coupled with effective style elimination, ensures that our extracted natural images provide pixel-aligned content inputs for style transfer while seamlessly aligning with the real-world data distribution. Fig. 8 further illustrates how DeStylePipe handles challenging styles through progressive refinement. For Clay Style and Needle Felted Style, Stage 1 yields unsatisfactory results with residual artifacts, while later stages progressively remove stylistic elements and produce photorealistic outputs. In the Needle Felted example, Stage 3 is required to fully recover the natural scene, showing that difficult styles benefit from the multi-stage pipeline.

Fig. 7: Our destylization results. Red bounding boxes highlight the preservation of fine-grained details and spatial structures when destylization, including text, keyboard keys, and intricate textures, without geometric distortion.

Fig. 8: Destylization performance across stages in challenging styles.

4.3 Validation of Stylization Performance

Unless otherwise noted, all experimental evaluations and visualizations are based on the fine-tuned Flux.2-Klein-9B model.

Qualitative Comparison with Style Transfer Models. As illustrated in Fig. 9, our method significantly outperforms state-of-the-art baselines in both stylistic expression and structural faithfulness. Regarding stylistic expression, baselines such as StyleShot and CSGO (Rows 1 and 3) suffer from "content leakage", where semantic elements from the style reference erroneously appear in the output, while USO (Rows 2 and 4) exhibits insufficient stylization for complex artistic style. In contrast, our model achieves a profound transformation, accurately capturing elements like thick brushstrokes and abstract geometric blocks. In terms of content preservation, our method effectively avoids the "structural collapse" or "content hallucinations" prevalent in Qwen-Style (Row 2 and 5), which often leads to distorted facial identities and inconsistent global scale. Notably, our approach maintains precise local details even in complex scenes (Rows 4 and 5); for instance, in Row 5, our result preserves consistent license plate information while all other methods fail. These advantages underscore the value of the DeStyle-350K dataset, whose high-fidelity, pixel-aligned nature provides the precise supervision necessary to overcome the traditional trade-off between style strength and content integrity.

Qualitative Comparison with Image Editing Models. As shown in Fig. 10, we first examine representative open-source models, including Qwen-Image-Edit and the FLUX2 series, under complex style conditions. As shown in Rows 1, 2, 4, and 5, these models exhibit recurring failure modes. Qwen-Image-Edit frequently suffers from *semantic leakage*, introducing unintended elements from the style reference. The FLUX2 variants demonstrate either incorrect style transfer (Row

Table 4: Quantitative comparison of style transfer methods across multiple metrics (**best** in bold, second-best underlined).

Method / Metric	DINO ↑	CLIP-TI ↑	CSD ↑	Style Loss ↓	Qwen-C ↑	Qwen-S ↑	Qwen-A ↑	Qwen-L ↓
OmniStyle	0.6579	0.2874	0.4551	0.1047	3.9131	2.8840	4.3089	0.0135
USO	**0.6871**	0.2975	0.3902	0.1106	4.4606	2.8612	4.4882	0.0098
Qwen-Style	0.5877	0.2861	0.5122	0.1152	3.6781	3.4156	4.2265	0.0765
AD	0.6111	0.2824	0.4523	0.1221	3.7790	2.7014	4.0808	0.3494
StyleID	0.6860	0.2901	0.3872	0.1194	3.9272	2.5330	4.2663	0.0067
CSGO	0.6022	0.2885	0.4750	0.1854	3.7027	3.3800	4.3867	0.0208
StyleShot	0.2963	0.2009	0.5308	**0.0833**	2.0560	3.0644	4.4452	0.0473
Ours	0.6441	**0.3034**	**0.5485**	0.0916	**4.5610**	**4.1140**	**4.5269**	**0.0059**

1) or insufficient stylization (Rows 2 and 4), failing to faithfully capture complex stylistic attributes. We further analyze closed-source systems, Nano-Banana-Pro and GPT-Image-1.5. While GPT-Image-1.5 achieves stronger stylization, it often alters the original content structure and disrupts global spatial scale (Rows 1 and 3). Nano-Banana-Pro preserves structure more reliably but still exhibits semantic leakage under complex scenarios. In contrast, our method consistently preserves semantic content and geometric structure while accurately transferring high-order stylistic patterns, even under highly complex style (Row 4).

Fig. 9: Qualitative comparison with other state-of-the-art style transfer methods. Note that our method keep the human skin tone and face identity without stylization distortions.

Quantitative Comparison with Style Transfer Models. As shown in Table 4, our method ranks best on six of eight metrics, leading in both style fidelity (CSD=0.5485, Qwen-S=4.1140) and content preservation (CLIP-TI=0.3034, Qwen-C=4.5610), with the lowest semantic leakage (Qwen-L=0.0059). In contrast, USO and StyleID preserve content well (DINO>0.68) but transfer style poorly (Qwen-S<2.90), whereas StyleShot achieves low Style Loss at the expense of severe content degradation (DINO=0.2963). Therefore, our approach strikes the best balance across all evaluation dimensions.

Quantitative Comparison with Image Editing Models. As shown in Table 5, our method achieves the best scores in CLIP-TI, Qwen-C, and Qwen-L,

Fig. 10: Qualitative comparison with the existing image editing models.

Table 5: Quantitative comparison of image editing methods across multiple metrics (**best** in bold, second-best underlined).

Model / Metric	DINO ↑	CLIP-TI ↑	CSD ↑	Style Loss ↓	Qwen-C ↑	Qwen-S ↑	Qwen-A ↑	Qwen-L ↓
Qwen-Image-Edit	0.5248	0.2337	0.4967	0.1482	3.2562	3.6436	4.5403	0.0088
FLUX.2-Klein9B	0.5191	0.2523	0.6240	0.0562	3.7121	4.0858	4.6108	0.4748
FLUX.2-Klein4B	0.4985	0.2451	**0.6362**	0.0504	4.0437	3.5424	4.6138	0.4380
GPT-Image-1.5	**0.6506**	0.2952	0.5305	0.0892	4.5541	4.4389	**4.7143**	0.0691
Nano-Banana-Pro	0.6047	0.2801	0.5674	**0.0434**	4.5378	**4.4552**	4.6092	0.1185
Ours	0.6441	**0.3034**	0.5485	0.0916	**4.5610**	4.1140	4.5269	**0.0059**

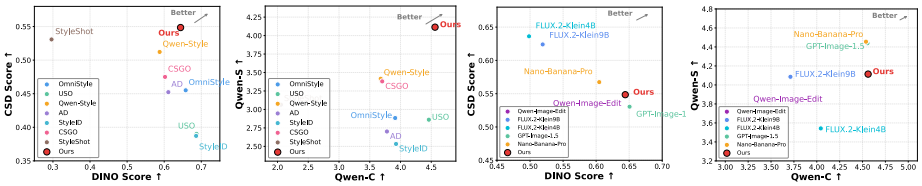

Fig. 11: Trade-off between content preservation and style fidelity across style transfer (left two) and image editing (right two) methods. Upper-right is better.

and ranks second in DINO, indicating strong content preservation and style-content disentanglement. It lags behind general-purpose editing models (e.g., Nano-Banana-Pro, GPT-Image-1.5) on CSD and Qwen-S, as those models are optimized for broader editing tasks, whereas ours is tailored for style transfer.

Content-Style Trade-off Analysis. Fig. 11 shows most methods exhibit a clear trade-off between content preservation and style fidelity, excelling in one dimension often comes at the expense of the other. In the style transfer comparison (left two), our method consistently occupies the upper-right region, achieving the best balance among all competitors. In the image editing comparison (right

Table 6: User study comparison between our method and representative style transfer approaches (**best** in bold, second-best underlined).

Metric / Method	Ours	OmniStyle	USO	Qwen-Style	AD	StyleID	CSGO	StyleShot
Rank 1 (%) ↑	**28.19**	10.82	9.65	13.68	9.17	11.19	8.12	9.18
Top 3 (%) ↑	**56.65**	50.45	24.11	51.27	31.25	33.68	25.58	27.01

Table 7: User study comparison between our method and representative image editing methods (**best** in bold, second-best underlined).

Metrics/Model	Ours	Nano-Banana-Pro	GPT-Image-1.5	Qwen-Image-Edit	FLUX.2-Klein9B	FLUX.2-Klein4B
Rank 1 (%) ↑	22.56	**24.64**	22.92	11.96	12.55	5.37
Top 3 (%) ↑	68.60	**70.12**	67.18	47.53	34.56	12.01

Table 8: Multiple baselines' performance gains from fine-tuning on DeStyle-350K.

Backbone	Setting	DINO ↑	CLIP-TI ↑	CSD ↑	Style Loss ↓	Qwen-C ↑	Qwen-S ↑	Qwen-A ↑	Qwen-L ↓
Flux.2-Klein4B	Before Finetuning	0.4985	0.2451	**0.6362**	**0.0504**	4.0437	**3.5424**	4.4138	0.4380
	After Finetuning	**0.6930**	**0.2987**	0.4999	0.1085	**4.5286**	3.1937	**4.4693**	**0.0025**
Flux.2-Klein9B	Before Finetuning	0.5191	0.2523	**0.6240**	**0.0562**	3.7121	4.0858	4.4108	0.4748
	After Finetuning	**0.6441**	**0.3034**	0.5485	0.0916	**4.5610**	**4.1140**	**4.5269**	**0.0059**
Qwen-Image-Edit	Before Finetuning	0.5248	0.2337	0.4967	0.1482	3.2562	3.6436	**4.5403**	0.0088
	After Finetuning (LoRA)	**0.6365**	**0.3059**	**0.5248**	**0.1105**	**4.5013**	**3.8954**	4.4982	**0.0054**

Fig. 12: Robustness of stylization in challenging scenarios. Red bounding boxes and zoomed-in patches highlight precise structure preservation.

two), our method attains competitive performance close to Nano-Banana-Pro and GPT-Image-1.5, despite being a style transfer model. These results show that DeStyle-350K enables well-disentangled content-style representations.

User Study. We conducted a user study to assess the quality of stylization results. Participants were shown outputs from all methods and asked to rank their top three favorites based on: (1) style preservation, how well the style of the reference image is reflected; (2) content preservation, the degree to which structural details of the content image are retained; and (3) aesthetic appeal, overall visual quality. To reduce bias, image order was randomized and zooming was enabled. We collected 1,620 votes from 30 participants. As shown in Table 6 and Table 7, we report both Rank-1 proportions and Top-3 selection rates. Results show a clear preference for our method: it outperforms existing style transfer methods and achieves performance close to the SOTA closed-source image editing model.

Further Analysis. *(1) Stylization in Complex Scenarios.* As illustrated in Fig. 12, our model demonstrates exceptional robustness in various challenging scenarios. In Outdoor-Complex Backgrounds, it anchors sharp architectural lines and individual silhouettes despite dense crowds. For Indoor-Object Occlusions, the model successfully maintains the original spatial occlusion relationships, ensuring distinct contours for partially hidden entities. Finally, in High-Frequency Details-Patterns & Text, it preserves strict regularity and stroke clarity without geometric warping. These results further demonstrate the value of our DeStyle-

350K dataset in ensuring structural integrity during stylization. *(2) Effectiveness of DeStyle-350K.* As shown in Table 8, fine-tuning three diverse backbones on DeStyle-350K yields consistent improvements across all metrics. Most notably, Qwen-L drops dramatically (e.g., 0.4380→0.0025 on Flux.2-Klein-4B), indicating that models learn to transfer style without leaking semantic content. The consistent gains across architectures of different scales and families confirm that DeStyle-350K provides universally effective and backbone-agnostic supervision.

5 Conclusion

This paper introduces a scalable paradigm for supervised style transfer built on destylization. By reducing stylistic elements from artistic images to recover their natural counterparts, we enable unaltered artworks to serve as authentic ground-truth targets, effectively addressing the long-standing absence of reliable supervision. The resulting DeStyle-350K dataset, comprising 350K triplets across over 500 style categories, provides authentic and scalable supervision for training. Extensive evaluations on BCS-Bench across multiple architectures demonstrate that models trained on DeStyle-350K achieve superior stylization quality while maintaining structural integrity. We hope that the destylization paradigm and DeStyle-350K will serve as a solid foundation for future research in style transfer and image editing.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62572212), the Science and Technology Development Plan of Jilin Province (No. 20260203049SF).

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv:2308.12966 **1**(2), 3 (2023)
2. Chung, J., Hyun, S., Heo, J.P.: Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8795–8805 (2024)
3. Deng, Y., Tang, F., Dong, W., Sun, W., Huang, F., Xu, C.: Arbitrary style transfer via multi-adaptation network. In: ACM Int. Conf. Multimedia. pp. 2719–2727 (2020)
4. Frenkel, Y., Vinker, Y., Shamir, A., Cohen-Or, D.: Implicit style-content separation using b-lora. arXiv:2403.14572 (2024)
5. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv:1508.06576 (2015)
6. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Eur. Conf. Comput. Vis. pp. 2414–2423 (2016)

7. Gatys, L.A., Ecker, A.S., Bethge, M., Hertzmann, A., Shechtman, E.: Controlling perceptual factors in neural style transfer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3985–3993 (2017)
8. Hong, K., Jeon, S., Lee, J., Ahn, N., Kim, K., Lee, P., Kim, D., Uh, Y., Byun, H.: Aespa-net: Aesthetic pattern-aware style transfer networks. In: Int. Conf. Comput. Vis. pp. 22758–22767 (2023)
9. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Int. Conf. Comput. Vis. pp. 1501–1510 (2017)
10. Junyao, G., Yanchen, L., Yanan, S., Yinhao, T., Yanhong, Z., Kai, C., Cairong, Z.: Styleshot: A snapshot on any style. arxiv:2407.01414 (2024)
11. Kolkin, N., Salavon, J., Shakhnarovich, G.: Style transfer by relaxed optimal transport and self-similarity. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10051–10060 (2019)
12. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2` (2025)
13. Li, W., Fang, M., Zou, C., Gong, B., Zheng, R., Wang, M., Chen, J., Yang, M.: Styletokenizer: Defining image style by a single instance for controlling diffusion models. arXiv:2409.02543 (2024)
14. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. In: Adv. Neural Inform. Process. Syst. (2017)
15. Liao, J., Yao, Y., Yuan, L., Hua, G., Kang, S.B.: Visual attribute transfer through deep image analogy. arXiv:1705.01088 (2017)
16. Ouyang, Z., Li, Z., Hou, Q.: K-lora: Unlocking training-free fusion of any subject and style loras. In: IEEE Conf. Comput. Vis. Pattern Recog. (2025)
17. Qi, T., Fang, S., Wu, Y., Xie, H., Liu, J., Chen, L., He, Q., Zhang, Y.: Deadiff: An efficient stylization diffusion model with disentangled representations. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8693–8702 (2024)
18. Shah, V., Ruiz, N., Cole, F., Lu, E., Lazebnik, S., Li, Y., Jampani, V.: Ziplora: Any subject in any style by effectively merging loras. In: arXiv preprint arxiv:2311.13600 (2023)
19. Sohn, K., Ruiz, N., Lee, K., Chin, D.C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., et al.: Styledrop: Text-to-image generation in any style. arXiv:2306.00983 (2023)
20. Somepalli, G., Gupta, A., Gupta, K., Palta, S., Goldblum, M., Geiping, J., Shrivastava, A., Goldstein, T.: Measuring style similarity in diffusion models. arXiv preprint arXiv:2404.01292 (2024)
21. Tan, W.R., Chan, C.S., Aguirre, H., Tanaka, K.: Improved artgan for conditional synthesis of natural image and artwork. IEEE Transactions on Image Processing **28**(1), 394–409 (2019)
22. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
23. Team, Z.I.: Z-image: An efficient image generation foundation model with single-stream diffusion transformer. arXiv preprint arXiv:2511.22699 (2025)
24. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: $p+$: Extended textual conditioning in text-to-image generation. arXiv:2303.09522 (2023)
25. Wang, B., Wang, W., Yang, H., Sun, J.: Efficient example-based painting and synthesis of 2d directional texture. IEEE Trans. Vis. Comput. Graph. **10**(3), 266–277 (2004)
26. Wang, H., Wang, Q., Bai, X., Qin, Z., Chen, A.: Instantstyle: Free lunch towards style-preserving in text-to-image generation. arXiv:2404.02733 (2024)

27. Wang, H., Xing, P., Huang, R., Ai, H., Wang, Q., Bai, X.: Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. arXiv:2407.00788 (2024)
28. Wang, Y., Bai, T., Xie, X., Yi, Z., Wang, Y., Ma, R.: Sigstyle: Signature style transfer via personalized text-to-image models. Proceedings of the AAAI Conference on Artificial Intelligence **39**(8), 8051–8059 (2025)
29. Wang, Y., Liu, R., Lin, J., Liu, F., Yi, Z., Wang, Y., Ma, R.: Omnistyle: Filtering high quality style transfer data at scale. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7847–7856 (2025)
30. Wang, Z., Zhao, L., Xing, W.: Stylediffusion: Controllable disentangled style transfer via diffusion models. In: Int. Conf. Comput. Vis. pp. 7677–7689 (2023)
31. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
32. Wu, S., Huang, M., Cheng, Y., Wu, W., Tian, J., Luo, Y., Ding, F., He, Q.: Uso: Unified style and subject-driven generation via disentangled and reward learning. arXiv preprint arXiv:2508.18966 (2025)
33. Xing, P., Wang, H., Sun, Y., Wang, Q., Bai, X., Ai, H., Huang, R., Li, Z.: Csgo: Content-style composition in text-to-image generation. arXiv 2408.16766 (2024)
34. Xu, Y., Wang, Z., Xiao, J., Liu, W., Chen, L.: Freetuner: Any subject in any style with training-free diffusion. arXiv:2405.14201 (2024)
35. Yang, A., Li, A., Yang, B., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
36. Yang, Q., Chen, D., Tan, Z., Liu, Q., Chu, Q., Bao, J., Yuan, L., Hua, G., Yu, N.: Hq-50k: A large-scale, high-quality dataset for image restoration. arXiv preprint arXiv:2306.05390 (2023)
37. Zhang, S., Huang, H., Zhang, C., Li, X.: Qwenstyle: Content-preserving style transfer with qwen-image-edit. arXiv preprint arXiv:2601.06202 (2026)
38. Zhang, W., Cao, C., Chen, S., Liu, J., Tang, X.: Style transfer via image component analysis. IEEE Trans. Multimedia **15**(7), 1594–1601 (2013)
39. Zhang, Y., Li, M., Li, R., Jia, K., Zhang, L.: Exact feature distribution matching for arbitrary style transfer and domain generalization. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8035–8045 (2022)
40. Zhang, Y., Fang, C., Wang, Y., Wang, Z., Lin, Z., Fu, Y., Yang, J.: Multimodal style transfer via graph cuts. In: Int. Conf. Comput. Vis. pp. 5943–5951 (2019)
41. Zhang, Y., Dong, W., Tang, F., Huang, N., Huang, H., Ma, C., Lee, T.Y., Deussen, O., Xu, C.: Prospect: Expanded conditioning for the personalization of attribute-aware image generation. arXiv:2305.16225 (2023)
42. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10146–10156 (2023)
43. Zhang, Y., Tang, F., Dong, W., Huang, H., Ma, C., Lee, T.Y., Xu, C.: Domain enhanced arbitrary image style transfer via contrastive learning. In: SIGGRAPH (2022)
44. Zhou, Y., Gao, X., Chen, Z., Huang, H.: Attention distillation: A unified approach to visual characteristics transfer. arXiv preprint arXiv:2502.20235 (2025)