CCDiT: One-Step Diffusion Transformer with Plug-and-Play Color Correction for Image Super-Resolution

Yitian Ma, Longhao Sun, Peilin Xiong, and Keiji Yanai*

The University of Electro-Communications, Tokyo, Japan
{ma-y,sun-l,xiong-p}@mm.inf.uec.ac.jp
yanai@cs.uec.ac.jp

Abstract. Real-world image super-resolution (SR) aims to recover high-resolution images from low-resolution inputs degraded by unknown blur, noise, compression, and sensor processes. One-step diffusion SR methods reduce sampling cost, but their generated details may still introduce chrominance drift. We address this issue from two complementary directions. First, we propose CCDiT, a one-step Diffusion Transformer SR framework that adapts Qwen-Image-Edit-2509 through LoRA fine-tuning from a bicubic reference image. Second, we introduce Alpha Mask fusion, a lightweight plug-and-play chrominance correction module that preserves restored luminance and adaptively mixes CbCr channels from the SR output and the bicubic reference image. On RealSR, CCDiT obtains the best LPIPS, DISTS, FID, PSNR, and Mean $\Delta E2000$ among the compared methods. On DRealSR, it obtains the best LPIPS, FID, PSNR, and Mean $\Delta E2000$; on DIV2K-Val, it obtains the best LPIPS, PSNR, and Mean $\Delta E2000$. Applied after the default AdaIN color alignment of OSEDiff and TVT, the same Alpha Mask checkpoint further reduces Mean $\Delta E2000$ on RealSR without modifying either backbone. These results support Alpha Mask fusion as a reusable local chrominance refinement module for existing diffusion SR pipelines.

Keywords: Image super-resolution · Diffusion Transformer · Color consistency · One-step diffusion

1 Introduction

Image super-resolution reconstructs high-resolution detail from limited evidence [4,18,26]. Real-world inputs additionally mix unknown blur, noise, compression, and sensor artifacts, so pixel fidelity alone cannot ensure natural restoration. Diffusion priors synthesize realistic texture [11,19,24,26,29,31] beyond many CNN/GAN approaches [18,32,33,41], but iterative denoising is costly. OSEDiff, TSD-SR, and ODTSR reduce inference to one step [7,9,37]; their generated detail can nevertheless shift input chrominance.

* Corresponding author.

Such drift matters in recognition, food analysis, and other color-sensitive settings [5,21,28]. Prior diffusion SR work notes color shift [20], yet efficient methods emphasize speed, quality, or control rather than a reusable correction module. We therefore propose CCDiT, which LoRA-adapts Qwen-Image-Edit-2509 [36] for one-step SR from a bicubic reference, and Alpha Mask fusion, which locally mixes SR/reference chrominance without changing the backbone.

The main contributions of this paper are summarized as follows:

- We identify chrominance drift as a remaining failure mode in one-step diffusion SR and propose Alpha Mask fusion as a lightweight plug-and-play correction module.
- We instantiate the framework as CCDiT by adapting Qwen-Image-Edit-2509 into a one-step Diffusion Transformer SR backbone through LoRA fine-tuning from a bicubic reference image.
- We evaluate CCDiT on RealSR, DRealSR, and DIV2K-Val, and further apply Alpha Mask fusion to OSEDiff and TVT outputs to verify cross-backbone color-control behavior.

2 Related Work

2.1 Real-World and Diffusion SR

Real-world SR must jointly handle blur, noise, compression, and sensor artifacts that bicubic degradation cannot reproduce [2,32,35]. RealSR and DRealSR therefore use real camera pairs to test restoration beyond synthetic settings [2,35]. Diffusion priors improve texture synthesis [11,24,29]; StableSR, SUPIR, DiT4SR, and DiffBIR exploit them for real-world restoration [8,19,31,41], but iterative sampling is expensive.

2.2 Efficient Diffusion Transformers

Efficient diffusion SR methods use direct prediction, LoRA adaptation, distillation, VAE transfer, or one-step Diffusion Transformers [7,9,30,36,37,40]. Diffusion Transformers provide scalable alternatives to UNets [3,22,25], while LoRA learns low-rank updates to frozen pretrained weights [13].

Our one-step SR formulation is inspired by ODTSR [9]. ODTSR adapts Qwen-Image through a Noise-hybrid Visual Stream and fidelity-aware adversarial training. We instead adapt Qwen-Image-Edit-2509 [36] and retain its pretrained semantic and appearance conditioning pathways. The bicubic reference is jointly encoded with the restoration prompt by Qwen2.5-VL and separately encoded into reference latents by the VAE. Our first-stage adaptation uses latent reconstruction and perceptual supervision. We further introduce Alpha Mask fusion for spatially adaptive chrominance correction while preserving restored luminance.

2.3 Color Consistency and Chrominance Fusion

RGB correction couples brightness, structure, and color, whereas YCbCr separates luminance from chrominance [23]. We therefore preserve restored Y and learn where to select CbCr from the SR output or bicubic reference. Inspired by spatial alpha blending [38], Alpha Mask Net predicts chrominance reliability rather than a foreground matte. It needs only an SR output and its bicubic reference and is thus applicable to different one-step backbones.

3 Method

3.1 Method Overview

CCDiT separates one-step detail restoration from color anchoring. As shown in Figs. 1 and 2, a LoRA-adapted Qwen-Image-Edit-2509 backbone restores structure and texture, while Alpha Mask Net performs YCbCr chrominance fusion and can also post-process other SR backbones.

Given a high-resolution image I_{HR}, we first generate the low-resolution input I_{LR} through an online degradation process $\mathcal{D}(\cdot)$, and then obtain a reference image I_{ref} with the same resolution as the target using bicubic interpolation:

$$I_{LR} = \mathcal{D}(I_{HR}), \qquad I_{ref} = \mathrm{Bicubic}(I_{LR}). \tag{1}$$

Stage 1 maps I_{ref} to an SR output I_q in one denoising step. With that backbone frozen, Stage 2 predicts a spatial map that mixes CbCr from I_q and I_{ref} while retaining the restored luminance, yielding I_f.

Stage 1 restores structure and texture, whereas Stage 2 estimates the reliability of two available chrominance sources. Freezing the backbone fixes the distribution that supplies I_q and makes fusion modular across direct CCDiT, AdaIN-aligned CCDiT, and external SR outputs.

3.2 One-Step Diffusion Transformer Super-Resolution

CCDiT starts from the bicubic latent rather than pure noise, retaining input structure while predicting missing detail. With VAE encoder E, decoder D, multimodal condition c, and the FlowMatch noise level σ_t, the one-step path is

$$\begin{aligned} z_{gt} &= E(I_{HR}), & z_{ref} &= E(I_{ref}), \\ z_t &= (1-\sigma_t)z_{ref} + \sigma_t\epsilon, & \epsilon &\sim \mathcal{N}(0, I), \\ v_\theta &= v_\theta(z_t, z_{ref}, c, t), & z_q &= z_t - \sigma_t v_\theta, \quad I_q = D(z_q). \end{aligned} \tag{2}$$

Following ODTSR [9], we select the discrete scheduler timestep $t = 750$ with `one_step_t_index`=250 in the descending 1000-step schedule. Here, σ_t is the FlowMatch noise level associated with that timestep, and v_θ is the predicted flow velocity. The condition c is obtained by jointly encoding I_{ref} and the BLIP-derived restoration prompt with Qwen2.5-VL; prompt generation is described in Section 4.1.

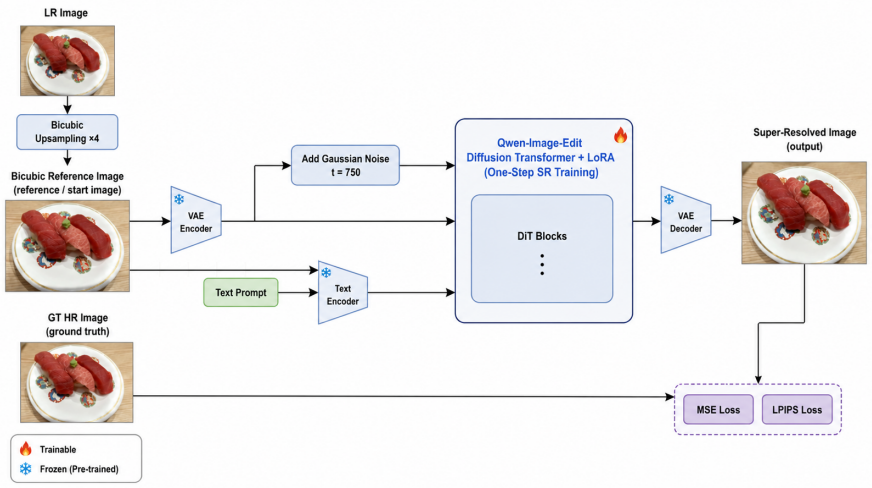


Fig. 1. Stage 1: LoRA adaptation of Qwen-Image-Edit-2509 for one-step SR from a bicubic reference.

Because z_t retains the low-frequency geometry and approximate color of I_{ref}, the one-step update can focus on missing high-frequency content. The clean reference latent z_{ref} supplies the native editing condition: its tokens are concatenated with the noisy restoration tokens inside the Transformer. A noised copy of z_{ref} initializes the restoration state z_t. The predicted velocity moves the current state directly to z_q at zero noise level, requiring one Diffusion Transformer evaluation without intermediate sampling.

We insert LoRA adapters into the query, key, value, and output projections of the attention blocks and into the image/text MLP and modulation projections [13]. The VAE and original transformer weights remain frozen, limiting trainable memory while preserving the broad image prior learned by Qwen-Image-Edit-2509. The adapters consequently learn the restoration-specific velocity field without replacing the generative prior or changing the latent representation used by the pretrained model.

LoRA adapts the pretrained editing backbone without introducing an additional SR-specific control stream.

3.3 First-Stage Training Objective

Stage 1 combines latent reconstruction and AlexNet-LPIPS supervision [42]:

$$\begin{aligned} \mathcal{L}_{stage1} &= \lambda_z \mathcal{L}_{latent} + \lambda_p \mathcal{L}_{perc}, \\ \mathcal{L}_{latent} &= \|z_q - z_{gt}\|_2^2, \qquad \mathcal{L}_{perc} = \mathrm{LPIPS}_{alex}(D(z_q), D(z_{gt})), \end{aligned} \tag{3}$$

with $\lambda_z = \lambda_p = 1.0$.

Latent supervision stabilizes the one-step update before decoding. LPIPS penalizes perceptually important discrepancies in decoded images. Together, they discourage weak texture from latent-only fitting and structural deviation from perceptual-only supervision.

3.4 Plug-and-Play Adaptive Chrominance Fusion

Alpha Mask fusion uses I_{ref} as a conditional color anchor without accessing backbone internals. YCbCr fusion preserves restored luminance and avoids the structural blur caused by direct RGB blending, while a learned map avoids the rigidity of full CbCr replacement [23]. AdaIN and Alpha Mask are complementary: AdaIN performs global color alignment, whereas Alpha Mask locally refines chrominance without altering luminance. We optionally apply AdaIN before Alpha Mask fusion:

$$\begin{aligned} I_q^A &= \mathrm{AdaIN}(I_q, I_{ref}), \\ I_s &= \begin{cases} I_q, & \text{direct CCDiT output,} \\ I_q^A, & \text{with AdaIN pre-alignment.} \end{cases} \end{aligned} \tag{4}$$

Default CCDiT uses $I_s = I_q$. The AdaIN-based CCDiT variants and the official OSEDiff and TVT pipelines use $I_s = I_q^A$ [14,37,40]. Fusion is

$$\begin{aligned} (Y_s, Cb_s, Cr_s) = \mathrm{YCbCr}(I_s), \quad (Y_r, Cb_r, Cr_r) = \mathrm{YCbCr}(I_{ref}), \\ \alpha = M_\phi(F) \in [0,1]^{H\times W}, \quad Y_f = Y_s, \\ Cb_f = \alpha \odot Cb_s + (1-\alpha) \odot Cb_r, \\ Cr_f = \alpha \odot Cr_s + (1-\alpha) \odot Cr_r, \\ I_f = \mathrm{RGB}(Y_f, Cb_f, Cr_f). \end{aligned} \tag{5}$$

Large α trusts SR chrominance; small α anchors it to the reference.

AdaIN and Alpha Mask are complementary. AdaIN aligns global channel means and variances but applies the same transformation over the entire image. It cannot express that a smooth region should inherit reference chrominance while a reconstructed edge should retain SR chrominance. Alpha Mask makes this decision locally between the two available CbCr candidates. Because $Y_f = Y_s$, the second operation preserves the luminance detail of either the direct or AdaIN-aligned SR output. The improvement of AdaIN + Alpha Mask in Table 2 empirically supports this compatibility rather than treating the two operations as substitutes.

The bicubic reference is a color anchor rather than an absolute target, because interpolation and degradation can make it locally unreliable. Fixed CbCr replacement assumes uniform reliability and can suppress useful color detail. The learned coefficient instead transitions continuously between both sources and retains SR chrominance where it better matches the target.

This distinction also explains why the mask is predicted from both candidates rather than only from the reference. A reference-only predictor could detect edges

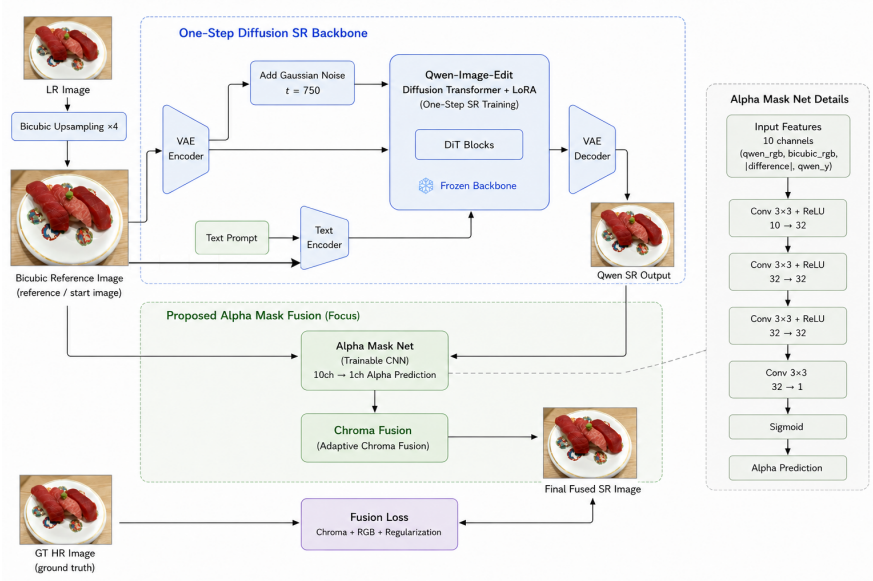


Fig. 2. Stage 2: training Alpha Mask Net with the SR backbone frozen.

or flat regions but could not determine whether the SR output agrees with them. The absolute RGB difference exposes this disagreement directly, while Y_s helps the network avoid treating luminance contrast as a chrominance error. The mask can consequently vary across materials, boundaries, and smooth regions without changing restored luminance.

3.5 Alpha Mask Net

Alpha Mask Net predicts chrominance-source reliability from the candidate images, their disagreement, and SR luminance:

$$F = \mathrm{Concat}(I_s, I_{ref}, |I_s - I_{ref}|, Y_s). \tag{6}$$

The three RGB groups in F describe both candidates and their absolute discrepancy, while Y_s provides a structural cue. The resulting 10-channel tensor passes through four 3×3 convolutions, $10 \to 32 \to 32 \to 32 \to 1$, with ReLU after the first three and Sigmoid after the last. The convolutional stack expands the local receptive field while retaining pixel alignment, and the final Sigmoid constrains the fusion coefficient without test-time clipping. A shallow network is sufficient because it estimates source reliability rather than generating texture or inferring new semantics. Its FP32 parameters occupy 0.08 MiB, and it adds no diffusion step; runtime and memory are reported in Table 5.

3.6 Second-Stage Training Objective

With the SR backbone frozen, Stage 2 optimizes only ϕ using five losses:

$$\mathcal{L}_{mask} = \lambda_c \mathcal{L}_{chroma} + \lambda_r \mathcal{L}_{rgb} + \lambda_{tv} \mathcal{L}_{tv} + \lambda_v \mathcal{L}_{var} + \lambda_o \mathcal{L}_{oracle}. \tag{7}$$

The objective combines direct image supervision with constraints on the spatial mask. $\mathcal{L}_{chroma}$ targets the quantity that the module is designed to change, whereas $\mathcal{L}_{rgb}$ verifies that chrominance correction improves the reconstructed RGB image rather than only an intermediate color space. The remaining terms regulate the geometry of the solution and provide a training-time target for an otherwise underdetermined mixing coefficient. The direct reconstruction and mask regularizers are

$$\begin{aligned} \mathcal{L}_{chroma} &= \|Cb_f - Cb_{gt}\|_1 + \|Cr_f - Cr_{gt}\|_1, \\ \mathcal{L}_{rgb} &= \|I_f - I_{HR}\|_1, \qquad \mathcal{L}_{tv} = \|\nabla_x \alpha\|_1 + \|\nabla_y \alpha\|_1, \\ \mathcal{L}_{var} &= \max(0, \nu - \mathrm{Var}(\alpha)). \end{aligned} \tag{8}$$

$\mathcal{L}_{tv}$ suppresses isolated pixel-to-pixel oscillation that could create chroma speckle around smooth regions. In contrast, $\mathcal{L}_{var}$ prevents the opposite degeneracy: a nearly constant mask that reduces the adaptive module to global blending. The hinge form applies a penalty only when spatial variance falls below ν, so masks that already exhibit sufficient spatial adaptation are not encouraged to become unnecessarily extreme.

Let q_c, r_c, g_c denote CbCr from I_s, I_{ref}, I_{HR}. A least-squares target directly supervises mixing:

$$\alpha_{oracle} = \mathrm{Clamp}\left(\frac{((g_c - r_c) \odot (q_c - r_c))_{\Sigma C}}{((q_c - r_c)^2)_{\Sigma C} + \epsilon}, 0, 1 \right), \qquad \mathcal{L}_{oracle} = \|\alpha - \alpha_{oracle}\|_1, \tag{9}$$

where $(\cdot)_{\Sigma C}$ sums over CbCr. We use $\alpha \in [0, 1]$ and $\nu = 0.020$. The weights are $\lambda_c = 2.0$, $\lambda_r = 0.05$, $\lambda_{tv} = 0.01$, $\lambda_v = 0.02$, and $\lambda_o = 0.10$. Component-wise effects appear in Table 4; fusion behavior appears in Tables 2 and 3.

The oracle coefficient is available only during training because it uses g_c. For every pixel, it projects the target chrominance onto the line segment between the SR and reference candidates and clamps the least-squares solution to the feasible interval. $\mathcal{L}_{oracle}$ therefore teaches the network a source-selection rule without requiring an externally annotated mask. At test time, I_{HR} and α_{oracle} are discarded; M_ϕ predicts α directly from the ten-channel feature tensor. The larger chroma weight reflects the module's specific purpose, while the lower RGB, smoothness, variance, and oracle weights act as complementary constraints rather than competing reconstruction objectives.

3.7 Inference Pipeline

At inference, BLIP captions the LR image, and bicubic upsampling produces I_{ref}. Qwen2.5-VL jointly encodes the reference and prompt into c, while the

VAE supplies z_{ref}. Its clean form conditions restoration, and its noised copy initializes the one-step update. One Transformer evaluation, VAE decode, and mask forward pass produce the output. For an external backbone, Alpha Mask requires only its output and I_{ref}, optionally after AdaIN; no hidden features or backbone modification are needed.

4 Experiments

4.1 Experimental Settings

We evaluate backbone quality, color strategies, cross-model transfer, the five losses, and mask overhead. CCDiT is trained on LSDIR [17] with an online $4\times$ degradation pipeline following OSEDiff and Real-ESRGAN [32,37]. The first stage samples mixed blur kernels, resizing, Gaussian or Poisson noise, and JPEG compression. The optional second stage repeats blur, resizing, and noise, then randomizes JPEG compression and final sinc filtering. Sampling these combinations online exposes the model to a broader degradation distribution than a fixed precomputed LR set. Bicubic upsampling is applied only after the final degradation, producing I_{ref} at the target resolution while retaining the LR image's low-frequency content.

For both training and inference, BLIP captions the LR image on CPU using `Salesforce/blip-image-captioning-large` [16]. We append "high-quality photo restoration, clearer and sharper details, increase resolution, preserve fine textures." The LR-derived caption supplies scene semantics without using the HR target, while the suffix specifies the restoration direction. Running BLIP on CPU leaves accelerator memory available to the Diffusion Transformer and keeps the conditioning procedure identical when ground truth is unavailable at test time.

Stage 1 runs for one epoch on six NVIDIA RTX PRO 6000 Blackwell Max-Q GPUs (about one day), with learning rate 5×10^{-5}, LoRA rank 16, accumulation 3, gradient checkpointing, and checkpoints every 4000 steps. Stage 2 freezes the SR model and trains the 32-channel mask network for one epoch on two GPUs (about one day), with learning rate 2×10^{-4}, accumulation 4, and checkpoints every 10000 steps.

For transfer, OSEDiff and TVT retain the AdaIN alignment enabled by their official pipelines [37,40]. The same Stage-2 checkpoint, trained only on CCDiT outputs, is applied without backbone-specific adaptation or backbone modification. This tests direct compatibility rather than each model's best attainable result; model-specific mask adaptation remains future work.

4.2 Evaluation Metrics

We report PSNR and SSIM [12,34], LPIPS and DISTS [6,42], MUSIQ and MANIQA [15,39], and FID [10]. Output–ground-truth Mean $\Delta E2000$ measures color difference [27]; lower is better. Following standard SR evaluation, Table 1

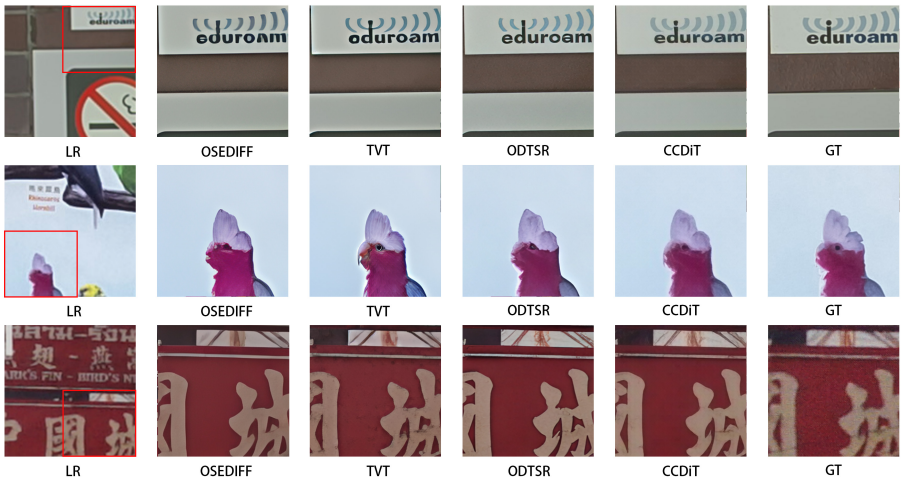


Fig. 3. Qualitative comparison on RealSR and DRealSR.

uses Y-channel PSNR/SSIM, whereas the three color-focused tables use RGB values. These forms are not directly comparable, while $\Delta E2000$ is defined identically in all four quality-comparison tables.

4.3 Comparison with Existing Methods

Table 1 tests how CCDiT performs with one-step inference. We compare CCDiT with one-step methods and multi-step diffusion restoration methods on RealSR, DRealSR, and DIV2K-Val [1].

With one step, CCDiT gives the best LPIPS, PSNR, and Mean $\Delta E2000$ on all three datasets; it also leads DISTS and FID on RealSR and FID on DRealSR. ODTSR remains better on several DIV2K perceptual/distribution metrics, so no method dominates every criterion. CCDiT nevertheless compares favorably with both one-step competitors and 40–50-step SUPIR/DiT4SR. On RealSR, CCDiT improves PSNR-Y over TVT by 0.49 dB and reduces Mean $\Delta E2000$ by 0.1545.

On DRealSR, CCDiT achieves LPIPS 0.2290, FID 104.45, PSNR-Y 28.66, and Mean $\Delta E2000$ 4.4190, while TVT retains the highest SSIM-Y. On DIV2K-Val, CCDiT ranks second in FID and SSIM-Y; ODTSR leads DISTS, MANIQA, and FID, whereas TSD-SR achieves the highest MUSIQ. These differences indicate dataset-dependent trade-offs between reconstruction, color fidelity, and perceptual quality.

4.4 Color Fusion and Loss Ablations with Cross-Model Generalization

Tables 2 and 3 evaluate color correction on RealSR. In Table 2, all correction variants start from CCDiT without Alpha Mask; default CCDiT includes the

Table 1. Comparison on three benchmarks. Bold/underline: best/second best.

Dataset	Method	Step(s)	LPIPS↓	DISTS↓	MUSIQ↑	MANIQA↑	FID↓	PSNR↑	SSIM↑	Mean $\Delta E2000$ ↓
RealSR	OSEDiff	1	0.2921	0.2128	69.09	0.6326	123.49	25.15	0.7341	5.1556
	PiSA-SR	1	0.2672	0.2044	70.15	0.6560	124.09	25.50	0.7417	4.9202
	TSD-SR	1	0.2743	0.2104	**71.19**	0.6347	114.45	24.81	0.7172	5.2863
	TVT	1	0.2597	0.2061	69.89	0.6232	109.90	25.81	**0.7596**	4.7171
	SUPIR (CFG=4.0)	50	0.3541	0.2488	62.09	0.5780	130.38	23.65	0.6620	5.2989
	DiT4SR (CFG=8.0)	40	0.3215	0.2251	67.76	0.6564	118.55	23.50	0.6657	6.0262
	ODTSR (f=1.0)	1	0.2398	0.1894	68.29	**0.6622**	101.49	25.07	0.7361	4.9832
	CCDiT	1	**0.1918**	**0.1768**	64.98	0.6305	**79.69**	**26.30**	0.7553	**4.5626**
DRealSR	OSEDiff	1	0.2968	0.2165	64.65	0.5895	135.29	27.92	0.7835	4.8380
	PiSA-SR	1	0.2960	0.2169	66.11	0.6146	130.48	28.32	0.7804	4.6880
	TSD-SR	1	0.2966	0.2136	**66.61**	0.5850	135.32	27.77	0.7559	4.9653
	TVT	1	0.2899	0.2205	65.56	0.5776	134.28	28.27	**0.7899**	4.6469
	SUPIR (CFG=4.0)	50	0.4243	0.2795	58.79	0.5471	169.48	25.09	0.6460	5.4355
	DiT4SR (CFG=8.0)	40	0.3869	0.2508	65.75	**0.6287**	163.05	25.40	0.6657	5.9711
	ODTSR (f=1.0)	1	0.2592	**0.1926**	62.86	0.6227	119.86	28.14	0.7736	4.6903
	CCDiT	1	**0.2290**	0.1977	61.46	0.6037	**104.45**	**28.66**	0.7764	**4.4190**
DIV2K-Val	OSEDiff	1	0.2942	0.1975	67.96	0.6147	26.34	23.72	0.6109	6.3291
	PiSA-SR	1	0.2823	0.1934	69.68	0.6401	25.09	23.87	0.6058	6.1938
	TSD-SR	1	0.2673	0.1821	**71.69**	0.6226	23.29	23.02	0.5808	6.8515
	TVT	1	0.2773	0.1860	68.67	0.6060	24.78	24.23	**0.6292**	5.9634
	SUPIR (CFG=4.0)	50	0.3919	0.2312	63.86	0.5903	31.40	22.13	0.5259	6.6931
	DiT4SR (CFG=8.0)	40	0.3448	0.2100	68.07	0.6430	30.57	21.78	0.5485	7.8203
	ODTSR (f=1.0)	1	0.2761	**0.1700**	67.94	**0.6438**	**21.47**	23.92	0.6108	5.9247
	CCDiT	1	**0.2515**	0.1943	64.57	0.6213	22.56	**24.56**	0.6178	**5.4659**

Table 2. Color correction on CCDiT outputs without Alpha Mask (RealSR). Default CCDiT includes Alpha Mask. Correction steps are applied from left to right. Bold/underline: best/second best; ties follow displayed precision.

Method	LPIPS↓	DISTS↓	MUSIQ↑	MANIQA↑	FID↓	PSNR-RGB↑	SSIM-RGB↑	Mean $\Delta E2000$ ↓
CCDiT w/o Alpha Mask	0.1926	0.1777	64.96	**0.6323**	79.80	24.79	0.7233	4.6296
CCDiT w/o Alpha Mask + Fixed CbCr	0.2076	**0.1758**	63.86	0.6236	92.99	24.64	0.7138	4.7790
CCDiT (with Alpha Mask)	0.1918	0.1768	64.98	0.6305	**79.69**	24.81	0.7233	4.5626
CCDiT w/o Alpha Mask + AdaIN	0.1918	0.1770	**64.99**	0.6301	80.76	25.73	0.7259	4.1688
CCDiT w/o Alpha Mask + Wavelet	0.1934	0.1783	64.83	0.6285	80.35	25.29	0.7237	4.3828
CCDiT w/o Alpha Mask + AdaIN + Alpha Mask	**0.1915**	0.1768	64.97	0.6280	80.66	**25.74**	0.7261	**4.1307**
CCDiT w/o Alpha Mask + AdaIN + AdaIN	0.1919	0.1770	**64.99**	0.6301	80.76	25.73	0.7259	4.1688
CCDiT w/o Alpha Mask + AdaIN + Wavelet	0.1925	0.1771	64.65	0.6274	80.87	25.73	**0.7273**	4.1874

mask. The second block applies AdaIN followed by the indicated correction. Table 3 tests the shared mask on OSEDiff and TVT without backbone changes.

In Table 2, fixed CbCr replacement degrades LPIPS, PSNR-RGB, SSIM-RGB, and Mean $\Delta E2000$ relative to no mask, confirming that spatial adaptation matters. The learned CCDiT mask improves LPIPS, PSNR-RGB, and Mean $\Delta E2000$ while preserving the best FID of 79.69. Applying AdaIN to CCDiT without Alpha Mask reaches PSNR-RGB 25.73 and Mean $\Delta E2000$ 4.1688; applying Alpha Mask afterward further improves them to 25.74 and 4.1307. Sequential AdaIN and Wavelet correction of the same backbone output instead leads SSIM-RGB, so the evidence supports complementary global and local correction without universal dominance.

The shared checkpoint in Table 3 lowers Mean $\Delta E2000$ from 5.1556 to 4.9010 for OSEDiff and from 4.7171 to 4.5687 for TVT after official AdaIN alignment. For OSEDiff, it also improves LPIPS, DISTS, MUSIQ, PSNR-RGB, and SSIM-

Table 3. Transfer on RealSR after each backbone's official AdaIN. One shared mask checkpoint is used; bold marks the better result per backbone.

Backbone	Method	LPIPS↓	DISTS↓	MUSIQ↑	MANIQA↑	FID↓	PSNR-RGB↑	SSIM-RGB↑	Mean $\Delta E2000$ ↓
OSEDiff	AdaIN	0.2921	0.2128	69.09	**0.6326**	**123.49**	23.59	0.6833	5.1556
	AdaIN + Alpha Mask	**0.2888**	**0.2104**	**69.25**	0.6274	129.16	**23.63**	**0.6851**	**4.9010**
TVT	AdaIN	**0.2597**	0.2061	69.89	**0.6232**	**109.90**	24.25	**0.7179**	4.7171
	AdaIN + Alpha Mask	0.2609	**0.2012**	**70.01**	0.6210	116.61	**24.28**	0.7173	**4.5687**

Table 4. Five-loss ablation on RealSR. Bold/underline: best/second best; ties follow displayed precision.

Setting	LPIPS↓	DISTS↓	MUSIQ↑	MANIQA↑	FID↓	PSNR-RGB↑	SSIM-RGB↑	Mean $\Delta E2000$ ↓
Full (all five losses)	**0.1918**	0.1768	64.98	0.6305	79.69	**24.8092**	0.7233	**4.5626**
w/o $\mathcal{L}_{var}$	0.1924	0.1770	64.99	<u>0.6311</u>	79.63	24.8048	<u>0.7234</u>	4.5762
w/o $\mathcal{L}_{oracle}$	0.1928	0.1771	64.96	**0.6313**	79.60	24.8031	**0.7235**	4.5832
w/o $\mathcal{L}_{tv}$	0.1924	**0.1765**	**65.02**	0.6310	**79.57**	24.8040	0.7232	<u>4.5633</u>
w/o $\mathcal{L}_{rgb}$	<u>0.1922</u>	<u>0.1767</u>	<u>65.01</u>	0.6308	<u>79.58</u>	<u>24.8074</u>	0.7233	4.5634
w/o $\mathcal{L}_{chroma}$	<u>0.1922</u>	0.1769	64.99	0.6308	79.64	24.8052	<u>0.7234</u>	4.5679

RGB; for TVT, it improves DISTS, MUSIQ, and PSNR-RGB. Other metrics decline slightly, indicating color-oriented rather than uniform quality enhancement. Because the mask was not adapted to either backbone, backbone-specific fine-tuning may help but remains untested.

In Table 4, the full objective achieves the best LPIPS (0.1918), highest PSNR-RGB (24.8092), and lowest Mean $\Delta E2000$ (4.5626). The w/o-TV variant leads DISTS, MUSIQ, and FID, but slightly reduces PSNR-RGB and increases color error; w/o Oracle leads MANIQA and SSIM-RGB. Full is therefore selected as the most balanced configuration for the method's reconstruction and color-fidelity objectives while retaining competitive perceptual quality.

4.5 Discussion

The experiments reveal a fidelity–perception trade-off rather than a universally dominant setting. Alpha Mask changes only chrominance, so gains in PSNR-RGB and $\Delta E2000$ need not improve every perceptual metric. Transfer after AdaIN further demonstrates compatibility with different output distributions. However, the module can only interpolate colors present in the SR output or reference, and the shared checkpoint has not been adapted to backbone-specific artifacts. Model-specific training and repeated-run uncertainty therefore remain important future evaluations.

The examples in Fig. 4 qualitatively show CCDiT color preservation and reduced OSEDiff tone drift after Alpha Mask fusion; they are not a benchmark.

4.6 Computational Efficiency

On an RTX PRO 6000, we time the complete fusion path for one 512×512 pair (batch 1) over 20 synchronized CUDA-event trials after 10 warm-ups. We include feature construction, Mask Net, Sigmoid, color conversion, and fusion, but exclude I/O, resize, transfer, and the SR backbone.

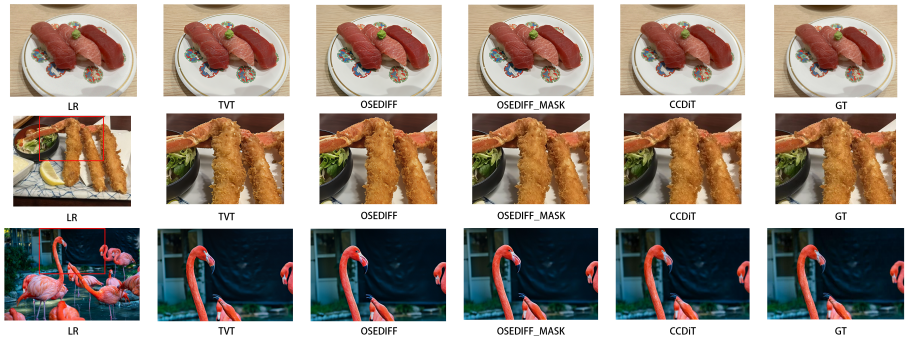


Fig. 4. Non-benchmark examples. Red boxes mark crops; the images support visual inspection only.

Table 5. Alpha Mask microbenchmark on RTX PRO 6000 (batch 1); memory uses MiB.

Resolution	Avg. latency (ms)	Min–Max (ms)	Std. (ms)	Equiv. FPS	Param. memory (MiB)	Peak allocated (MiB)	Extra peak (MiB)
512 × 512	2.2704	2.1821–3.6680	0.3212	440.46	0.08	87.18	81.10

Table 5 is a module microbenchmark, not end-to-end CCDiT or cross-device latency.

5 Conclusion

We presented CCDiT, which adapts Qwen-Image-Edit-2509 for one-step super-resolution, and Alpha Mask fusion for adaptive chrominance correction with restored luminance preserved. The correction module requires no additional iterative diffusion sampling steps. The mask predicts spatially varying fusion weights from the restored output and bicubic reference, allowing local chrominance refinement while keeping the SR backbone fixed. Experiments on RealSR, DRealSR, and DIV2K-Val support its reconstruction and color-fidelity performance. The ablations further support spatially adaptive fusion over fixed CbCr replacement on CCDiT outputs. A shared Alpha Mask checkpoint also improves color consistency after the official AdaIN alignment of OSEDiff and TVT without backbone modification. These results support combining global AdaIN alignment with local Alpha Mask refinement, while showing that improved color fidelity does not uniformly improve perceptual metrics. Its effectiveness depends on reference-color reliability, and backbone-specific adaptation and broader cross-dataset evaluation remain future work.

Acknowledgements This work was supported by JST CRONOS JPMJCS24K4 and JSPS KAKENHI 26K23875.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: Dataset and study. In: CVPR Workshops. pp. 1122–1131 (2017)
2. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: ICCV. pp. 3086–3095 (2019)
3. Chen, J., Yu, J., Ge, C., et al.: PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: ICLR (2024)
4. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: CVPR. pp. 22367–22377 (2023)
5. Chiu, M.C., Wang, Y., Kim, D.E.G., Chen, P.Y., Ma, X.: ColorSense: A study on color vision in machine visual recognition. In: IEEE SaTML. pp. 681–697 (2025)
6. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE TPAMI **44**(5), 2567–2581 (2022)
7. Dong, L., Fan, Q., Guo, Y., et al.: TSD-SR: One-step diffusion with target score distillation for real-world image super-resolution. In: CVPR. pp. 23174–23184 (2025)
8. Duan, Z.P., Zhang, J., Jin, X., et al.: DiT4SR: Taming diffusion transformer for real-world image super-resolution. In: ICCV. pp. 18948–18958 (2025)
9. Fang, Y., Chen, Y., Yin, S., et al.: One-step diffusion transformer for controllable real-world image super-resolution. In: CVPR. pp. 23440–23450 (2026)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: NeurIPS. vol. 30, pp. 6626–6637 (2017)
11. Ho, J., Jain, A.N., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. vol. 33, pp. 6840–6851 (2020)
12. Hore, A., Ziou, D.: Image quality metrics: PSNR vs. SSIM. In: ICPR. pp. 2366–2369 (2010)
13. Hu, E.J., Shen, Y., Wallis, P., et al.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
14. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV. pp. 1501–1510 (2017)
15. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: MUSIQ: Multi-scale image quality transformer. In: ICCV. pp. 5148–5157 (2021)
16. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML. vol. 162, pp. 12888–12900. PMLR (2022)
17. Li, Y., Zhang, K., Liang, J., et al.: LSDIR: A large scale dataset for image restoration. In: CVPR Workshops. pp. 1775–1787 (2023)
18. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SwinIR: Image restoration using Swin transformer. In: ICCV Workshops. pp. 1833–1844 (2021)
19. Lin, X., He, J., Chen, Z., et al.: DiffBIR: Toward blind image restoration with generative diffusion prior. In: ECCV. pp. 430–448. Springer (2024)
20. Moser, B.B., Frolov, S., Raue, F., Palacio, S., Dengel, A.: Dynamic attention-guided diffusion for image super-resolution. In: WACV. pp. 451–460 (2025)
21. Ofli, F., Aytar, Y., Weber, I., al Hammouri, R., Torralba, A.: Is Saki #delicious? The food perception gap on Instagram and its relation to health. In: WWW. pp. 509–518. International World Wide Web Conferences Steering Committee (2017)
22. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4195–4205 (2023)

23. Poynton, C.: Digital Video and HDTV: Algorithms and Interfaces. Morgan Kaufmann, San Francisco (2003)
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
25. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
26. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE TPAMI **45**(4), 4713–4726 (2023)
27. Sharma, G., Wu, W., Dalal, E.N.: The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. Color Research and Application **30**(1), 21–30 (2005)
28. Singh, A., Bay, A., Mirabile, A.: Assessing the importance of colours for CNNs in object recognition. arXiv preprint arXiv:2012.06917 (2020)
29. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
30. Sun, L., Wu, R., Ma, Z., Liu, S., Yi, Q., Zhang, L.: Pixel-level and semantic-level adjustable super-resolution: A dual-LoRA approach. In: CVPR. pp. 2333–2343 (2025)
31. Wang, J., Yue, Z., Zhou, S., Chan, K.C.K., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. IJCV **132**(12), 5929–5949 (2024)
32. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In: ICCV Workshops. pp. 1905–1914 (2021)
33. Wang, X., Yu, K., Wu, S., et al.: ESRGAN: Enhanced super-resolution generative adversarial networks. In: ECCV Workshops. pp. 63–79. Springer (2019)
34. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004)
35. Wei, P., Xie, Z., Lu, H., et al.: Component divide-and-conquer for real-world image super-resolution. In: ECCV. pp. 101–117. Springer (2020)
36. Wu, C., Li, J., Zhou, J., et al.: Qwen-Image technical report. arXiv preprint arXiv:2508.02324 (2025)
37. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. In: NeurIPS. vol. 37, pp. 92529–92553 (2024)
38. Xu, N., Price, B., Cohen, S., Huang, T.: Deep image matting. In: CVPR. pp. 2970–2979 (2017)
39. Yang, S., Wu, T., Shi, S., et al.: MANIQA: Multi-dimension attention network for no-reference image quality assessment. In: CVPR Workshops. pp. 1191–1200 (2022)
40. Yi, Q., Li, S., Wu, R., Sun, L., Wu, Y., Zhang, L.: Fine-structure preserved real-world image super-resolution via transfer VAE training. In: ICCV. pp. 12415–12426 (2025)
41. Yu, F., Gu, J., Li, Z., et al.: Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In: CVPR. pp. 25669–25680 (2024)
42. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)