REFPLC: LATENT RECOVERY AND STATE-CONDITIONED PIXEL MODULATION FOR REFERENCE-BASED SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Severe degradation removes the very cues needed to identify missing image detail. A sharp reference can supply related evidence, but differences in pose, viewpoint, and appearance make direct transfer unreliable. Reference-guided generation must therefore address two linked questions: how to turn an unaligned observation into a target-compatible condition, and how to determine its additional contribution as reconstruction evolves. We propose RefPLC, which treats reference guidance as bounded restoration increments at the latent-condition and RGB-velocity interfaces of pixel diffusion. Restoration-Oriented Latent Transport (ROLT) aligns detail-sensitive VAE differences and interprets them relative to the degraded target, preserving the LQ latent as the condition's residual base. State-Conditioned Velocity Modulation (SCVM) compares shared-decoder responses with and without reference evidence at the same sampling state, then spatially modulates the resulting correction. The two stages connect reference preparation to its use within one restoration trajectory. On Single717, RefPLC achieves 25.81 dB PSNR and 67.89 MUSIQ, improving over the equally adapted PiD baseline by 1.89 dB and 4.02, respectively. External evaluation yields 45.88 FID on WR-SR and 28.17 dB PSNR on Face. In a paired 90-target study, Top references improve PSNR by 0.761 dB over the same trained model without a reference. These results support separating target-conditioned evidence recovery from state-dependent reference application.

1 INTRODUCTION

Restoring a severely degraded image requires more than generating plausible texture. The result should retain the particular people, objects, and layout observed in the target. A generative prior helps resolve ambiguous appearance, but information removed by blur, noise, and downsampling cannot be uniquely inferred from the LQ image alone. Reference-based super-resolution (RefSR) offers an additional observation of related content, potentially supplying instance-specific evidence that the target no longer reveals (Zhang et al., 2019; Yang et al., 2020; Guo et al., 2024).

That extra observation creates a tension: it is informative precisely because it contains detail absent from the LQ image, yet this absence makes its relevance difficult to verify. Changes in viewpoint, pose, illumination, and visible content further separate the reference from the desired output. Correspondence learning addresses where to retrieve information (Jiang et al., 2021; Cao et al., 2022; Zhou et al., 2025), while diffusion-based methods regulate reference features through spatial or correlation-aware conditioning (Guo et al., 2024; Wang et al., 2026). These advances motivate a more specific question for iterative high-resolution generation: what reference-derived change should be made to the target condition, and what remains useful to the current prediction?

A correspondence is not yet a restoration condition. A matched reference region contains both useful detail and appearance tied to its own image. Transferring its full representation can mix target-relevant cues with different lighting, pose, or background. Strong degradation also weakens the target features used to distinguish these components. The challenge is therefore not resolved by a match score alone: the retrieved value must be mapped to the target grid and interpreted relative to the information already retained by the LQ condition. This requires separating the representation used to locate evidence from the content used to recover the target.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Available evidence is not the same as a needed update. A reference remains fixed during sampling, whereas the generated image and the base model's prediction change. A region that supplies missing structure in one state may overlap with detail already produced in another; a strong similarity score does not quantify this incremental need. Existing reference-attention gates address related reliability concerns (Guo et al., 2024; Wang et al., 2026). Our question is how to express the additional reference response at a fixed sampling state, distinct from the decoder's reference-independent response, and connect that update to an already recovered target condition.

Our organizing principle is to express reference guidance as a *restoration increment* at two connected interfaces. PiD conditions RGB diffusion on a compact latent (Lu et al., 2026). We first recover that condition with target-grid reference evidence, then adapt the reference contribution to each pixel update. Unlike regulation within reference-attention features (Guo et al., 2024; Wang et al., 2026), our formulation explicitly parameterizes a latent-condition residual and an RGB-velocity residual. The recovered condition determines the base prediction seen by pixel modulation, connecting the two modules within one trajectory.

Restoration-Oriented Latent Transport (ROLT) implements the first interface with transported VAE differences, target-conditioned mapping, and a bounded update. **State-Conditioned Velocity Modulation (SCVM)** implements the second by differencing the same decoder's responses with and without evidence, then applying a spatial gain. Single717, four external benchmarks, and a paired 90-target study evaluate reconstruction, perceptual quality, and reference dependence. With the same training split, 10-epoch adaptation, and frozen base Pixel Diffusion network, RefPLC improves PSNR over PiD by 1.89 dB. The ablation evaluates the complete modules and their combination.

2 Related Work

Correspondence and reference restoration. SRNTT transfers neural textures across image scales, and TTSR learns texture retrieval and transfer through attention (Zhang et al., 2019; Yang et al., 2020). C^2-Matching learns transformation-robust correspondence; MASA-SR combines coarse-to-fine search with spatial adaptation; DATSR uses deformable aggregation; and SSMTF introduces multi-scale state-space texture fusion (Jiang et al., 2021; Lu et al., 2021; Cao et al., 2022; Zhou et al., 2025). ROLT separates localization from recovery: matching descriptors locate related content, while transported sharp-minus-blurred VAE values modify the target condition.

Reference-guided diffusion. Ref-Diff uses change-aware conditional diffusion in remote sensing, and iRAG retrieves real or generated reference patches with hallucination filtering (Dong et al., 2024; Lee et al., 2025). ReFIR regulates cross-image attention through spatial gating and distribution alignment, whereas Ada-RefSR gates reference attention using implicit correlation (Guo et al., 2024; Wang et al., 2026). TriFlowSR and CRefDiff explore high-resolution and local–global reference guidance (Shi et al., 2025; Wang & Sun, 2025). RefPLC combines a transported latent residual before condition projection with a same-state decoder-difference residual at the RGB velocity output; Appendix I compares these interfaces.

Generative restoration. SR3 demonstrates iterative diffusion-based super-resolution (Saharia et al., 2021), while latent diffusion moves the generative state to a compressed representation (Rombach et al., 2022). SeeSR, S3Diff, OSEDiff, PiSA-SR, and SUPIR develop semantic guidance, efficient sampling, and large restoration priors (Wu et al., 2024c; Zhang et al., 2024a; Wu et al., 2024b; Sun et al., 2025; Yu et al., 2024). PiD instead conditions pixel-space generation on latents (Lu et al., 2026). Our two reference paths exploit this distinction: the latent is a condition, and the evolving state remains an RGB image.

3 Data and Evaluation Sets

Our internal training split contains **97,589 GT–Ref pairs and 64,427 GT targets**. **Single717** evaluates overall restoration on 717 targets with one reference each: 300 people (41.84%), 145 products (20.22%), 122 scene-text images (17.02%), and 150 video frames (20.92%). Video frames use their Top reference. Each target contributes once to the aggregate evaluation.

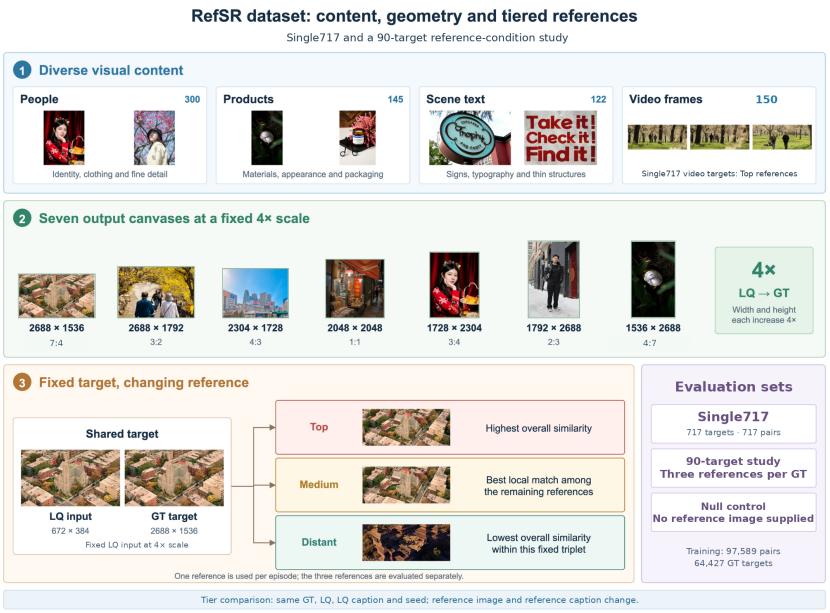

Figure 1: **Training and test organization.** Single717 measures one-reference restoration. The reference-condition study uses 90 GT targets, three references per target, and a no-reference control. The $4\times$ task covers seven native output canvases.

A separate **90-target reference study** associates each GT with Top, Medium, and Distant references, supplied one at a time. Top prioritizes overall similarity, Medium local correspondence among remaining candidates, and Distant the least globally similar member of the selected triplet. **Null means that no reference image is provided**; it is an additional control, not a fourth similarity tier. Within each target group, the GT, LQ, LQ caption, and seed are fixed. The three reference conditions contain 270 GT–Ref pairs; the no-reference control uses the same 90 targets. Figure 1 summarizes the organization.

We retain square, landscape, and portrait output sizes up to a longest side of 2688 pixels, with a fixed $4\times$ spatial scale. LQ inputs are generated with the official Real-ESRGAN degradation procedure (Wang et al., 2021). The Single717 main comparison uses the same LQ input for every method on each target. Exact canvas dimensions and evaluation groups are listed in Appendix G.

4 Method

4.1 Overall framework

Given LQ $I_L \in \mathbb{R}^{3\times H\times W}$ and an unaligned reference I_R, RefPLC predicts an RGB image of size $4H \times 4W$. Captions $c_L = C_L(I_L)$ and $c_R = C_R(I_R)$ come from the input images (Zhang et al., 2023; Wu et al., 2024c). The LQ caption conditions PiD; the reference caption regulates SCVM evidence. One frozen VAE encoder E encodes the LQ, sharp reference, and blurred reference, with $z_L = E(I_L)$. ROLT produces a target-grid latent increment $\Delta z_{L\leftarrow R}$ with the shape of z_L. SCVM produces a state-dependent RGB-velocity correction $\Delta v_{R,t}$ with the output image shape. With condition projection P and PiD velocity predictor f_θ,

$$\begin{aligned} v_t^{\text{base}} &= f_\theta(x_t, t; P(z_L + \Delta z_{L\leftarrow R}), c_L), \\ v_t &= v_t^{\text{base}} + \kappa \Delta v_{R,t}, \qquad \kappa = 0.25. \end{aligned} \tag{1}$$

Here x_t is the current RGB sample. We use noise time $x_t = (1-t)y + t\epsilon$ for clean image y and Gaussian noise ϵ. Sampling proceeds from 1 to 0, and $x_t - tv_t$ estimates the clean endpoint. ROLT is evaluated once per image pair; SCVM is reevaluated with the sampling state (Figure 2).

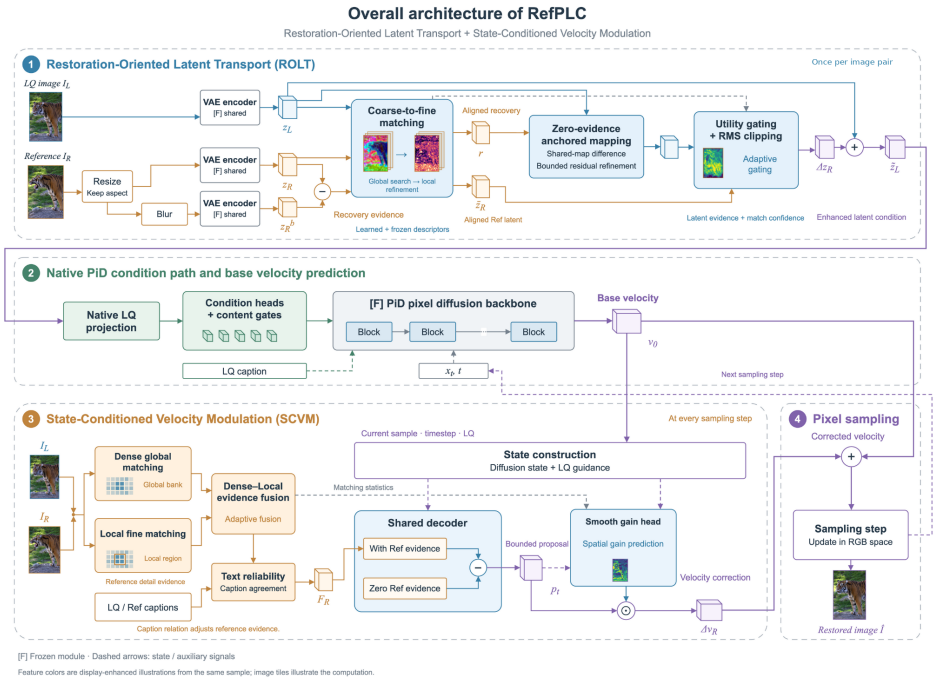

Figure 2: **RefPLC architecture.** ROLT prepares the condition once per image pair; SCVM adjusts each RGB update. Compact labels r, Δz_R, and v_0 denote the transported residual $\hat{r}$, target-grid increment $\Delta z_{L\leftarrow R}$, and current base velocity v_t^{base}. The sampler receives $v_t = v_t^{\text{base}} + 0.25\Delta v_{R,t}$. [F]: frozen.

4.2 Restoration-Oriented Latent Transport

ROLT transfers detail-sensitive values to the target condition (Figure 3a). T_R preserves the reference aspect ratio, and B_R is a fixed blur. The same frozen encoder processes both reference versions. Their difference is transported to target location i using affinities A_{ij} over reference candidates $\mathcal{N}_i$:

$$d_R = E(T_R(I_R)) - E(B_R(T_R(I_R))),$$
$$\hat{r}_i = \sum_{j\in\mathcal{N}_i} A_{ij} d_{R,j}, \qquad A_{ij} \geq 0, \quad \sum_j A_{ij} = 1. \tag{2}$$

Normalization applies to nonempty valid candidate sets; empty support gives zero evidence. Learned latent descriptors and frozen C^2 features select a coarse reference region. Fine affinities combine local similarity, that coarse prior, and valid support. The same affinities align the sharp reference latent for gating. Descriptors thus determine *where* to retrieve, while VAE differences determine *what* is transferred.

A local map H, comprising three 3×3 convolutions, interprets this evidence relative to z_L. Its two evaluations use identical parameters and differ only in reference evidence:

$$u_i = \hat{r}_i + \gamma s_i \tanh[H(z_L/s, \hat{r}/s)_i - H(z_L/s, \mathbf{0})_i]\,. \tag{3}$$

The detached channel RMS s_i normalizes the target, with a numerical floor. The skip and map difference vanish for zero evidence. A learned spatial gate g uses latent relationships and matching peak and concentration statistics to weight each target location. Its one-channel output lies in $[0, 1]$ and is broadcast over latent channels. Gating precedes an image-level RMS cap:

$$a = \lambda_z g \odot u, \qquad \Delta z_{L\leftarrow R} = a \min\left(1, \frac{\rho\,\mathrm{RMS}(z_L)}{\mathrm{RMS}(a) + \varepsilon}\right),$$
$$\tilde{z}_L = z_L + \Delta z_{L\leftarrow R}. \tag{4}$$

RMS is computed per image over channels and positions. We use $\gamma = 0.25$, $\lambda_z = 0.5$, and $\rho = 0.15$. Both skip and mapping pass through the gate and cap. $\mathcal{L}_{\text{geo}}$ supervises correspondence and confidence using known training relations. $\mathcal{L}_{\text{lat}}$ supervises the recovered latent and gate utility against clean targets (Appendices B and E).

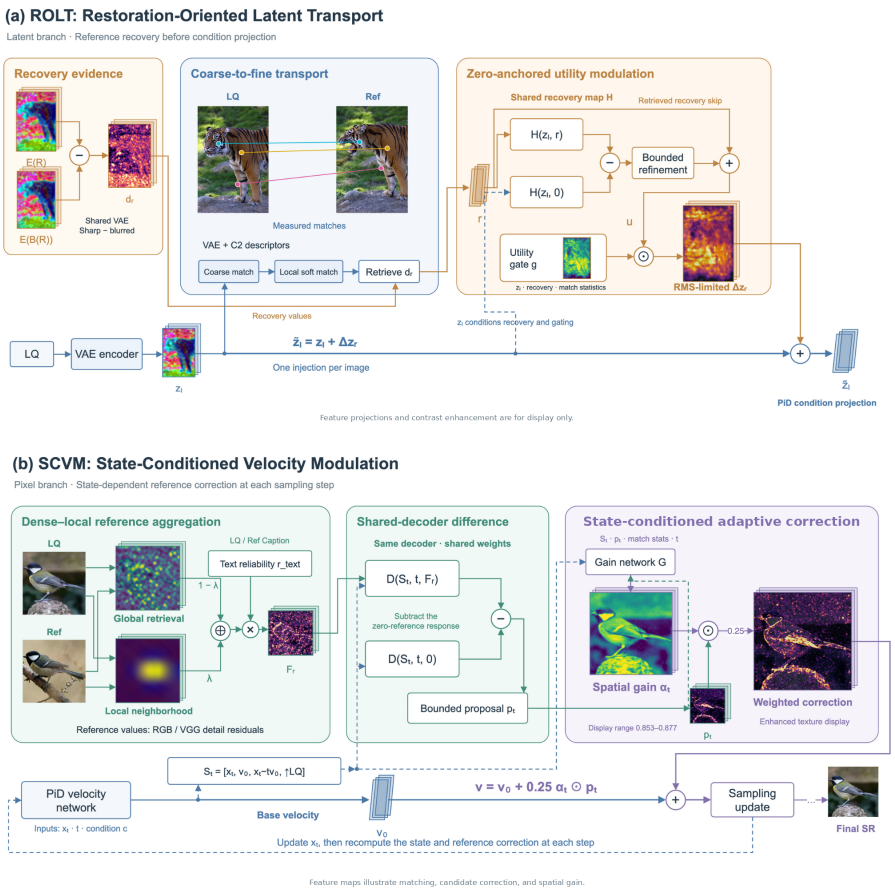

Figure 3: **Reference modules.** (a) ROLT: $r = \hat{r}$ is transported evidence; $\Delta z_r = \Delta z_{L \leftarrow R}$ is the gated, RMS-limited condition update. (b) SCVM: $v_0 = v_t^{\text{base}}$, $\lambda = \omega_i$, and $v = v_t$. The bounded proposal p_t is weighted by spatial gain α_t and added to the base velocity with scale 0.25.

4.3 State-Conditioned Velocity Modulation

SCVM adapts reference evidence to the current sample (Figure 3b). Dense f_i^G aggregates RGB and VGG-feature differences (Simonyan & Zisserman, 2015) over the valid reference bank; local f_i^L uses a 5×5 neighborhood of a coarse match. Both lie on the target grid. A learned function M maps correspondence confidence and concentration statistics m_i to a mixing weight:

$$f_i = (1 - \omega_i) f_i^G + \omega_i f_i^L, \qquad \omega_i = 0.75\,\text{sigmoid}(M(m_i)). \tag{5}$$

Fusion retains at least 0.25 global weight. A bounded caption-compatibility signal $r_{\text{text}} = h_{\text{text}}(c_L, c_R)$ adjusts the fused evidence as $F_R = \mathcal{R}(f; r_{\text{text}})$, without replacing PiD's LQ caption.

The state $S_t = \text{Concat}[x_t, v_t^{\text{base}}, x_t - t v_t^{\text{base}}, \mathcal{U}(I_L)]$ concatenates the sample, base velocity, base clean estimate, and upsampled LQ into 12 channels. $\mathcal{U}$ aligns the canvas and RGB range. The same decoder D is evaluated with reference and shape-matched zero evidence, holding parameters, state, and time fixed:

$$p_t = \beta \tanh\left(\frac{D(S_t, t, F_R) - D(S_t, t, \mathbf{0}_{F_R})}{\beta}\right), \qquad \beta = 0.5. \tag{6}$$

The RGB-shaped proposal measures the decoder's additional evidence response in a state whose base prediction already incorporates ROLT.

A gain head uses S_t, detached p_t, aligned matching statistics, and time. A coarse sigmoid-bounded prediction is bilinearly upsampled to a one-channel gain, broadcast over RGB:

$$\Delta v_{R,t} = \alpha_t \odot p_t, \qquad \alpha_t \in [0, 1]^{1 \times 4H \times 4W}. \tag{7}$$

The proposal specifies the update; the gain controls its spatial use. Both are recomputed at every vector-field evaluation (Appendix C).

4.4 Learning and inference

The base recovery loss supervises the composed velocity and its one-step clean estimate at GT resolution. The total training objective adds the two ROLT supervision terms:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \lambda_{\text{lat}}\mathcal{L}_{\text{lat}}. \tag{8}$$

Both auxiliary weights are 0.05. The base loss updates ROLT, SCVM, and selected PiD conditioning parameters; auxiliary losses update ROLT. Other backbone and feature-extractor parameters remain frozen. Image-pair evidence is cached. RefPLC uses 25 sampling steps and guidance scale 5, recomputing state-dependent corrections at each vector-field evaluation (Appendices D–F).

5 Experiments

5.1 Setup

Single717 methods share the same LQ inputs from the official Real-ESRGAN procedure. SSMTF/Ada-RefSR use RefPLC's training split and optimization updates (10 epochs). PiD and module variants use the same split and duration; PiD adapts its condition pathway at the corresponding RefPLC learning rate, with Pixel Diffusion frozen. PiD appears in the main and ablation tables. On the four external benchmarks, P marks published Ada-RefSR scores (Wang et al., 2026); other rows, including Nano Banana Pro (Raisinghani, 2025), use official LQ images. Non-FT SSMTF/Ada-RefSR use released weights; FT uses RefPLC's training data and updates. Each method follows its official inference pipeline.

Single717 reports RGB PSNR/SSIM, LPIPS-VGG, and eight no-reference measures. External tables use Y-channel PSNR/SSIM and report LPIPS, FID, NIQE, MUSIQ, and CLIPIQA. Appendix G.4 specifies their provenance. Multi-method tables mark the two best recorded values in red bold and blue underline; two-method tables bold the better value. External LPIPS is not ranked across sources.

5.2 Quantitative comparison

Table 1: Single717: 717 targets with common LQ inputs. SSMTF/Ada-RefSR use RefPLC's fine-tuning data and updates (10 epochs). PiD adapts its condition pathway at the corresponding learning rate, with Pixel Diffusion frozen. Red bold/blue underlined: best/second-best recorded values.

Method	PSNR RGB↑	SSIM RGB↑	LPIPS VGG↓	MUSIQ↑	NIQE↓	DeQA↑	MANIQA↑	Q-Align↑	UniPercept IAA↑	UniPercept IQA↑	VisualQuality R1↑
FLUX.1 Kontext	15.22	0.5530	0.5890	67.42	6.031	3.720	0.3140	4.140	60.56	66.56	4.116
BAGEL	16.82	0.5940	0.6390	57.28	9.141	2.589	0.2260	2.659	44.15	42.58	2.533
Qwen-Image-Edit-2511	16.24	0.5853	0.5406	63.74	6.799	3.806	0.2953	4.273	61.01	65.70	4.203
Nano Banana Pro	17.14	0.5829	0.5550	67.02	5.669	**3.981**	**0.3227**	4.392	64.45	70.30	**4.405**
OmniGen2	11.24	0.4549	0.6513	66.12	6.895	3.866	0.3124	4.335	60.37	68.69	4.268
PiD	23.92	0.6796	0.4731	63.87	6.089	3.749	0.2850	4.151	61.90	66.71	4.172
SSMTF	**25.99**	0.6984	0.5916	54.06	9.313	2.356	0.2373	2.425	48.97	40.74	2.329
Ada-RefSR	25.52	**0.7575**	**0.4319**	60.94	6.218	3.261	0.2998	3.539	53.18	54.95	3.341
RefPLC	25.81	0.7079	0.4463	**67.89**	**5.325**	3.964	0.2969	**4.401**	**65.67**	**70.53**	4.345

Single717. Among the nine methods in Table 1, RefPLC records five first-place and ten top-two values across eleven metrics. Its 25.81 dB PSNR and 67.89 MUSIQ exceed the adapted PiD baseline by 1.89 dB and 4.02. Against Ada-RefSR, PSNR is 0.29 dB higher, MUSIQ 6.95 higher, and Q-Align 0.862 higher. SSMTF reaches 25.99 dB PSNR; RefPLC combines 0.4463 LPIPS-VGG and 67.89 MUSIQ versus SSMTF's 0.5916 and 54.06. Compared with Nano Banana Pro, RefPLC obtains 8.67 dB higher PSNR and higher MUSIQ, Q-Align, and UniPercept values. These comparisons jointly describe target fidelity and perceptual quality.

Table 2: External benchmarks. P: scores from Ada-RefSR (Wang et al., 2026); other rows use official LQ inputs. Non-FT SSMTF/Ada-RefSR: official weights; FT: RefPLC training data and updates (10 epochs). S1/S50: inference steps. Red bold/blue underlined: best/second-best recorded values per dataset, excluding cross-source LPIPS ranking.

Dataset	Method	PSNR-Y↑	SSIM-Y↑	LPIPS↓	FID↓	NIQE↓	MUSIQ↑	CLIPIQA↑
CUFED5	Real-ESRGAN-S1^{P}	20.31	**0.5543**	0.36970	175.91	3.886	64.71	0.4214
	S3Diff-S1^{P}	20.46	0.5234	0.35440	160.13	**3.710**	63.85	0.5238
	OSEDiff-S1^{P}	18.87	0.4935	0.36690	150.75	4.001	65.24	0.5077
	PiSA-SR-S1^{P}	20.34	0.5286	0.37070	164.69	4.199	64.28	0.5043
	SUPIR+ReFIR-S50^{P}	19.00	0.4711	0.43590	151.27	4.275	62.74	0.4235
	SeeSR+ReFIR-S50^{P}	20.22	0.5255	0.34520	137.37	3.743	**70.54**	0.5538
	Nano Banana Pro	19.58	0.4250	0.50700	251.80	4.101	62.90	0.4020
	SSMTF	21.03	0.5315	0.82450	294.30	9.906	16.83	0.1421
	SSMTF FT	**21.06**	0.5354	0.83230	293.39	10.195	16.84	0.1569
	Ada-RefSR-S1^{P}	20.48	0.5461	0.28940	127.89	3.870	67.91	0.5174
	Ada-RefSR FT	20.54	0.5474	0.28920	**126.38**	3.851	68.14	0.5162
	RefPLC	21.00	0.5446	0.38822	142.59	3.797	65.71	**0.5588**
WR-SR	Real-ESRGAN-S1^{P}	22.15	**0.5974**	0.36310	97.91	3.698	64.52	0.6076
	S3Diff-S1^{P}	21.91	0.5620	0.35420	63.82	4.417	63.82	0.6149
	OSEDiff-S1^{P}	21.33	0.5603	0.32610	71.51	3.736	71.19	0.6899
	PiSA-SR-S1^{P}	21.88	0.5682	0.33830	63.00	4.036	69.52	0.6407
	SUPIR+ReFIR-S50^{P}	21.01	0.5378	0.39490	76.50	3.641	70.12	0.6219
	SeeSR+ReFIR-S50^{P}	21.83	0.5673	0.34350	61.96	3.892	72.15	0.7120
	Nano Banana Pro	20.21	0.4600	0.53000	143.11	3.841	67.93	0.6190
	SSMTF	22.95	0.5860	0.77610	153.33	10.330	19.76	0.1675
	SSMTF FT	**23.05**	0.5937	0.78230	152.79	10.319	19.46	0.1896
	Ada-RefSR-S1^{P}	21.97	0.5777	0.30610	53.28	3.743	69.86	0.6612
	Ada-RefSR FT	22.08	0.5805	0.30580	54.22	3.769	69.90	0.6563
	RefPLC	22.45	0.5929	0.30606	**45.88**	**3.610**	**72.24**	**0.7169**
Bird	Real-ESRGAN-S1^{P}	25.08	0.7044	0.53060	178.88	8.850	29.31	0.3571
	S3Diff-S1^{P}	24.84	0.7106	0.29030	60.95	5.320	73.11	0.8285
	OSEDiff-S1^{P}	23.79	0.7086	0.29490	49.19	4.980	74.76	0.8451
	PiSA-SR-S1^{P}	24.49	0.7003	0.28160	68.00	4.851	74.36	0.8373
	SUPIR+ReFIR-S50^{P}	23.83	0.6974	0.38110	94.93	5.442	63.67	0.6478
	SeeSR+ReFIR-S50^{P}	23.72	0.6985	0.29450	52.40	5.618	**75.11**	**0.8617**
	Nano Banana Pro	23.51	0.6180	0.43700	72.20	**4.132**	73.26	0.7890
	SSMTF	25.29	0.7068	0.67110	178.65	13.153	17.05	0.1426
	SSMTF FT	25.39	0.7192	0.67940	175.49	12.421	16.90	0.1577
	Ada-RefSR-S1^{P}	25.30	0.7292	0.25360	**36.42**	5.177	72.00	0.8303
	Ada-RefSR FT	25.43	0.7331	0.25450	36.55	5.258	71.69	0.8246
	RefPLC	**25.81**	**0.7460**	0.28865	40.27	4.447	73.46	0.8307
Face	Real-ESRGAN-S1^{P}	26.17	0.7142	0.51130	162.43	8.950	26.48	0.3630
	S3Diff-S1^{P}	26.69	0.7344	0.21890	51.12	5.060	74.11	0.6963
	OSEDiff-S1^{P}	25.53	0.7434	0.22630	55.04	4.955	71.77	0.6459
	PiSA-SR-S1^{P}	26.24	0.7254	0.21260	48.92	4.542	**75.99**	0.7191
	SUPIR+ReFIR-S50^{P}	25.72	0.7062	0.26230	57.14	3.969	73.09	0.6260
	SeeSR+ReFIR-S50^{P}	26.32	0.7375	0.22930	57.29	5.289	75.42	**0.7236**
	Nano Banana Pro	25.43	0.6530	0.33200	79.40	**3.351**	75.04	0.7030
	SSMTF	26.50	0.7202	0.65270	177.16	12.009	17.91	0.1783
	SSMTF FT	26.66	0.7352	0.65820	178.89	11.571	17.64	0.2059
	Ada-RefSR-S1^{P}	27.13	0.7523	0.17490	**42.70**	4.488	74.38	0.6453
	Ada-RefSR FT	27.27	**0.7562**	0.17560	42.98	4.556	74.12	0.6409
	RefPLC	**28.17**	0.7524	0.25378	47.35	4.485	69.81	0.6466

External benchmarks. Table 2 reports 45.88 FID, 3.610 NIQE, 72.24 MUSIQ, and 0.7169 CLIP-IQA for RefPLC on WR-SR. Relative to the Ada-RefSR FT configuration, FID is 15.38% lower and PSNR is 0.37 dB higher. RefPLC records the highest PSNR and SSIM on Bird and the highest PSNR on Face among the listed configurations; CUFED5 CLIPIQA is 0.5588. Nano Banana Pro has the lowest NIQE on Bird and Face. Across all four datasets, RefPLC records 1.42–2.75 dB higher PSNR and lower FID than Nano Banana Pro. These comparisons retain method-specific inference settings, rather than equating sampling budgets. Full scores and their provenance appear in Appendix H.

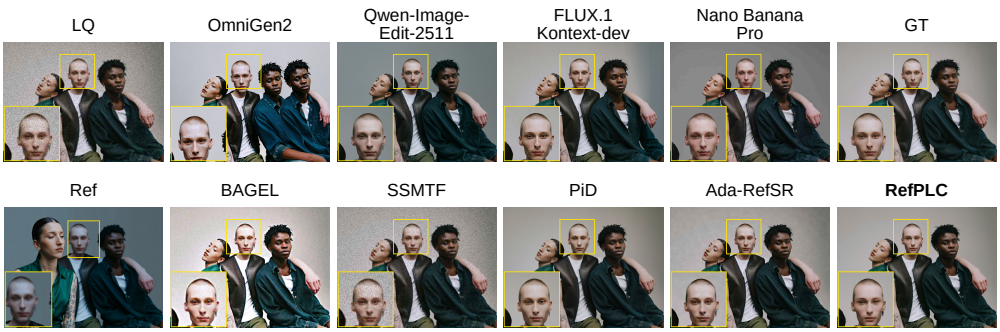

Figure 4: **Group portrait.** The nine methods in Table 1, including adapted PiD, share the same Real-ESRGAN LQ. Insets show facial detail; Appendix J.1 specifies the checkpoints.

Table 3: Reference conditions on 90 GT targets. Each reference is supplied separately; Null omits the reference. Bold marks the better recorded value within each metric and condition.

Metric	Method	Top	Medium	Distant	Null
PSNR↑	RefPLC	**30.8553**	**30.7319**	30.1697	30.0939
	Ada-RefSR	30.6515	30.5444	**30.4356**	**30.4413**
LPIPS↓	RefPLC	0.2118	0.2119	0.2069	0.2158
	Ada-RefSR	**0.1772**	**0.1811**	**0.1989**	**0.2150**

5.3 Reference Conditions and Component Ablation

Reference dependence. Table 3 separates reference-induced change from absolute cross-method quality. Relative to its own Null output, RefPLC gains 0.7614 dB with Top, 0.6380 dB with Medium, and 0.0758 dB with Distant; all three reduce its LPIPS. Top and Medium PSNR exceed Ada-RefSR by 0.2037 and 0.1874 dB, while Ada-RefSR has lower absolute LPIPS in all conditions. RefPLC's strongest PSNR response is to Top, whereas its lowest LPIPS occurs with Distant. Similarity tiers thus need not induce the same ordering in both metrics. Figure J.9 shows corresponding outputs.

Table 4: Ablation of ROLT and SCVM on Single717. ROLT only and SCVM only add the respective module to PiD; RefPLC combines both. The same training split and 10 epochs are used; the base Pixel Diffusion network stays frozen. Red bold and blue underlined entries denote the best and second-best recorded values; exact ties share a rank.

Method	PSNR RGB↑	SSIM RGB↑	LPIPS VGG↓	MUSIQ↑	NIQE↓	DeQA↑	MANIQA↑	Q-Align↑	UniPercept IAA↑	UniPercept IQA↑	VisualQuality R1↑
PiD	23.92	0.6796	0.4731	63.87	6.089	3.749	0.2850	4.151	61.90	66.71	4.172
ROLT only	24.73	0.6972	0.4587	66.10	5.632	3.880	0.2920	4.300	64.20	68.90	4.270
SCVM only	25.68	**0.7151**	**0.4382**	65.40	5.840	3.885	0.2820	4.290	63.70	68.20	4.250
RefPLC	**25.81**	0.7079	0.4463	**67.89**	**5.325**	**3.964**	**0.2969**	**4.401**	**65.67**	**70.53**	**4.345**

Contributions of ROLT and SCVM. Table 4 enables ROLT and SCVM as complete modules. ROLT prepares the target condition; SCVM adjusts RGB velocity using the current prediction. Their combination improves PSNR over PiD by 1.89 dB, SSIM by 0.0283, and MUSIQ by 4.02, and reduces LPIPS-VGG by 0.0268. RefPLC records the best PSNR and all eight no-reference values in this comparison. SCVM alone records the highest SSIM and lowest LPIPS-VGG. The four configurations establish the modules' reconstruction and quality profiles, while the zero-evidence and amplitude properties follow from their parameterization.

5.4 Qualitative Results

Figures 4 and 5 show group and adult–child scenes for the nine methods in Table 1, including the adapted PiD baseline. With common Real-ESRGAN LQ inputs, the displayed RefPLC outputs retain the target arrangement and visible face and hair boundaries despite the reference's different composition. Appendix J.1 specifies checkpoints and adds further examples.

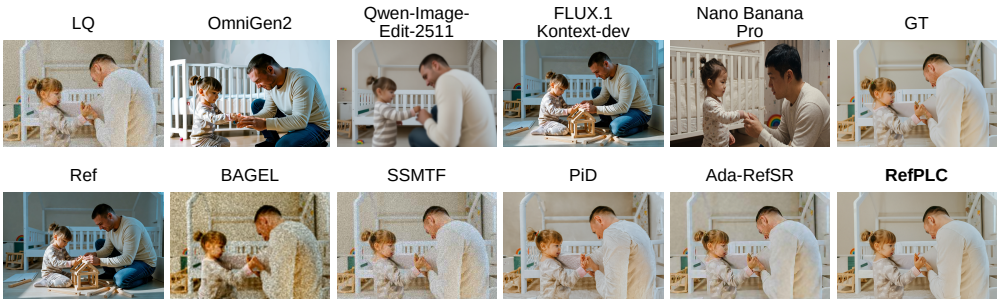

Figure 5: **Adult–child scene.** Complete model comparison showing target composition, adult pose, and child placement.

Figure 6 compares face and bird restoration. In the displayed RefPLC outputs, eye boundaries are visible within the target face, and the bird has a continuous neck contour with localized plumage. Figure J.8 adds an egret and a portrait.

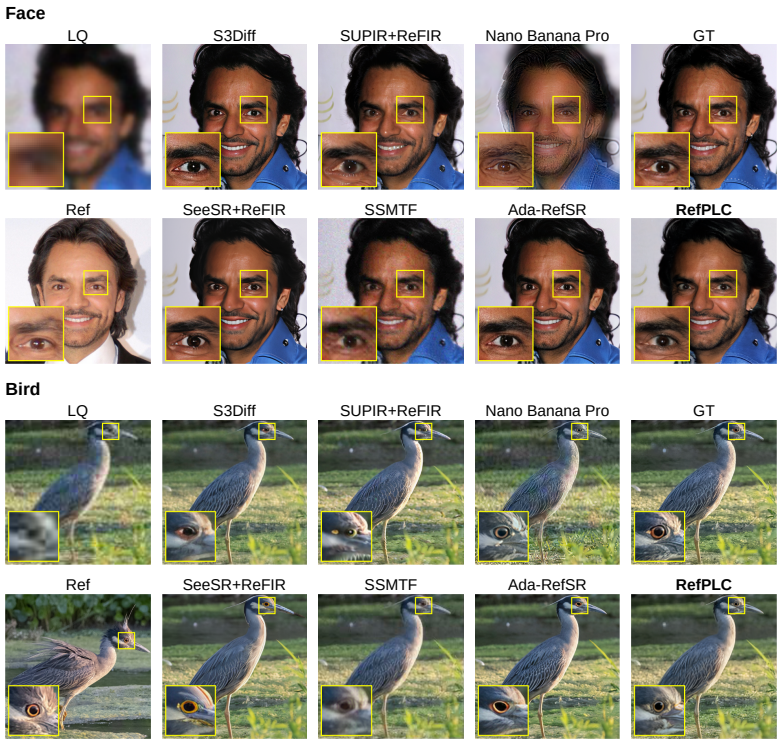

Figure 6: **External qualitative comparison.** Face and bird restoration. Yellow boxes mark the enlarged regions. Additional examples appear in Figure J.8.

6 Conclusion

RefPLC combines target-grid latent recovery with state-conditioned RGB-velocity modulation. ROLT produces a transported condition increment, while SCVM evaluates the additional pixel response to reference evidence within the resulting sampling state. Single717, external-benchmark, and module results characterize the resulting fidelity–quality profile. The paired 90-target study measures the benefit of supplying related references to the same trained system. By separating evidence recovery from state-dependent application, RefPLC connects unaligned reference observations to target-conditioned RGB restoration.

AI USE STATEMENT

Generative AI tools assisted with drafting and translation, literature discovery, method and experiment discussions, mathematical exposition, conceptual figure drafts, table and LaTeX preparation, interpretation of experimental records, and critical review of the text. The authors are responsible for the final scientific content, experimental results, citations, and submitted materials.

ETHICS STATEMENT

The intended application is image restoration. Synthesized details estimate missing visual information and should be distinguished from original measurements. Reference-image use should respect privacy, consent, source attribution, and applicable rights.

REPRODUCIBILITY STATEMENT

Section 4 specifies the two reference interfaces and objective. The supplement provides tensor coordinates, expanded computation, loss definitions, gradient paths, parameter settings, and training and inference algorithms. The source package includes original-precision result tables and a figure manifest. Paired-input, checkpoint, and metric conventions are specified with the corresponding comparisons.

REFERENCES

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. URL https://arxiv.org/abs/2506.15742.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Blau_The_Perception-Distortion_Tradeoff_CVPR_2018_paper.html.

Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *European Conference on Computer Vision*, 2022. URL https://arxiv.org/abs/2207.11938.

Shuo Cao, Jiayang Li, Xiaohui Li, Yuandong Pu, Kaiwen Zhu, Yuanting Gao, Siqi Luo, Yi Xin, Qi Qin, Yu Zhou, Xiangyu Chen, Wenlong Zhang, Bin Fu, Yu Qiao, and Yihao Liu. UniPercept: Towards unified perceptual-level image understanding across aesthetics, quality, structure, and texture. *arXiv preprint arXiv:2512.21675*, 2025. URL https://arxiv.org/abs/2512.21675.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. URL https://arxiv.org/abs/2505.14683.

Runmin Dong, Shuai Yuan, Bin Luo, Mengxuan Chen, Jinxiao Zhang, Lixian Zhang, Weijia Li, Juepeng Zheng, and Haohuan Fu. Building bridges across spatial and temporal resolutions: Reference-based super-resolution via change priors and conditional diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27684–27694, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Dong_Building_Bridges_across_Spatial_and_Temporal_Resolutions_Reference-Based_Super-Resolution_via_CVPR_2024_paper.html.

Hang Guo, Tao Dai, Zhihao Ouyang, Taolin Zhang, Yaohua Zha, Bin Chen, and Shutao Xia. ReFIR: Grounding large restoration models with retrieval augmentation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 46593–46621, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/52a599ccd90f6830ef316c65a60de0a7-Abstract-Conference.html.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.08500.

Yuming Jiang, Kelvin C. K. Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via C^2-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://arxiv.org/abs/2106.01863.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. URL https://cs.stanford.edu/people/jcjohns/eccv16/.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5128–5137, 2021. URL https://arxiv.org/abs/2108.05997.

Byeonghun Lee, Hyunmin Cho, Hong Gyu Choi, Soo Min Kang, Iljun Ahn, and Kyong Hwan Jin. Reference-based super-resolution via image-based retrieval-augmented generation diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10764–10774, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Lee_Reference-based_Super-Resolution_via_Image-based_Retrieval-Augmented_Generation_Diffusion_ICCV_2025_paper.html.

Xiaoming Li, Shiguang Zhang, Shangchen Zhou, Lei Zhang, and Wangmeng Zuo. Learning dual memory dictionaries for blind face restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. URL https://arxiv.org/abs/2210.08160.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Siyu Liu, Zheng-Peng Duan, Jia OuYang, Jiayi Fu, Hyunhee Park, Zikun Liu, Chun-Le Guo, and Chongyi Li. FaceMe: Robust blind face restoration with personal identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. URL https://arxiv.org/abs/2501.05177.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.03003.

Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. MASA-SR: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6368–6377, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Lu_MASA-SR_Matching_Acceleration_and_Spatial_Adaptation_for_Reference-Based_Image_Super-Resolution_CVPR_2021_paper.html.

Yifan Lu, Qi Wu, Jay Zhangjie Wu, Zian Wang, Huan Ling, Sanja Fidler, and Xuanchi Ren. PiD: Fast and high-resolution latent decoding with pixel diffusion. *arXiv preprint arXiv:2605.23902*, 2026. URL https://arxiv.org/abs/2605.23902.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. URL https://live.ece.utexas.edu/research/Quality/niqe_spl.pdf.

Qwen Team. Qwen-Image-Edit-2511. Official model card, version 2511, 2025. URL https://huggingface.co/Qwen/Qwen-Image-Edit-2511.

Naina Raisinghani. Introducing Nano Banana Pro. Google DeepMind, November 2025. URL https://blog.google/innovation-and-ai/products/nano-banana-pro/.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*, 2021. URL https://arxiv.org/abs/2104.07636.

Zhenning Shi, Zizheng Yan, Yuhang Yu, Clara Xue, Jingyu Zhuang, Qi Zhang, Jinwei Chen, Tao Li, and Qingnan Fan. Ultra-high-definition reference-based landmark image super-resolution with generative diffusion prior. *arXiv preprint arXiv:2508.10779*, 2025. URL https://arxiv.org/abs/2508.10779.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1409.1556.

Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-LoRA approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2412.03017.

Ce Wang and Wanjie Sun. Controllable reference guided diffusion with local global fusion for real world remote sensing image super resolution. *arXiv preprint arXiv:2506.23801*, 2025. URL https://arxiv.org/abs/2506.23801.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. URL https://arxiv.org/abs/2207.12396.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1905–1914, 2021. URL https://arxiv.org/abs/2107.10833.

Yuan Wang, Yuhao Wan, Siming Zheng, Bo Li, Qibin Hou, and Peng-Tao Jiang. Trust but verify: Adaptive conditioning for reference-based diffusion super-resolution via implicit reference correlation modeling. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.01864.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13 (4):600–612, 2004. URL https://ece.uwaterloo.ca/~z70wang/publications/ssim.html.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. OmniGen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871v3*, 2025a. URL https://arxiv.org/abs/2506.18871v3.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the International Conference on Machine Learning*, volume 235, pp. 54015–54029, 2024a. URL https://proceedings.mlr.press/v235/wu24ah.html.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *Advances in Neural Information Processing Systems*, 2024b. URL https://arxiv.org/abs/2406.08177.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. SeeSR: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024c. URL https://arxiv.org/abs/2311.16518.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. VisualQuality-R1: Reasoning-induced image quality assessment via reinforcement learning to rank. In *Advances in Neural Information Processing Systems*, 2025b. URL https://papers.nips.cc/paper_files/paper/2025/hash/7f8f7bf2d357a09bdfff04b2e0f5c4e-Abstract-Conference.html.

Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/2006.04139.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1191–1200, 2022. URL https://openaccess.thecvf.com/content/CVPR2022W/NTIRE/html/Yang_MANIQA_Multi-Dimension_Attention_Network_for_No-Reference_Image_Quality_Assessment_CVPRW_2022_paper.html.

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2501.11561.

Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2401.13627.

Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-guided one-step image super-resolution with diffusion priors. *arXiv preprint arXiv:2409.17058*, 2024a. URL https://arxiv.org/abs/2409.17058.

Howard Zhang, Yuval Alaluf, Sizhuo Ma, Achuta Kadambi, Jian Wang, and Kfir Aberman. InstantRestore: Single-step personalized face restoration with shared-image attention. *arXiv preprint arXiv:2412.06753*, 2024b. URL https://arxiv.org/abs/2412.06753.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. URL https://arxiv.org/abs/1801.03924.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, Yandong Guo, and Lei Zhang. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023. URL https://arxiv.org/abs/2306.03514.

Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7982–7991, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Zhang_Image_Super-Resolution_by_Neural_Texture_Transfer_CVPR_2019_paper.html.

Hongyang Zhou, Xiaobin Zhu, Jingyan Qin, Yu Xu, Roberto M. Cesar, Jr., and Xu-Cheng Yin. Multi-scale texture fusion for reference-based image super-resolution: New dataset and solution. *International Journal of Computer Vision*, 133:6971–6992, 2025. URL https://link.springer.com/article/10.1007/s11263-025-02514-3.

Shangchen Zhou, Kelvin C. K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2206.11253.

SUPPLEMENTARY MATERIAL

The supplement expands the eight main-text equations into the complete reference-interaction and learning formulation. Appendix A defines tensor coordinates and shapes; Appendices B and C describe the two modules; Appendices D–F give the time convention, losses, computation order, and fixed settings. Data organization, the full benchmark table, and further visual comparisons follow in Appendices G–J.

A NOTATION AND COORDINATE SYSTEMS

An episode consists of an LQ image, a single reference image, their inference-time captions, and a target used only for supervision or evaluation. For clarity, spatial indices i and j refer to target and reference positions, respectively; batch indices are omitted unless needed. The full target-resolution sampling canvas may include padding. A geometric operator Π removes padding and applies the evaluation-resolution mapping.

Table A.1: Tensor roles. B is batch size, $C_z = 32$, and C_F is the reference-evidence width. Pixel shapes are written before temporary padding.

Quantity	Shape	Meaning
I_L	$B \times 3 \times H \times W$	LQ observation
I_R	$B \times 3 \times H_R \times W_R$	Unaligned reference
z_L	$B \times C_z \times h \times w$	Target latent grid
z_R, z_R^b, d_R	$B \times C_z \times h_R \times w_R$	Reference latent grid
$\hat{r}, \bar{z}_R, u, \Delta z_{L\leftarrow R}$	$B \times C_z \times h \times w$	Transported target-grid tensors
g	$B \times 1 \times h \times w$	Latent gate, channel-broadcast
F_R	$B \times C_F \times h_F \times w_F$	Target-aligned reference evidence
$x_t, v_t^{\mathrm{base}}, p_t, \Delta v_{R,t}$	$B \times 3 \times 4H \times 4W$	RGB states and velocities
S_t	$B \times 12 \times 4H \times 4W$	Channel-concatenated state
α_t	$B \times 1 \times 4H \times 4W$	RGB-broadcast spatial gain

Reference-grid features can differ in size and aspect ratio from target-grid features. Transport maps d_R to $\hat{r}_{L\leftarrow R}$ before it enters the LQ condition. As defined in the figure captions, compact r denotes $\hat{r}_{L\leftarrow R}$, Δz_R or Δz_r denotes $\Delta z_{L\leftarrow R}$, and v_0 denotes the current base velocity v_t^{base}, not sampling time zero. In the SCVM panel, λ denotes ω_i and v denotes v_t. The condition projection is P; parameter groups are named by their modules. Encoder sharing means that all image versions use the same frozen VAE. Decoder sharing means that the evidence and zero-evidence evaluations use the same decoder parameters.

B FULL ROLT FORMULATION

B.1 REFERENCE GEOMETRY AND DETAIL-SENSITIVE VALUES

Let T_R retain the complete reference content under proportional resizing and let $a_j \in [0, 1]$ denote the effective valid area represented by reference-grid location j. The internal reference side limit is $L_R = \min(1024, \max(H, W))$. The internal resize factor is

$$\eta_R = \min\left(1, \frac{L_R}{\max(H_R, W_R)}\right). \tag{B.1}$$

Thus, the internal stage downsizes an oversized reference but does not enlarge an already smaller one. Input-adapter resizing and temporary padding are distinct geometry operations. Masks record which locations contain valid image support.

The reference perturbation uses a normalized 9×9 Gaussian with $\sigma_B = 1.2$:

$$K_B(a, b) = \frac{\exp[-(a^2 + b^2)/(2\sigma_B^2)]}{\sum_{u=-4}^{4} \sum_{v=-4}^{4} \exp[-(u^2 + v^2)/(2\sigma_B^2)]}, \quad -4 \le a, b \le 4. \tag{B.2}$$

Writing $\bar{I}_R = T_R(I_R)$, the shared frozen encoder produces

$$z_L = E(I_L), \qquad z_R = E(\bar{I}_R), \qquad z_R^b = E(B_R(\bar{I}_R)), \qquad d_R = z_R - z_R^b. \tag{B.3}$$

All latent operations use the unpacked 32-channel representation. The same image geometry and encoder are used for the sharp and blurred reference. The difference measures the encoder's response to blur-induced detail attenuation and retains the reference-grid coordinates until transport. Because E is nonlinear, this representation is an encoder-space difference rather than an exact Fourier high-pass projection.

B.2 Coarse correspondence and target-grid transport

Learned descriptors provide queries and keys $q_i^L, k_j^L \in \mathbb{R}^{128}$. Frozen C^2 features provide corresponding descriptors $q_i^{C^2}, k_j^{C^2}$. The combined similarity is

$$s_{ij} = 0.25\langle q_i^L, k_j^L\rangle + 0.75\langle q_i^{C^2}, k_j^{C^2}\rangle. \tag{B.4}$$

This mixture is used to organize correspondences; the transferred values remain d_R and z_R, not the C^2 descriptors.

The coarse stage uses an area-weighted grid with maximum side 16. For side $X \in \{L, R\}$, its descriptor field Q^X, valid areas a^X, and coarse cell $\mathcal{C}_X$, area-weighted aggregation is

$$\mathrm{Pool}_{a^X}(Q^X, \mathcal{C}_X) = \frac{\sum_{j\in\mathcal{C}_X} a_j^X Q_j^X}{\sum_{j\in\mathcal{C}_X} a_j^X}, \qquad \sum_{j\in\mathcal{C}_X} a_j^X > 0. \tag{B.5}$$

Cells with no valid support are excluded. Coarse correspondence probabilities P_{ab}^c define the selected reference cell $b^*(a) = \arg\max_b P_{ab}^c$. A target location i in cell $a(i)$ searches the corresponding fine-grid neighborhood $\mathcal{N}_i$ around $b^*(a(i))$.

Let ℓ_{ij} denote the fine logit incorporating the descriptor similarity, coarse correspondence prior, and effective spatial support. The normalized valid-window affinities and transported tensors are

$$\begin{aligned} A_{ij} &= \frac{\mathbf{1}[j \in \mathcal{N}_i,\ a_j > 0]\exp(\ell_{ij})}{\sum_{k\in\mathcal{N}_i,\ a_k>0}\exp(\ell_{ik})}, \\ \hat{r}_{L\leftarrow R,i} &= \sum_j A_{ij} d_{R,j}, \qquad \bar{z}_{R,i} = \sum_j A_{ij} z_{R,j}. \end{aligned} \tag{B.6}$$

The expression applies to a nonempty valid window; empty-support locations produce zero evidence. Coarse localization is discrete, while the executed fine aggregation is soft. Both transported tensors have the target latent shape, so the subsequent mapping does not assume aligned input images.

Matching concentration provides additional gate information. For example, the peak and entropy of the normalized distribution are

$$\mathrm{peak}_i = \max_j A_{ij}, \qquad \mathrm{ent}_i = -\sum_{j:A_{ij}>0} A_{ij}\log A_{ij}. \tag{B.7}$$

The gate combines learned descriptors, peak, entropy, maximum similarity, and the relationship of z_L to $\bar{z}_R$. A high similarity alone therefore does not prescribe the final latent update strength.

B.3 Zero-anchored mapping and RMS constraint

The location-wise scale is the detached channel RMS, with a numerical floor:

$$s_i = \mathrm{sg}\left[\max\left(\sqrt{\frac{1}{C_z}\sum_{c=1}^{C_z} z_{L,ci}^2}, \varepsilon_s\right)\right]. \tag{B.8}$$

The local map H consists of three 3×3 convolutions. Its target-conditioned reference contribution includes the direct transported-value skip:

$$u_i = \hat{r}_i + 0.25 s_i \tanh[H(z_L/s, \hat{r}/s)_i - H(z_L/s, \mathbf{0})_i]\,. \tag{B.9}$$

The two evaluations share parameters and target normalization. The gated candidate is $a = 0.5g \odot u$. For a sample tensor X containing N entries, define

$$\mathrm{RMS}(X) = \sqrt{\frac{1}{N}\sum_{n=1}^{N} X_n^2}, \qquad \mathcal{P}_{z_L}(a) = a \min\left(1, \frac{0.15\,\mathrm{RMS}(z_L)}{\mathrm{RMS}(a) + \varepsilon}\right). \tag{B.10}$$

The reduction is over the channels and spatial positions of each image, independently across the batch. The complete update is

$$\Delta z_{L\leftarrow R} = \mathcal{P}_{z_L}(0.5g \odot u), \qquad \widetilde{z}_L = z_L + \Delta z_{L\leftarrow R}. \tag{B.11}$$

All reference-derived paths, including the direct skip, enter the gate and the RMS projection.

Latent-amplitude property. For $A = \mathrm{RMS}(a)$ and $Z = \mathrm{RMS}(z_L)$,

$$\mathrm{RMS}(\Delta z_{L\leftarrow R}) = A \min\left(1, \frac{0.15Z}{A + \varepsilon}\right) \le 0.15Z. \tag{B.12}$$

When the scale factor equals one, its selection implies $A + \varepsilon \le 0.15Z$; otherwise the inequality follows from $A/(A + \varepsilon) \le 1$. This property bounds the RMS amplitude added at the condition interface. Squaring both sides gives the corresponding mean-square bound.

Zero-evidence property. If $\hat{r} = \mathbf{0}$, the two map evaluations coincide and the direct skip vanishes. Hence $u = 0$, $a = 0$, and $\Delta z_{L\leftarrow R} = 0$. The model then follows its own base-conditioning path. When that conditioning path has been adapted during learning, zero reference does not imply equality to an unrelated, pre-adaptation checkpoint.

C Full SCVM Formulation

C.1 Visual evidence and global–local fusion

The visual values combine reference RGB detail with differences in a fixed VGG representation (Simonyan & Zisserman, 2015). Let ϕ_ℓ denote a VGG feature field and $\mathcal{A}_\ell$ its alignment to the evidence grid. The available value components are

$$\begin{aligned} d_R^{\mathrm{rgb}} &= \bar{I}_R - B_R^{\mathrm{pix}}(\bar{I}_R), \\ d_R^{\phi,\ell} &= \mathcal{A}_\ell\left[\phi_\ell(\bar{I}_R) - \phi_\ell(B_R^{\mathrm{pix}}(\bar{I}_R))\right], \\ \nu_j &= \mathrm{Pack}\left(d_{R,j}^{\mathrm{rgb}}, \{d_{R,j}^{\phi,\ell}\}_\ell\right). \end{aligned} \tag{C.1}$$

Here Pack collects spatially aligned RGB and perceptual residual channels into the common evidence representation, and B_R^{pix} is its low-pass reference operator. It need not be identified with the latent-branch perturbation B_R. Both correspondence paths operate on this same value representation.

For dense affinities A^G over the reference bank and local affinities A^L in a 5×5 window around the coarse correspondence,

$$f_i^G = \sum_{j\in\Omega_R} A_{ij}^G \nu_j, \qquad f_i^L = \sum_{j\in\mathcal{N}_i^{5\times5}} A_{ij}^L \nu_j. \tag{C.2}$$

Both sets of weights are normalized over their respective valid candidate supports. Let m_i denote the statistics supplied by the matching stage. Their fusion is

$$\omega_i = 0.75\,\mathrm{sigmoid}(M(m_i)), \qquad f_i = (1 - \omega_i) f_i^G + \omega_i f_i^L. \tag{C.3}$$

Since $0 \le \omega_i \le 0.75$, the dense contribution is at least 0.25. The fusion is a convex combination of evidence vectors at the same target location, not a blend of independently generated SR images.

C.2 CAPTION PATH AND EVIDENCE RELIABILITY

Captions are generated from the actual LQ and reference images:

$$c_L = C_{\text{DAPE-RAM}}(I_L), \qquad c_R = C_{\text{RAM}}(I_R). \tag{C.4}$$

The caption sources use RAM (Zhang et al., 2023) and its degradation-aware adaptation in SeeSR (Wu et al., 2024c). The LQ caption enters PiD's text-conditioning path. SCVM also derives a bounded reliability signal from the caption relation, denoted $r_{\text{text}} = h_{\text{text}}(c_L, c_R)$. The evidence passed to the pixel decoder is

$$F_R = \mathcal{R}(f; r_{\text{text}}), \qquad r_{\text{text}} \in \mathcal{B}_{\text{text}}, \tag{C.5}$$

where $\mathcal{B}_{\text{text}}$ is the bounded reliability range and $\mathcal{R}$ applies that signal to visual evidence. The text signal and local-fusion weight ω have different roles: ω selects between correspondence ranges, whereas the caption relation modulates reliability of the resulting visual evidence. The Ref caption does not enter ROLT or replace c_L in the base generator. Once computed, captions and the image-pair evidence are reused across the sampling trajectory.

C.3 SHARED-DECODER DIFFERENCE AND SPATIAL GAIN

Let $\mathcal{U}(I_L)$ be the LQ image resampled to the sampling canvas and expressed in the model's RGB range. The complete state is

$$S_t = \text{Concat}_{\text{ch}}\left[x_t, v_t^{\text{base}}, \hat{y}_t^{\text{base}}, \mathcal{U}(I_L)\right], \qquad \hat{y}_t^{\text{base}} = x_t - t v_t^{\text{base}}. \tag{C.6}$$

Each component has three channels, giving 12 channels in total. The base velocity already uses $\widetilde{z}_L$. The decoder's true-evidence and zero-evidence evaluations therefore share not only their weights but also the same ROLT-informed sampling state:

$$\begin{aligned} d_t &= D(S_t, t, F_R) - D(S_t, t, \mathbf{0}_{F_R}), \\ p_t &= 0.5 \tanh(d_t / 0.5). \end{aligned} \tag{C.7}$$

The zero tensor matches F_R in shape. The subtraction measures the change of the pixel decoder's response when this evidence is present, holding its other inputs fixed.

Matching statistics are first mapped from the evidence grid to the RGB sampling grid by the spatial alignment operator $\mathcal{U}_m$, preserving their four channels. The gain input then concatenates 12 state channels, three detached proposal channels, four aligned statistic channels, and one broadcast timestep, where $\mathbf{1}_{4H\times 4W}$ is a spatial all-ones field:

$$\widetilde{m} = \mathcal{U}_m(m), \qquad J_t = \text{Concat}_{\text{ch}}[S_t, \text{sg}(p_t), \text{sg}(\widetilde{m}), t\,\mathbf{1}_{4H\times 4W}], \qquad J_t \in \mathbb{R}^{B\times 20\times 4H\times 4W}. \tag{C.8}$$

The spatial gain is computed as

$$\begin{aligned} J_t^{\downarrow} &= \text{Pool}_{1/8}(J_t), \\ \alpha_t^{\downarrow} &= \text{sigmoid}\Big(\text{Conv}_{24\to 1}^{1\times 1}\left[\text{SiLU}(\text{Conv}_{20\to 24}^{3\times 3}(J_t^{\downarrow}))\right]\Big), \\ \alpha_t &= \text{BilinearUp}(\alpha_t^{\downarrow}), \qquad \Delta v_{R,t} = \alpha_t \odot p_t. \end{aligned} \tag{C.9}$$

The one-channel gain is broadcast over RGB. Pooling and bilinear interpolation provide a coarse spatial control field, while the proposal retains output-resolution detail. Detaching the proposal and statistics at the gain input leaves the direct proposal path differentiable through the final product. It prevents the gain-input branch from supplying an additional gradient to those input tensors.

Pixel-amplitude and null properties. The proposal satisfies $\|p_t\|_\infty \le 0.5$, and bilinear interpolation of the sigmoid output preserves $0 \le \alpha_t \le 1$. Therefore,

$$\|v_t - v_t^{\text{base}}\|_\infty = \|0.25\alpha_t \odot p_t\|_\infty \le 0.125. \tag{C.10}$$

This is a pointwise bound on the additive velocity correction in model units. If $F_R = 0$, the shared-decoder difference is zero, as is the pixel correction. In the experiment, Null denotes absence of a reference image at the inference interface. The zero-evidence identity describes the module tensors; no constant reference image is introduced for this control.

D Flow Coordinates and Sampling

D.1 Noise time and clean time

Flow matching and rectified flow learn vector fields along paths connecting data and noise (Lipman et al., 2023; Liu et al., 2023). We use noise time to make the clean endpoint explicit. Given a clean sampling-canvas image y and independent Gaussian noise ϵ,

$$x_t = (1-t)y + t\epsilon, \qquad \frac{\partial x_t}{\partial t} = \epsilon - y, \qquad x_t - t(\epsilon - y) = y. \tag{D.1}$$

Thus sampling starts at $x_1 = \epsilon$ and ends at $x_0 = y$. The current clean estimate under a predicted velocity is $\hat{y}_t = x_t - tv_t$.

PiD's published exposition uses the equivalent clean-time coordinate τ (Lu et al., 2026). Under $\tau = 1 - t$,

$$\tilde{x}_\tau = \tau y + (1-\tau)\epsilon = x_{1-\tau}, \qquad w_\tau = \frac{\partial \tilde{x}_\tau}{\partial \tau} = y - \epsilon = -v_{1-\tau}. \tag{D.2}$$

The clean estimate is invariant to this change:

$$x_t - tv_t = \tilde{x}_\tau + (1-\tau)w_\tau. \tag{D.3}$$

Consequently, a base predictor represented in clean time corresponds mathematically to $f_t(x, t; c) = -f_\tau(x, 1-t; c)$ in noise time. This relation specifies the notation at the wrapper interface; it is not an additional network module or a change to the generated trajectory.

D.2 Sampling update and cached evidence

Let $1 = t_0 > t_1 > \cdots > t_K = 0$ be the sampler's time grid, with $K = 25$ for the evaluation configuration. Denote the PiD sampler update at step k by Ψ_k. RefPLC supplies the full vector field to that update:

$$x_{t_{k+1}} = \Psi_k(x_{t_k}, v(\cdot,\cdot), t_k, t_{k+1}), \qquad v(x,t) = v^{\text{base}}(x,t) + 0.25\alpha(x,t) \odot p(x,t). \tag{D.4}$$

Every velocity evaluation, including an internal solver stage, recomputes the corresponding state tensor, proposal, and gain. For an explicit single-evaluation integration step, the coordinate relation reduces to

$$x_{t_{k+1}} = x_{t_k} + (t_{k+1} - t_k)v(x_{t_k}, t_k). \tag{D.5}$$

Equation D.5 makes the time-direction sign explicit; Equation D.4 accommodates the configured PiD integration routine. The reference correction coefficient scales only the extra velocity. The backbone's guidance setting, here scale 5, belongs to the base-prediction interface and is distinct from that coefficient.

Latent transport, static correspondence evidence, and captions are computed once for a target–reference pair. The state tensor, decoder difference, proposal, and gain are recomputed whenever the sampler evaluates the vector field. This separation permits reuse of image-pair calculations without replacing the state-dependent correction by a fixed image residual.

E Learning Objectives and Gradient Routing

E.1 Primary recovery target

The primary loss operates on the composed velocity and the one-step clean estimate, with the geometry map Π applied at supervision resolution. For $t > 0$,

$$\widehat{I}_t = \Pi(x_t - tv_t), \qquad v_t^* = \frac{\Pi(x_t) - I_{GT}}{t}, \qquad \mathcal{L}_{\text{vel}} = \mathbb{E}\left[\|\Pi(v_t) - v_t^*\|_2^2\right]. \tag{E.1}$$

Cropping and fixed image resampling act linearly on the relevant image fields. The identity $\Pi(x_t) - tv_t^* = I_{GT}$ aligns the supervised velocity with the GT endpoint. On an unpadded matching-resolution canvas, this reduces to the usual target $\epsilon - y$.

Algorithm D.1 RefPLC inference with a supplied reference

Require: LQ I_L, reference I_R, time grid $\{t_k\}_{k=0}^{K}$, sampling operators $\{\Psi_k\}$
1: Generate $c_L = C_L(I_L)$ and $c_R = C_R(I_R)$; prepare image geometry and valid masks.
2: Encode z_L, z_R, z_R^b with the shared frozen VAE; set $d_R = z_R - z_R^b$.
3: Compute coarse-to-fine correspondences and transport d_R, z_R to $\hat{r}, \bar{z}_R$.
4: Evaluate the zero-anchored map, confidence gate, and RMS cap; set $\widetilde{z}_L = z_L + \Delta z_{L\leftarrow R}$.
5: Prepare dense/local reference values, matching statistics m, and caption-regulated evidence F_R.
6: Draw $x_{t_0} \sim \mathcal{N}(0, I)$ at the output sampling resolution.
7: **for** $k = 0, \ldots, K-1$ **do**
8: Compute the guided base velocity using $P(\widetilde{z}_L)$ and c_L.
9: Form $S_{t_k} = [x_{t_k}, v_{t_k}^{\text{base}}, x_{t_k} - t_k v_{t_k}^{\text{base}}, \mathcal{U}(I_L)]$.
10: Evaluate the shared decoder with F_R and 0; obtain p_{t_k} via scaled $\tanh$.
11: Predict the coarse gain and upsample it to α_{t_k}.
12: Supply $v_{t_k} = v_{t_k}^{\text{base}} + 0.25\alpha_{t_k} \odot p_{t_k}$ to Ψ_k and advance the sample.
13: **end for**
14: Remove padding and return the restored RGB image.

The image reconstruction term is the entrywise mean of the Charbonnier penalty:

$$\mathcal{L}_{\text{Charb}} = \mathbb{E}\left[\frac{1}{N_I}\sum_{n=1}^{N_I}\sqrt{(\widehat{I}_{t,n} - I_{GT,n})^2 + 10^{-6}}\right]. \tag{E.2}$$

Horizontal and vertical first-difference operators ∇_x and ∇_y form the edge comparison. With ℓ_{edge} denoting the edge-distance criterion and ℓ_{feat} the fixed perceptual-feature criterion, the conventional structural terms are written as

$$\begin{aligned}
\mathcal{L}_{\text{edge}} &= \mathbb{E}\left[\sum_{d\in\{x,y\}} \ell_{\text{edge}}(\nabla_d \widehat{I}_t, \nabla_d I_{GT})\right], \\
\mathcal{L}_{\text{VGG}} &= \mathbb{E}\left[\ell_{\text{feat}}(\Phi(\widehat{I}_t), \Phi(I_{GT}))\right], \\
\mathcal{L}_{\text{base}} &= \mathcal{L}_{\text{vel}} + 0.10\mathcal{L}_{\text{Charb}} + 0.03\mathcal{L}_{\text{edge}} + 0.02\mathcal{L}_{\text{VGG}}.
\end{aligned} \tag{E.3}$$

Φ is the fixed VGG representation (Simonyan & Zisserman, 2015). Feature-space supervision follows perceptual-loss formulations for image transformation and super-resolution (Johnson et al., 2016). These recovery criteria support the final image objective. Supervision is evaluated at the actual GT resolution; a resized GT canvas can supply the diffusion carrier without replacing the GT used for scoring the clean estimate.

E.2 Geometry supervision on the executed correspondence hierarchy

Training references with known full-image or crop relationships provide correspondence targets Y^c, Y^f and confidence targets Y^q. Let P^c, P^f, q be the predictions, and let Ω_c, Ω_q denote valid supervision locations. Let j_i^* be the ground-truth reference-grid correspondence of target position i under the known training relation. The fine supervision mask reflects the window selected by the network:

$$\chi_i = \mathbf{1}[i \text{ is valid}]\,\mathbf{1}[j_i^* \in \mathcal{N}_i], \qquad N_f = \sum_i \chi_i. \tag{E.4}$$

For the coarse-correspondence, fine-correspondence, and confidence criteria ℓ_c, ℓ_f, and ℓ_q, respectively, the objective is

$$\begin{aligned}
\mathcal{L}_c &= \mathbb{E}\left[\frac{1}{|\Omega_c|}\sum_{a\in\Omega_c}\ell_c(P_a^c, Y_a^c)\right],\\
\mathcal{L}_f &= \mathbb{E}\left[\frac{1}{\max(1,N_f)}\sum_i \chi_i \ell_f(P_i^f, Y_i^f)\right],\\
\mathcal{L}_q &= \mathbb{E}\left[\frac{1}{|\Omega_q|}\sum_{i\in\Omega_q}\ell_q(q_i, Y_i^q)\right],\\
\mathcal{L}_{\text{geo}} &= \mathcal{L}_c + \mathcal{L}_f + \mathcal{L}_q.
\end{aligned} \tag{E.5}$$

When no fine target enters an executed window, the corresponding masked fine contribution is zero. This objective trains the matching hierarchy without substituting a ground-truth reference or correspondence into ordinary inference.

E.3 Latent recovery and gate utility target

Let z_C be the shared-VAE encoding of the clean target resampled to the LQ scale, and let Ω_L exclude padded target-latent positions. The latent objective combines recovery error and gate supervision:

$$\begin{aligned}
\mathcal{L}_{\text{value}} &= \mathbb{E}\left[\frac{1}{|\Omega_L|}\sum_{i\in\Omega_L}\left\|\frac{\widetilde{z}_L(i) - z_C(i)}{s_i}\right\|_2^2\right],\\
\mathcal{L}_{\text{gate}} &= \mathbb{E}\left[\frac{1}{|\Omega_L|}\sum_{i\in\Omega_L}\mathrm{BCE}(g_i, g_i^*)\right],\\
\mathcal{L}_{\text{lat}} &= \mathcal{L}_{\text{value}} + 0.05\mathcal{L}_{\text{gate}}.
\end{aligned} \tag{E.6}$$

The label depends on the target latent residual $e_i = z_C(i) - z_L(i)$ and ungated candidate $b_i = 0.5u_i$:

$$g_i^* = \mathrm{sg}\left[\mathrm{clip}_{[0,1]}\left(\frac{\langle e_i, b_i\rangle}{\|b_i\|_2^2 + \varepsilon}\right)\right]. \tag{E.7}$$

The projection has a direct least-squares interpretation. For the scalar regularized fit $\min_{0\le a\le 1}\|e_i - ab_i\|_2^2 + \varepsilon a^2$, the stationary point is

$$a_i^{\text{unconstrained}} = \frac{\langle e_i, b_i\rangle}{\|b_i\|_2^2 + \varepsilon}; \qquad a_i^* = \mathrm{clip}_{[0,1]}(a_i^{\text{unconstrained}}). \tag{E.8}$$

Thus g_i^* represents a location-wise utility proxy before image-level RMS limiting. The stop-gradient prevents the target construction from becoming another optimization path through the candidate direction. The proxy supervises latent recovery strength and is distinct from a label of final image-quality improvement.

E.4 Total objective and trainable paths

The complete objective uses fixed auxiliary coefficients:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + 0.05\mathcal{L}_{\text{geo}} + 0.05\mathcal{L}_{\text{lat}}. \tag{E.9}$$

Let θ_{ROLT}, θ_{SCVM}, and θ_{cond} denote ROLT, trainable SCVM, and selected PiD conditioning parameters. The gradient paths are

$$\begin{aligned}
&\nabla_{\theta_{\text{ROLT}}}\mathcal{L} = \nabla_{\theta_{\text{ROLT}}}\big(\mathcal{L}_{\text{base}} + 0.05\mathcal{L}_{\text{geo}} + 0.05\mathcal{L}_{\text{lat}}\big),\\
&\nabla_{\theta_{\text{SCVM}}}\mathcal{L} = \nabla_{\theta_{\text{SCVM}}}\mathcal{L}_{\text{base}}, \qquad \nabla_{\theta_{\text{cond}}}\mathcal{L} = \nabla_{\theta_{\text{cond}}}\mathcal{L}_{\text{base}}.
\end{aligned} \tag{E.10}$$

The frozen backbone and feature extractors receive no parameter updates. The primary training signal uses a sampled state and a one-step clean estimate, rather than requiring a full differentiable rollout.

Algorithm E.1 Reference-coupled learning at one sampled noise time

Require: Training tuple (I_L, I_R, I_{GT}), captions from the input images, known-relation auxiliary pair

1: Encode image latents and prepare ROLT/SCVM image-pair evidence.
2: Sample noise ϵ and $t > 0$ from the training time distribution; form $x_t = (1-t)y + t\epsilon$.
3: Compute $\widetilde{z}_L$, v_t^{base}, S_t, p_t, and α_t in the same order as inference.
4: Form $v_t = v_t^{\text{base}} + 0.25\alpha_t \odot p_t$ and $\widehat{I}_t = \Pi(x_t - tv_t)$.
5: Evaluate the velocity, image, edge, and fixed perceptual recovery criteria.
6: Evaluate coarse/fine/confidence geometry supervision with the executed-window mask.
7: Form detached scales and gate targets; evaluate latent recovery and gate losses.
8: Backpropagate Equation E.9 along the parameter paths in Equation E.10.

Table F.1: Reference-module settings and parameter groups.

Component	Setting
Shared VAE latent width	32 unpacked channels; encoder frozen
ROLT descriptors	128 learned dimensions plus frozen C^2 features
Descriptor similarity mixture	0.25 learned, 0.75 C^2
Reference perturbation for ROLT	Gaussian 9×9, $\sigma_B = 1.2$
Coarse correspondence grid	Maximum side 16; valid-area weighting
Zero-anchored local map	Three 3×3 convolutions; residual coefficient 0.25
Pre-cap latent scaling	0.5 times the confidence-gated candidate
Latent correction cap	0.15 times LQ latent RMS per image
SCVM local window	5×5 fine-grid neighborhood
Local evidence weight	$0.75\,\mathrm{sigmoid}(M(m))$
RGB proposal bound	$p_t = 0.5\tanh(d_t/0.5)$
Gain network	Pool to $1/8$; Conv $20 \to 24$, SiLU, Conv $24 \to 1$, sigmoid, bilinear upsample
Reference velocity coefficient	$\kappa = 0.25$
Evaluation sampling configuration	25 steps; guidance scale 5

F Implementation Settings and Computation

All adapted configurations are trained for 10 epochs on the 97,589-pair internal training split. The base learning rates are 10^{-5} for ROLT, 2×10^{-6} for trainable SCVM parameters, and 10^{-7} for selected PiD condition parameters. PiD-only fine-tuning uses the same learning rate for its conditioning parameters as RefPLC (10^{-7}); the base Pixel Diffusion network is not updated. The selected PiD projection-output heads, condition heads, and content gates remain fixed for the first 100 updates, then follow their learning schedule. Other PiD weights, the VAE, text encoders, and frozen matching and perceptual networks remain fixed. The selected trainable PiD condition group contains 12,619,784 parameters; this number excludes reference-module parameters and is not the total model size.

Cached versus step-dependent work. Shared-VAE encoding, ROLT, reference feature extraction, and static correspondence/evidence preparation are image-pair computations. Each vector-field evaluation requires the base prediction, two shared-decoder evaluations, and one gain-head evaluation. Let N_{FE} count evaluations of the composed vector field, separately from the number of sampler intervals K. Then

$$C_{\text{total}} = C_{\text{pair}} + N_{\text{FE}}\big(C_{\text{base}} + 2C_D + C_G\big) + C_{\text{sampler}}. \tag{F.1}$$

This expression separates reuse from repeated work without identifying operation counts with measured wall-clock time. A materialized dense correspondence matrix over N_L and N_R positions has $N_L N_R$ entries and matching arithmetic of order $N_L N_R d$. The SCVM local aggregation uses at most 25 reference positions per target location. The choice of implementation and chunking controls peak working memory.

G Evaluation Protocols and Complete Numerical Results

G.1 Data organization

The training split contains 97,589 GT–Ref pairs corresponding to 64,427 distinct GT targets: 34,109 people images, 3,909 product images, 9,828 scene-text images, and 16,581 video-frame targets. Each non-video target is paired with one reference. Each video-frame target is paired separately with Top, Medium, and Distant references, yielding 49,743 video-based training pairs. Each training example uses one reference. Single717 is the common evaluation set for the cross-method comparison and the four component configurations. The reference-condition study uses 90 distinct GT targets, each evaluated with three related references and without a reference. Scores from these two evaluation tasks are reported separately.

Table G.1: Single717 composition. Video frames use the Top reference, and each target contributes one pair.

Content	GT targets	Fraction
People	300	41.84%
Products	145	20.22%
Scene text	122	17.02%
Video frames	150	20.92%
Total	717	100%

Table G.2: Native $4\times$ restoration canvases. All dimensions are width $\times$ height.

LQ size	Target size	Aspect ratio
672×384	2688×1536	$7:4$
672×448	2688×1792	$3:2$
576×432	2304×1728	$4:3$
512×512	2048×2048	$1:1$
432×576	1728×2304	$3:4$
448×672	1792×2688	$2:3$
384×672	1536×2688	$4:7$

G.2 Weights, training duration, and inference

For all Single717 quantitative and qualitative results, SSMTF and Ada-RefSR use checkpoints fine-tuned on the same internal training set as RefPLC for 10 epochs and the same number of optimization updates. The component study compares PiD, ROLT only, SCVM only, and RefPLC using the same training split, duration, and Single717 evaluation. PiD contains neither reference module. Its conditioning pathway is fine-tuned for 10 epochs on the same training data and with the same learning rate as the corresponding PiD parameters in RefPLC, while the pretrained Pixel Diffusion network is frozen. ROLT only enhances the latent condition without SCVM velocity correction; SCVM only retains the unmodified latent condition and applies state-conditioned velocity modulation. RefPLC combines both modules. Their ordered (ROLT, SCVM) indicators are $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(1, 1)$, respectively. The same adapted PiD results appear in the cross-method and component tables. The individual-module configurations appear only in the component comparison.

All entries in the Single717 main table use a common LQ input for each target. These inputs are obtained with the official Real-ESRGAN degradation procedure. For the external benchmarks, SSMTF and Ada-RefSR each have two configurations: non-FT rows use their official released weights, while FT rows use checkpoints fine-tuned on the RefPLC training set for 10 epochs with the same number of optimization updates as RefPLC. Sampling follows each method's official inference pipeline. S1 and S50 denote one and fifty inference steps. For diffusion methods these are denoising steps; for a feed-forward method S1 denotes one restoration pass. These labels are separate from epochs or optimization updates. Adaptation duration is specified in epochs, and model-specific inference settings are preserved. Rows marked P reproduce the numbers published in the Ada-RefSR comparison (Wang et al., 2026); they are not new measurements by us. Every unmarked external row is evaluated from the official benchmark LQ images, without a method-specific regeneration of the degradation. Weight adaptation and inference-step counts are separate attributes. Retaining each

official inference procedure gives a configuration-level quality comparison, not a matched-compute experiment.

G.3 General-purpose generation and editing baselines

The Single717 comparison includes general-purpose models alongside specialized restoration systems. Table G.3 gives the publication or versioned model source for each named baseline. FLUX.1 Kontext studies in-context generation and editing, BAGEL and OmniGen2 study unified multimodal generation, and Qwen-Image-Edit-2511 and Nano Banana Pro supply image-editing interfaces. These citations identify the models; Table 1 reports their Single717 results, and Nano Banana Pro also appears in the four external-benchmark comparisons.

Table G.3: Sources for the general-purpose baselines in the Single717 comparison. Versioned sources are retained where the evaluated name identifies a particular release.

Model	Primary source
FLUX.1 Kontext	In-context image generation and editing with flow matching (Black Forest Labs et al., 2025).
BAGEL	Unified multimodal pretraining (Deng et al., 2025).
Qwen-Image-Edit-2511	Official model card for release 2511 (Qwen Team, 2025).
Nano Banana Pro	Official Google DeepMind model introduction (Raisinghani, 2025).
OmniGen2	Multimodal generation report, version 3 (Wu et al., 2025a).

G.4 Metric families and ranking

Single717 uses RGB PSNR/SSIM (Wang et al., 2004) and LPIPS-VGG (Zhang et al., 2018). Its eight no-reference measures are MUSIQ, NIQE, DeQA, MANIQA, Q-Align, UniPercept-IAA, UniPercept-IQA, and VisualQuality-R1 (Ke et al., 2021; Mittal et al., 2013; You et al., 2025; Yang et al., 2022; Wu et al., 2024a; Cao et al., 2025; Wu et al., 2025b). External PSNR/SSIM use the reported Y-channel convention; RefPLC's paired perceptual records use LPIPS-Alex. CLIPIQA measures no-reference quality using language–image representations (Wang et al., 2023). The released Ada-RefSR evaluator instantiates `pyiqa.create_metric('lpips')`, whose default is AlexNet, version 0.1, and crops a four-pixel border before paired scoring.[1] These source definitions distinguish the external metric family from Single717's LPIPS-VGG. Official LQ inputs alone do not establish identical evaluator versions, cropping, or reductions for every archived result. External LPIPS values are therefore retained as reported, without cross-source rank annotations or rank-based claims. FID is computed over output and GT sets, not as an average of individual-image FIDs (Heusel et al., 2017). The reference-condition table reports its paired PSNR and LPIPS records.

Rankings use the original scores and each metric's stated direction. Multi-method tables mark first place in red bold and second place in blue with an underline. Exact ties share a rank. The two-method reference-condition tables instead use black bold for the better value. The main external table ranks the common method set, while the full tables rank each dataset's complete set, excluding cross-source LPIPS ranking. These are recorded point-estimate rankings, not significance tests. The no-reference measures describe separate quality criteria rather than independent replications. Tables G.4–G.6 retain the precision used for differences in the text.

Fidelity and perceptual quality. Pixel agreement, paired feature similarity, and the naturalness of a restored image measure different aspects of the task. The perception–distortion tradeoff provides a general reason to examine these objectives jointly (Blau & Michaeli, 2018); it does not turn the eleven reported scores into a single common scale. We therefore retain the individual measures and report concrete changes, such as the full model's 1.89 dB PSNR increase over PiD and its 4.02 MUSIQ increase. The component comparison distinguishes the additional latent and pixel interfaces using the same evaluation targets. The reference-condition comparison instead changes the available reference for a fixed trained model and fixed target input.

[1] Official evaluation script: `https://github.com/vivoCameraResearch/AdaRefSR/blob/main/main_code/test/cufed5/cal_iqa.py`; metric configuration: `https://github.com/chaofengc/IQA-PyTorch/blob/main/pyiqa/default_model_configs.py`.

Scope of the comparisons. The cross-method tables compare complete systems under their specified weights and inference procedures. The four Single717 module configurations evaluate enabling ROLT and SCVM as complete branches; they do not isolate the contributions of each gate, correspondence feature, or caption operation. The reference study holds the trained model fixed and varies reference availability. Distant is a related reference selected from a triplet, whereas Null omits the reference image. Neither is a random or adversarial-reference test. Pixel, feature, and distribution scores characterize output quality; they do not label the origin of individual synthesized details. The experiments concern the stated synthetic degradation protocol.

G.5 Exact Single717 and component records

Table G.4: Original-precision Single717 results on common LQ inputs. SSMTF/Ada-RefSR use RefPLC's fine-tuning data and updates. PiD uses the same training split and 10-epoch condition adaptation, with Pixel Diffusion frozen. Red bold and blue underlined entries denote the best and second-best recorded values; exact ties share a rank.

Method	PSNR RGB↑	SSIM RGB↑	LPIPS VGG↓	MUSIQ↑	NIQE↓	DeQA↑	MANIQA↑	Q-Align↑	UniPercept IAA↑	UniPercept IQA↑	VisualQuality R1↑
FLUX.1 Kontext	15.224	0.553	0.589	67.416	6.031	3.720	0.314	4.140	60.562	66.558	4.116
BAGEL	16.818	0.594	0.639	57.275	9.141	2.589	0.226	2.659	44.152	42.582	2.533
Qwen-Image-Edit-2511	16.24	0.5853	0.5406	63.74	6.799	3.806	0.2953	4.273	61.01	65.70	4.203
Nano Banana Pro	17.14	0.5829	0.5550	67.02	5.669	**3.981**	**0.3227**	4.392	64.45	70.30	**4.405**
OmniGen2	11.24	0.4549	0.6513	66.12	6.895	3.866	0.3124	4.335	60.37	68.69	4.268
PiD	23.92	0.6796	0.4731	63.87	6.089	3.749	0.2850	4.151	61.90	66.71	4.172
SSMTF	**25.99**	0.6984	0.5916	54.06	9.313	2.356	0.2373	2.425	48.97	40.74	2.329
Ada-RefSR	25.52	**0.7575**	**0.4319**	60.94	6.218	3.261	0.2998	3.539	53.18	54.95	3.341
RefPLC	25.81	0.7079	0.4463	**67.89**	**5.325**	3.964	0.2969	**4.401**	**65.67**	**70.53**	4.345

Table G.5: Original-precision Single717 ablation. ROLT only and SCVM only enable one reference module; RefPLC enables both. Red bold and blue underlined entries denote the best and second-best recorded values; exact ties share a rank.

Method	PSNR RGB↑	SSIM RGB↑	LPIPS VGG↓	MUSIQ↑	NIQE↓	DeQA↑	MANIQA↑	Q-Align↑	UniPercept IAA↑	UniPercept IQA↑	VisualQuality R1↑
PiD	23.92	0.6796	0.4731	63.87	6.089	3.749	0.2850	4.151	61.90	66.71	4.172
ROLT only	24.73	0.6972	0.4587	66.10	5.632	3.880	0.2920	4.300	64.20	68.90	4.270
SCVM only	25.68	**0.7151**	**0.4382**	65.40	5.840	3.885	0.2820	4.290	63.70	68.20	4.250
RefPLC	**25.81**	0.7079	0.4463	**67.89**	**5.325**	**3.964**	**0.2969**	**4.401**	**65.67**	**70.53**	**4.345**

G.6 Reference conditions on 90 GT targets

Each target is paired separately with Top, Medium, and Distant references. The LQ image, target, LQ caption, and random seed remain fixed. The reference caption changes with its source image. In Null, the reference image is omitted rather than replaced by a black or constant image. These observations characterize the effect of a reference supplied through the inference interface. The zero-evidence identities in Appendices B and C separately describe the internal residual parameterization.

Table G.6: Original-precision reference conditions on 90 GT targets. Bold marks the better recorded value within each metric and condition.

Metric	Model	Top	Medium	Distant	Null
PSNR	RefPLC	**30.855277**	**30.731862**	30.169705	30.093904
PSNR	Ada-RefSR	30.651530	30.544441	**30.435631**	**30.441269**
LPIPS	RefPLC	0.211761	0.211860	0.206943	0.215815
LPIPS	Ada-RefSR	**0.177204**	**0.181102**	**0.198942**	**0.215021**

RefPLC improves PSNR and LPIPS over its own Null result in all three supplied-reference conditions. Its Top–Null PSNR change is 0.761373 dB, compared with 0.210261 dB for Ada-RefSR. This within-model change is distinct from absolute LPIPS, for which Ada-RefSR has lower values in all four conditions. Top gives RefPLC's largest PSNR change and Distant its lowest LPIPS; these metrics do not induce the same reference ordering. Null is an input intervention on the same trained model, whereas PiD in Tables 1 and 4 is an independently adapted baseline.

Table G.7: Within-model changes from Null. A positive PSNR change or LPIPS reduction denotes improvement.

Model	Reference	PSNR change (dB)	LPIPS reduction
RefPLC	Top	0.761373	0.004054
RefPLC	Medium	0.637958	0.003955
RefPLC	Distant	0.075801	0.008872
Ada-RefSR	Top	0.210261	0.037817
Ada-RefSR	Medium	0.103172	0.033919
Ada-RefSR	Distant	-0.005638	0.016079

H Complete External-Benchmark Comparison

Tables H.1 and H.2 give all 56 method–dataset configurations across CUFED5, WR-SR, Bird, and Face at the original reporting precision. Dataset-specific face methods include CodeFormer, DMDNet, FaceMe, and InstantRestore (Zhou et al., 2022; Li et al., 2022; Liu et al., 2025; Zhang et al., 2024b). C^2-Matching and DATSR remain in CUFED5 and WR-SR. The complete and common-method comparisons use the same source scores; rankings are computed within the method set of each table. Nano Banana Pro is included on every dataset, matching its presence in the qualitative panels. RefPLC increases PSNR over Nano Banana Pro by 1.422, 2.238, 2.292, and 2.748 dB on CUFED5, WR-SR, Bird, and Face, respectively, and reduces FID by 43.37%, 67.94%, 44.22%, and 40.37%. On Bird and Face, Nano Banana Pro attains the lowest NIQE (4.132 and 3.351); RefPLC attains the highest PSNR. These results distinguish pixel reconstruction, paired perceptual similarity, and no-reference naturalness.

Table H.1: Complete external results, part I: CUFED5 and WR-SR. P: scores from Wang et al. (2026); other entries use official LQ inputs. Non-FT: official weights; FT: RefPLC fine-tuning data and updates (10 epochs). Red bold/blue underlined: best/second-best recorded values per dataset, excluding cross-source LPIPS ranking.

Dataset	Method	PSNR-Y↑	SSIM-Y↑	LPIPS↓	FID↓	NIQE↓	MUSIQ↑	CLIPIQA↑
CUFED5	C^2-Matching^P	21.0199	0.5303	0.8040	292.5921	9.5425	16.8868	0.1227
	DATSR^P	21.0196	0.5306	0.8071	292.7726	9.6305	16.8736	0.1269
	Real-ESRGAN-S1^P	20.3084	**0.5543**	0.3697	175.9116	3.8861	64.7123	0.4214
	S3Diff-S1^P	20.4594	0.5234	0.3544	160.1337	**3.7096**	63.8519	0.5238
	OSEDiff-S1^P	18.8748	0.4935	0.3669	150.7504	4.0005	65.2446	0.5077
	PiSA-SR-S1^P	20.3371	0.5286	0.3707	164.6927	4.1992	64.2840	0.5043
	SUPIR+ReFIR-S50^P	18.9994	0.4711	0.4359	151.2700	4.2745	62.7384	0.4235
	SeeSR+ReFIR-S50^P	20.2169	0.5255	0.3452	137.3693	3.7426	**70.5375**	0.5538
	Nano Banana Pro	19.577	0.425	0.507	251.800	4.101	62.896	0.402
	SSMTF	21.0253	0.5315	0.8245	294.3038	9.9061	16.8306	0.1421
	SSMTF FT	**21.0638**	0.5354	0.8323	293.3923	10.1954	16.8447	0.1569
	Ada-RefSR-S1^P	20.4843	0.5461	0.2894	127.8945	3.8701	67.9070	0.5174
	Ada-RefSR FT	20.5424	0.5474	0.2892	**126.3837**	3.8513	68.1366	0.5162
	RefPLC	20.999031	0.544612	0.388216	142.589789	3.797353	65.705989	**0.558830**
WR-SR	C^2-Matching^P	22.8766	0.5784	0.7471	153.1606	9.0121	19.9772	0.1527
	DATSR^P	22.9017	0.5820	0.7583	153.5259	9.3999	20.1124	0.1613
	Real-ESRGAN-S1^P	22.1491	**0.5974**	0.3631	97.9142	3.6976	64.5176	0.6076
	S3Diff-S1^P	21.9122	0.5620	0.3542	63.8237	4.4171	63.8237	0.6149
	OSEDiff-S1^P	21.3263	0.5603	0.3261	71.5148	3.7364	71.1935	0.6899
	PiSA-SR-S1^P	21.8813	0.5682	0.3383	62.9958	4.0359	69.5187	0.6407
	SUPIR+ReFIR-S50^P	21.0134	0.5378	0.3949	76.4993	3.6407	70.1242	0.6219
	SeeSR+ReFIR-S50^P	21.8334	0.5673	0.3435	61.9597	3.8919	72.1520	0.7120
	Nano Banana Pro	20.214	0.460	0.530	143.113	3.841	67.933	0.619
	SSMTF	22.9542	0.5860	0.7761	153.3310	10.3300	19.7599	0.1675
	SSMTF FT	**23.0508**	0.5937	0.7823	152.7906	10.3188	19.4570	0.1896
	Ada-RefSR-S1^P	21.9722	0.5777	0.3061	53.2811	3.7429	69.8608	0.6612
	Ada-RefSR FT	22.0784	0.5805	0.3058	54.2231	3.7687	69.8980	0.6563
	RefPLC	22.451501	0.592870	0.306060	**45.882290**	**3.610163**	**72.235892**	**0.716945**

Table H.2: Complete external results, part II: Bird and Face. P: published scores; other entries use official LQ inputs. Non-FT: official weights; FT: RefPLC fine-tuning data and updates (10 epochs). Red bold/blue underlined: best/second-best recorded values per dataset, excluding cross-source LPIPS ranking.

Dataset	Method	PSNR-Y↑	SSIM-Y↑	LPIPS↓	FID↓	NIQE↓	MUSIQ↑	CLIPIQA↑
Bird	Real-ESRGAN-S1^{P}	25.0809	0.7044	0.5306	178.8849	8.8496	29.3109	0.3571
	S3Diff-S1^{P}	24.8422	0.7106	0.2903	60.9452	5.3195	73.1125	0.8285
	OSEDiff-S1^{P}	23.7861	0.7086	0.2949	49.1911	4.9796	74.7567	0.8451
	PiSA-SR-S1^{P}	24.4891	0.7003	0.2816	67.9959	4.8513	74.3633	0.8373
	SUPIR+ReFIR-S50^{P}	23.8344	0.6974	0.3811	94.9298	5.4422	63.6740	0.6478
	SeeSR+ReFIR-S50^{P}	23.7211	0.6985	0.2945	52.4038	5.6182	**75.1056**	**0.8617**
	Nano Banana Pro	23.514	0.618	0.437	72.196	**4.132**	73.257	0.789
	SSMTF	25.2893	0.7068	0.6711	178.6543	13.1533	17.0455	0.1426
	SSMTF FT	25.3910	0.7192	0.6794	175.4906	12.4206	16.9004	0.1577
	Ada-RefSR-S1^{P}	25.2998	0.7292	0.2536	**36.4249**	5.1771	72.0039	0.8303
	Ada-RefSR FT	25.4341	0.7331	0.2545	36.5540	5.2584	71.6932	0.8246
	RefPLC	**25.805537**	**0.746003**	0.288653	40.270590	4.446717	73.460886	0.830723
Face	CodeFormer^{P}	26.2677	0.7082	0.3798	99.6599	5.4105	63.9886	0.5192
	DMDNet^{P}	26.8993	0.7472	0.2316	56.6262	4.6030	72.9495	0.6395
	Real-ESRGAN-S1^{P}	26.1678	0.7142	0.5113	162.4299	8.9504	26.4788	0.3630
	S3Diff-S1^{P}	26.6881	0.7344	0.2189	51.1219	5.0599	74.1095	0.6963
	OSEDiff-S1^{P}	25.5302	0.7434	0.2263	55.0428	4.9550	71.7721	0.6459
	PiSA-SR-S1^{P}	26.2433	0.7254	0.2126	48.9154	4.5423	**75.9855**	0.7191
	SUPIR+ReFIR-S50^{P}	25.7231	0.7062	0.2623	57.1432	3.9686	73.0854	0.6260
	SeeSR+ReFIR-S50^{P}	26.3191	0.7375	0.2293	57.2910	5.2885	75.4230	**0.7236**
	Nano Banana Pro	25.426	0.653	0.332	79.403	**3.351**	75.035	0.703
	FaceMe-S50^{P}	26.6591	0.7245	0.2500	54.3278	4.7295	72.4144	0.6483
	InstantRestore-S1^{P}	26.2245	0.7350	0.2070	51.2283	5.9473	69.7209	0.5726
	SSMTF	26.4983	0.7202	0.6527	177.1558	12.0090	17.9103	0.1783
	SSMTF FT	26.6575	0.7352	0.6582	178.8888	11.5711	17.6381	0.2059
	Ada-RefSR-S1^{P}	27.1271	0.7523	0.1749	**42.7042**	4.4880	74.3828	0.6453
	Ada-RefSR FT	27.2747	**0.7562**	0.1756	42.9781	4.5556	74.1203	0.6409
	RefPLC	**28.173506**	0.752425	0.253776	47.346693	4.485465	69.811805	0.646570

I Reference Interfaces and Interpretation

Reference-based restoration separates the availability of external detail from the mechanism used to apply it. SRNTT and TTSR transfer reference texture, while MASA-SR and C^2-Matching refine the correspondence machinery (Zhang et al., 2019; Yang et al., 2020; Lu et al., 2021; Jiang et al., 2021). Generative approaches further regulate reference information inside a restoration model. Table I.1 summarizes the representations and intervention sites most closely related to RefPLC.

Table I.1: Architectural comparison of reference representations, intervention sites, and regulation.

Method	Reference representation	Intervention	Regulation
MASA-SR (Lu et al., 2021)	Matched multi-scale features	Restoration feature aggregation	Coarse-to-fine search and spatial adaptation
Ref-Diff (Dong et al., 2024)	Reference image and change priors	Conditional diffusion denoising	Change-aware conditioning
iRAG (Lee et al., 2025)	Real/generated reference patches	Retrieval-augmented restoration	Hash retrieval and hallucination filtering
ReFIR (Guo et al., 2024)	Cross-image features	Restoration-model attention	Spatial gating and distribution alignment
Ada-RefSR (Wang et al., 2026)	Reference attention and summary tokens	S3Diff attention features	Implicit correlation gating
PiD (Lu et al., 2026)	Target latent condition	Latent-conditioned RGB diffusion	Condition adapter
RefPLC	Transported VAE differences and aligned detail evidence	Before latent projection and at RGB velocity output	Zero-anchored maps, latent RMS cap, and state-conditioned gain

Connected rather than independent corrections. ROLT changes the latent condition supplied to PiD. The resulting base velocity enters SCVM both directly and through the one-step clean estimate. Thus SCVM interprets the reference against a state already influenced by ROLT; the two branches do not independently generate images that are blended afterwards. This computational dependence defines the two-interface connection. The Single717 ablation evaluates the complete branches: RefPLC records the best PSNR and no-reference values, while SCVM alone records the highest SSIM and lowest LPIPS-VGG. The results describe these configurations, not the isolated effect of each internal operation.

Correspondence, evidence, and usage strength. The affinity weights determine the source of a retrieved value. The sharp-minus-blurred difference defines its content. Target-conditioned mapping and spatial gating determine its contribution. This separation allows matching descriptors to serve localization without making the descriptor itself the transferred restoration signal. In SCVM, the caption relation adjusts already aligned visual evidence; it is separate from the global–local mixing weight and from the time-dependent spatial gain.

Residual amplitudes and reference availability. The latent RMS cap bounds the RMS amplitude of the condition increment relative to the LQ latent. The scaled $\tanh$ and spatial gain bound the additional velocity at a given sample state. These are local properties of the parameterization, while image quality is measured on the reconstructed output. The shared-map cancellation applies to zero evidence at the module input. Null specifies a different, observable condition: no reference image is supplied to the complete inference pipeline. The 90-target comparison uses that condition as each model's own baseline and reports the changes induced by three available references.

J Additional Qualitative Results

J.1 Single717 examples and display configurations

The grids retain a common two-row, six-column layout. The upper row contains LQ, OmniGen2, Qwen-Image-Edit-2511, FLUX.1 Kontext-dev, Nano Banana Pro, and GT; the lower row contains Ref, BAGEL, SSMTF, PiD, Ada-RefSR, and RefPLC. Full views show target composition and object geometry. Where included, enlarged target regions use corresponding normalized crops across outputs, while reference crops are selected in reference coordinates.

All Single717 qualitative and quantitative comparisons use the same LQ input across methods for each target, generated by the official Real-ESRGAN degradation procedure. The PiD results use 10-epoch fine-tuning on the same training dataset, with the same learning rate as the corresponding PiD parameters in RefPLC. This adaptation updates the PiD conditioning pathway and leaves the base Pixel Diffusion network frozen. Displayed SSMTF and Ada-RefSR results use checkpoints fine-tuned on RefPLC's training dataset for 10 epochs with the same number of optimization updates as RefPLC, consistent with the Single717 quantitative comparison. The selected scenes, method order, and target regions are consistent across the comparison grids. Panels preserve the supplied outputs and aspect ratios.

Portrait structure and local appearance. Complementing the main-text group portrait and adult–child scene, Figure J.2 shows the knitted hat and facial contour. Figure J.6 highlights the glasses rim, moustache, and cheek texture. Figure J.7 shows the eyes, hair, and mask boundary. These descriptions concern the displayed output regions.

Lettering and surface geometry. Figure J.3 juxtaposes reflective edges with wall texture. Figure J.4 shows character boundaries against a textured surface. Figure J.5 shows narrow edges and interior gaps. The illuminated manuscript in Figure J.1 complements these views through its rolled-paper edge, paper-surface appearance, and shadow transitions.

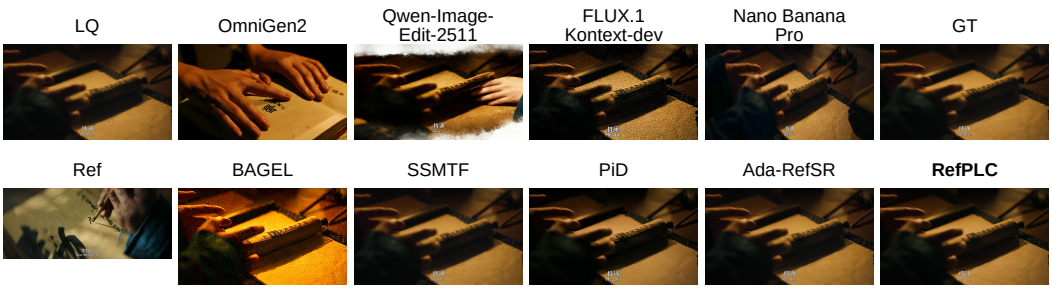

Figure J.1: **Illuminated manuscript.** Complete model grid showing the rolled-paper edge, paper-surface appearance, and shadow structure. Full views show the paper boundary and its surrounding illumination.

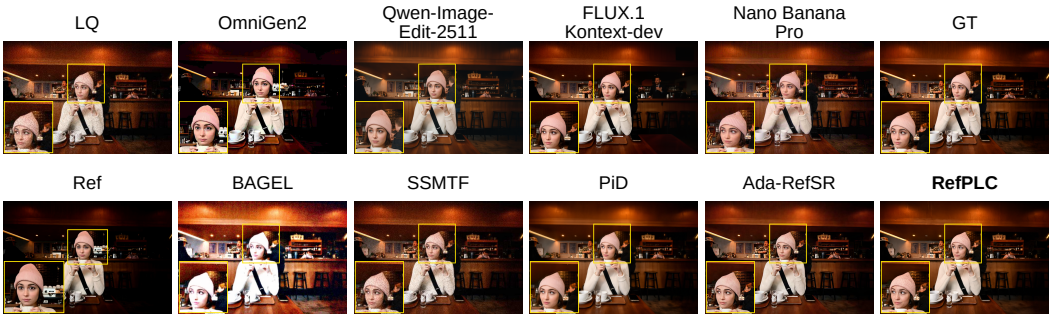

Figure J.2: **Cafe portrait.** Complete comparison of the knitted hat, face, and surrounding scene. Insets expose local texture and facial contours.

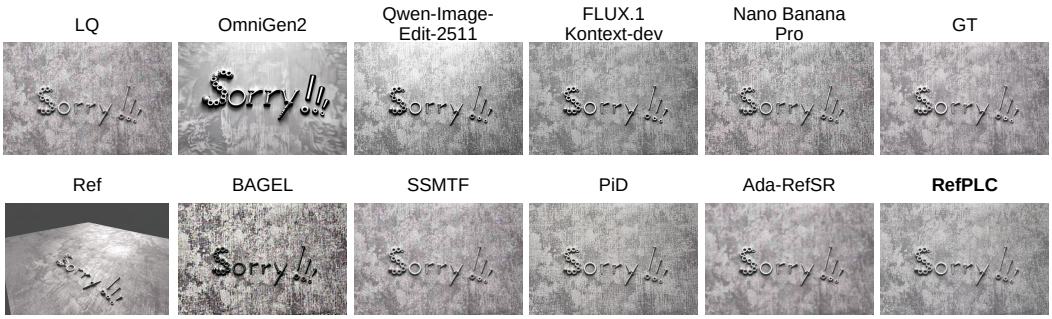

Figure J.3: **Metal lettering.** Complete comparison of reflective contours and the surrounding wall. Full views show the placement of reflective letters against the wall.

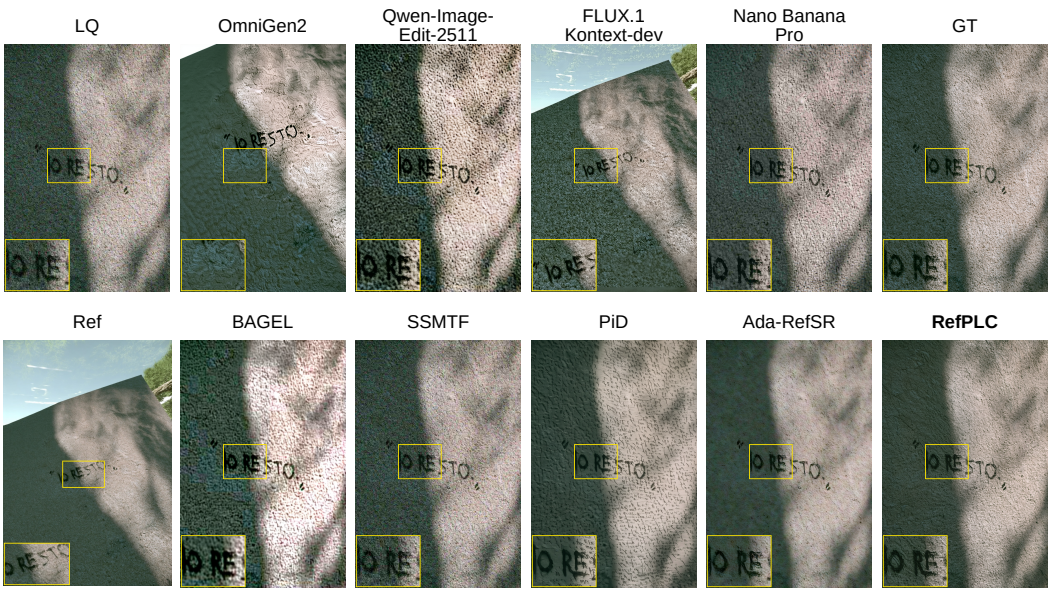

Figure J.4: **Wall lettering.** Complete model grid showing character boundaries and the textured surface. Enlargements use corresponding target regions.

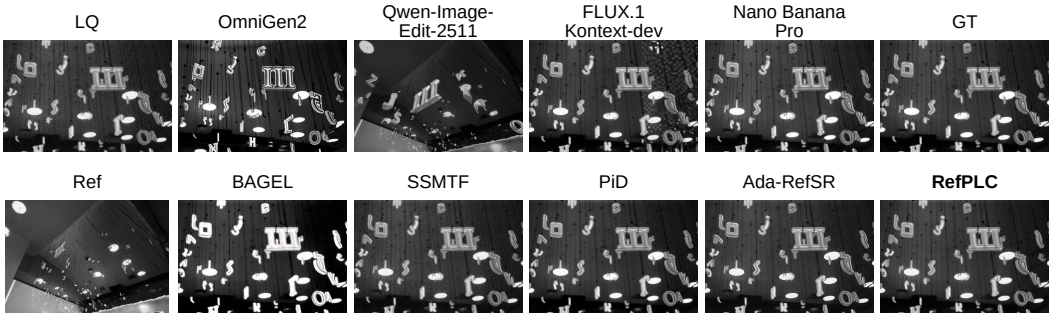

Figure J.5: **Suspended Roman numerals.** Full-image comparison of narrow strokes, edge profiles, and interior gaps.

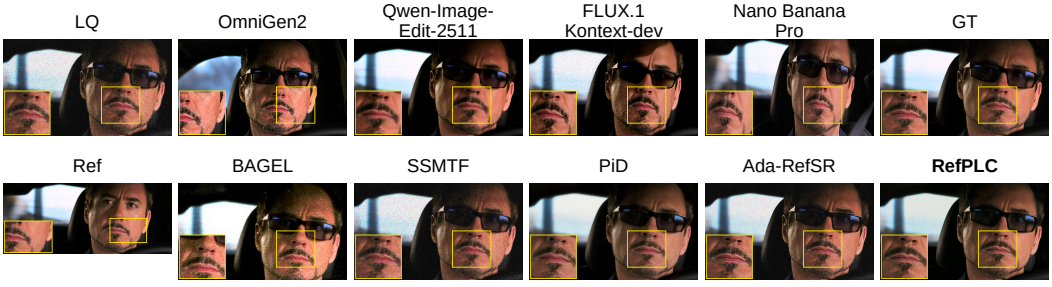

Figure J.6: **Glasses and beard.** Complete model grid with enlarged glasses, cheek, and facial-hair detail.



Figure J.7: **Masked portrait.** Full-image comparison of the target pose, clothing, hair, and mask boundary.

J.2 External examples

Figure J.8 supplies the complementary egret and portrait examples. Eye-region insets expose local shape and texture while the full views show consistency with target pose and composition. The example-level descriptions concern the displayed outputs.

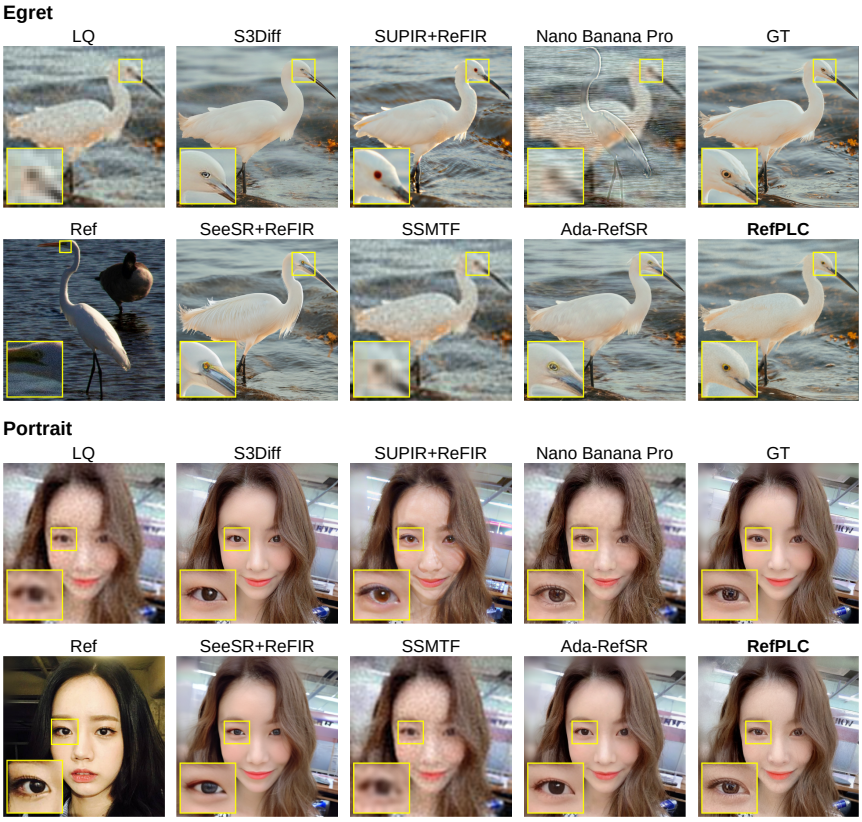

Figure J.8: **Additional external examples.** Egret and portrait restoration with changes in appearance and viewpoint between reference and target. Yellow boxes mark the enlarged regions.

J.3 REFERENCE-CONDITION VISUALIZATION

Figure J.9 fixes the target and compares the three reference choices and no-reference output. Its two enlarged regions emphasize hair, glasses, and facial outlines. The different reference images provide varying views of the scene; the outputs are inspected at the same target locations. The corresponding aggregate experiment uses 90 GT targets and is reported in Table 3.



Figure J.9: **Reference conditions.** GT and Top/Medium/Distant inputs, followed by RefPLC and Ada-RefSR outputs under all three references and Null. Enlargements use the same target regions. Null means no reference image is supplied.