GATE3DGEN: A PRE-PROCESSING MODEL THAT SELECTIVELY PURIFIES REAL-WORLD IMAGES FOR 3D GENERATION

Anonymous authors
Paper under double-blind review

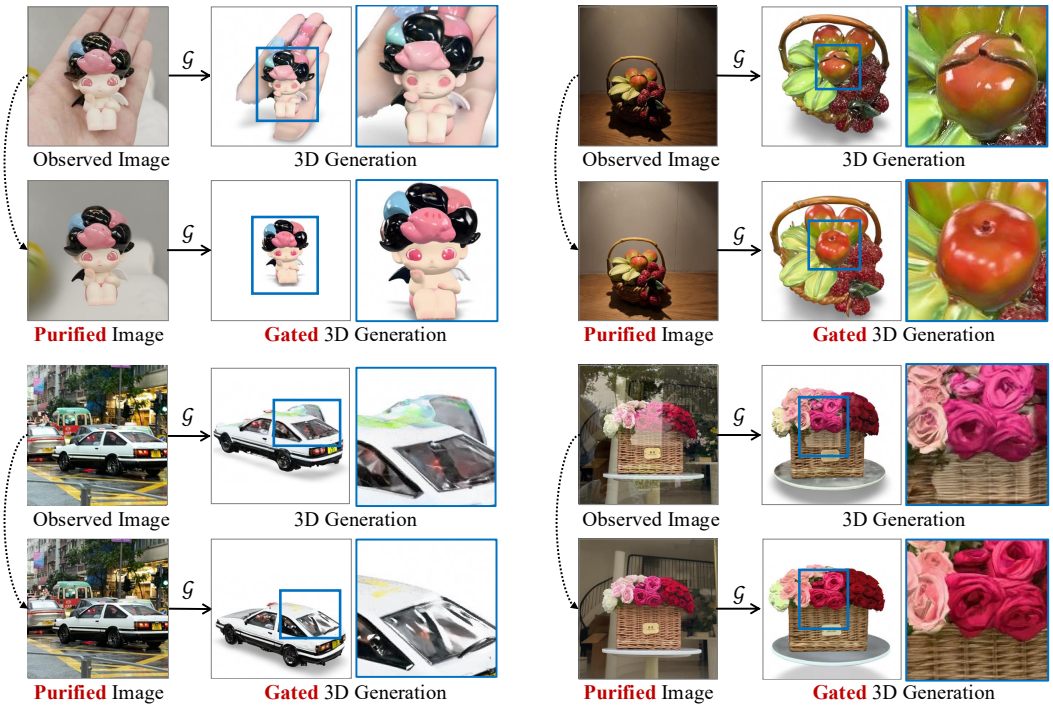


Figure 1: **Gate3DGen improves 3D generation by purifying real-world image inputs.** Each example compares the same frozen 3D generator $\mathcal{G}$ conditioned on the observed image (top) and our purified image (bottom). The examples show occlusion removal, illumination correction, and reflection suppression, with reduced extraneous geometry, contour distortions, and surface artifacts.

ABSTRACT

Single-image 3D generation has achieved high fidelity on curated benchmarks, yet degrades sharply on real-world images with uneven illumination, blur, and occlusions. Purifying such inputs requires deciding when to intervene and preserving the visual cues needed for 3D generation. We introduce **Gate3DGen**, a selective pre-processing model that gates image inputs to existing 3D generators without retraining them. A degradation-aware router bypasses inputs classified as clean and extracts features to guide purification of the remaining inputs. Conditioned on these features, a 3D-oriented one-step purifier purifies degraded inputs with spatial losses that encourage preservation of object identity, contours, and textural cues. To provide paired supervision, we construct a synthetic data engine that renders clean views of 3D assets and synthesizes composite degradations. On the synthetic benchmark, the purifier improves image fidelity and downstream 3D geometry, reducing TRELLIS.2 mesh distance by 61.5% relative to degraded inputs. VLM evaluation and human preferences on real photographs further support identity preservation and degradation correction. Code and model will be released.

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.

1 INTRODUCTION

Single-image 3D generative models have achieved remarkable progress on curated benchmarks (Hong et al., 2024; Tochilkin et al., 2024; Xu et al., 2024; Xiang et al., 2026; Li et al., 2026), yet degrade sharply on real-world images. This vulnerability stems from restrictive assumptions about input images. Contemporary models assume centered, sharp, and fully visible objects under uniform studio lighting (Xu et al., 2024; Xiang et al., 2025; 2026). In real-world scenarios, however, unconstrained photographs routinely exhibit blur, uneven illumination, or occlusions (Cho et al., 2025; Qu et al., 2026), corrupting essential visual cues and inducing severe artifacts in 3D outputs.

Open-source image-to-3D pipelines commonly use background removal and recentering (Tochilkin et al., 2024; Xu et al., 2024; Xiang et al., 2025; Zhao et al., 2025), which do not directly address photometric corruptions or foreground occlusions. Specialized preprocessing also targets individual factors such as illumination through delighting (Li et al., 2025a; Zhao et al., 2025). When multiple degradations coexist, correcting one factor can leave other sources of ambiguity in the conditioning image. Our comparisons show that generic 2D image editing can alter object appearance or contours, with adverse effects on downstream geometry (Section 4.4). The objective is therefore to recover a suitable observation of the same object, retaining the visible structural evidence that the 3D generator uses to infer shape and appearance. This motivates selective input purification that preserves visual cues for 3D generation while avoiding unnecessary changes to already suitable clean inputs.

We therefore introduce Gate3DGen, a selective pre-processing model that gates real-world images to existing 3D generators (Figure 1). It routes each image through purification or an unchanged bypass, keeping the downstream generators frozen. Gate3DGen couples a *degradation-aware router* with a *3D-oriented one-step purifier*. The router bypasses inputs classified as clean and selects the remaining inputs for purification, while extracting continuous degradation-aware features for the purifier. Conditioned on these features, the one-step purifier purifies degraded inputs in a single forward pass with spatial losses that encourage preservation of object identity and essential geometric and textural cues. A central challenge in training this model is the absence of matched ground-truth references for real-world images. To provide paired training supervision without manual collection, we construct a synthetic data engine from 3D assets. Rendering multi-view anchors from 3,961 3D assets (Zhang et al., 2025) and synthesizing composite corruptions (e.g., uneven illumination, blur, and occlusions) yields 43,936 aligned image pairs and 1,082 held-out cases with 3D ground truth.

We evaluate the purifier on this synthetic benchmark with two pretrained 3D generative models, TRELLIS.2 (Xiang et al., 2026) and Pixal3D (Li et al., 2026). For 2D purification, Gate3DGen purifies degraded inputs, reducing foreground LPIPS by 39.1% while preserving subject silhouettes, whereas the compared generic image editing methods can alter object identity and reduce reconstruction fidelity. When used to condition these 3D generative models, the purified images consistently improve all geometry and appearance metrics over the original observations. Specifically, Gate3DGen achieves a 61.5% reduction in all-surface mesh distance, closes 55.5% of the geometry gap to clean references on average (up to 83.1% on TRELLIS.2), and reduces rendering LPIPS by 10.9%. Evaluated separately, the degradation-aware router identifies degraded inputs with 88.45% recall and 93.47% specificity on clean inputs, sending inputs classified as degraded for purification. On real photographs, VLM scores and human preferences further support identity preservation and degradation correction, complementing the reference-based evaluation on synthetic inputs.

Our contributions are as follows.

- We formulate robust single-image 3D generation as selective input purification, improving existing 3D generators through purified inputs without modifying their architectures or retraining.
- We introduce Gate3DGen, combining a router that bypasses inputs classified as clean with a one-step purifier. Degradation-aware features and spatial losses guide purification toward preserving object identity, contours, and geometric and textural cues for downstream 3D generation.
- We construct a synthetic data engine that provides paired supervision for purification from 3D assets and a benchmark with 3D ground truth for evaluating downstream reconstruction. Our method reduces foreground LPIPS by 39.1% and TRELLIS.2 mesh distance by 61.5% relative to degraded inputs, demonstrating gains in both input-image fidelity and downstream 3D geometry.

2 Related Work

Robustness in Single-Image 3D Generation. Modern single-image 3D methods, including feed-forward reconstruction and iterative diffusion or flow models, achieve high fidelity on curated inputs (Hong et al., 2024; Tochilkin et al., 2024; Xu et al., 2024; Boss et al., 2025; Xiang et al., 2025; Zhao et al., 2025; Li et al., 2025b; Lai et al., 2025; Xiang et al., 2026; Li et al., 2026). Blur, uneven illumination, and occlusions can reduce their generation quality. Common input preprocessing operations include foreground segmentation and recentering, which do not directly correct illumination, blur, or foreground occlusions. Specialized methods address illumination through lighting decomposition or delighting (Li et al., 2025a; Lai et al., 2025). Related work investigates recovery of occluded geometry through volumetric reconstruction or multi-view synthesis (Cho et al., 2025; Qu et al., 2026; Shi et al., 2025). Our work focuses on selective input purification before frozen 3D generators, addressing composite degradations involving uneven illumination, blur, and occlusions.

2D Image Editing and Degradation Routing. Recent advances in instruction-guided editing (Labs et al., 2025; Wu et al., 2025; 2026; Liu et al., 2025; Zhang et al., 2026; Qiao et al., 2026) and all-in-one restoration (Potlapalli et al., 2023; Conde et al., 2024; Lin et al., 2023; Yu et al., 2024; Guo et al., 2024) synthesize visually compelling images. Their objectives do not explicitly measure downstream 3D fidelity. In our comparisons, generic image editing methods can remove visible degradations while changing object appearance or contours, with adverse effects on downstream geometry (Section 4.4). Editing clean photographs can introduce unwanted changes, motivating selective restoration. No-reference quality metrics (Mittal et al., 2013; Yang et al., 2022; Zhang et al., 2023) and general vision-language models (You et al., 2024; Bai et al., 2025; OpenAI, 2026) provide candidate quality signals, but their suitability for routing requires evaluation (Section 4.2). Gate3DGen couples a degradation-aware router with a one-step purifier to preserve identity and contours while selectively purifying inputs.

3 Method

Gate3DGen is a selective pre-processing model that purifies real-world input images without modifying the downstream 3D generator. It targets images degraded by uneven illumination, blur, and occlusions while encouraging preservation of geometric and textural cues. Figure 2 summarizes the model. The following sections detail the task formulation, data engine, router, and one-step purifier.

3.1 Problem Formulation

Let $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$ be a pretrained generator mapping a photograph to a 3D representation $y \in \mathcal{Y}$. A matched clean reference x^c shows a centered, sharp, fully visible subject under uniform lighting, with $\mathcal{G}(x^c) \approx y$. An observed image x^o may be clean or degraded. Uneven illumination, blur, and occlusions can violate these assumptions. These degradations can distort geometry ($\mathcal{G}(x^o) \not\approx y$).

We learn to purify input images under supervision from matched clean references. The purifier maps x^o toward the matched clean reference x^c using reconstruction and spatial losses that encourage preservation of object identity, contours, and textural cues. The objective is given by

$$\min_{\mathcal{C}_\theta} \mathbb{E}_{(x^o, x^c)} \left[\ell\big(\mathcal{C}_\theta(x^o),\, x^c\big) \right]. \tag{1}$$

Here, ℓ combines the latent, photometric, perceptual, region-weighted reconstruction, and edge losses defined in Section 3.4. To avoid unnecessary purification of clean inputs ($x^o \approx x^c$), the degradation-aware router $\mathcal{R}_\phi$ predicts a routing decision $\widehat{\gamma} \in \{0, 1\}$. The routing decision determines the pre-processed image as follows

$$\widehat{x}^c = \begin{cases} x^o, & \widehat{\gamma} = 0 \quad (\textsc{Keep}), \\ \mathcal{C}_\theta(x^o), & \widehat{\gamma} = 1 \quad (\textsc{Purify}), \end{cases}. \tag{2}$$

The *pre-processed image* $\widehat{x}^c$ is either a *purified image* from $\mathcal{C}_\theta$ (Purify) or the unchanged input (Keep). To supervise this objective where matched pairs (x^o, x^c, γ) are unavailable for real-world images, we construct a synthetic data engine from 3D assets (Section 3.2) to supply training data.

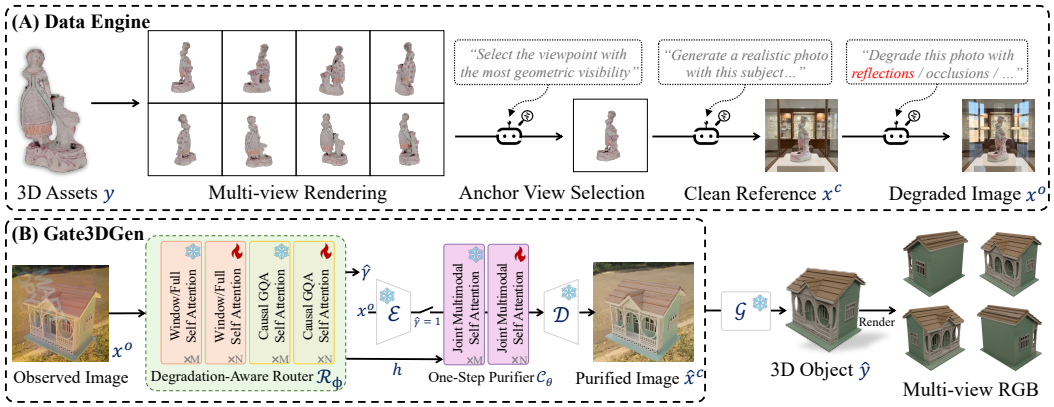


Figure 2: **Overview of Gate3DGen. (A) Data Engine.** We render 3D assets from multi-view, select informative anchor views, and apply composite degradations to construct aligned observation-reference pairs (x^o, x^c). **(B) Gate3DGen.** Router $\mathcal{R}_\phi$ bypasses inputs classified as clean ($\widehat{\gamma} = 0$). Otherwise ($\widehat{\gamma} = 1$), its features h guide the one-step DiT purifier $\mathcal{C}_\theta$. The image $\widehat{x}^c$ conditions the frozen 3D generator $\mathcal{G}$, with clean references and spatial losses used only during training.

3.2 Synthetic Data Engine

Learning $\mathcal{C}_\theta$ requires paired images (x_i^o, x_i^c) that share identical 3D geometry and viewpoints while differing only in capture quality, whereas learning $\mathcal{R}_\phi$ requires examples labeled by whether purification is necessary. Because real photographs provide neither matched clean counterparts nor controlled routing labels, we use high-fidelity 3D assets to construct this supervision (Figure 2A).

Anchor view selection. From about $4k$ high-fidelity 3D assets in TexVerse (Zhang et al., 2025), we render eight views under uniform lighting. A vision-language model (VLM) evaluates and ranks these viewpoints according to geometric visibility and self-occlusion. The four most informative views per asset are retained as multi-view anchors and serve as matched clean references x_i^c.

Degradation synthesis. For each anchor, instruction-guided editing produces a matched observation x_i^o with composite corruptions, including uneven illumination, blur, and occlusions. Each pair is audited by Qwen3-VL (Qwen Team, 2025) for structural consistency, and RMBG-2.0 extracts a soft foreground mask m_i from the clean reference x_i^c. This pipeline produces roughly $44k$ image pairs $\mathcal{D}_{\text{train}} = \{(x_i^o, x_i^c, m_i)\}$ for $\mathcal{C}_\theta$ and $100k$ routing pairs $\mathcal{D}_{\text{route}} = \{(x_j, \gamma_j)\}$ for $\mathcal{R}_\phi$, where $\gamma_j = 1$ denotes PURIFY and $\gamma_j = 0$ denotes KEEP. Training and evaluation assets are disjoint.

3.3 Degradation-Aware Router

Purifying clean inputs can introduce unwanted changes. To guide purification of images with spatially varying degradations, we use features extracted from the input image. We therefore design the degradation-aware router $\mathcal{R}_\phi$ based on Qwen2.5-VL (Bai et al., 2025). Trained on $\mathcal{D}_{\text{route}}$, the router predicts whether purification is needed and extracts continuous degradation-aware features.

Logit-based routing. Given input x^o, the router returns $(p, h) = \mathcal{R}_\phi(x^o)$, where $p \in [0, 1]$ is the predicted probability that purification is needed and h denotes continuous degradation-aware features. Rather than using greedy decoding, we derive the purification probability p directly from the unnormalized logits for the routing decisions PURIFY and KEEP, denoted by o_{purify} and o_{keep}.

$$p = \sigma(o_{\text{purify}} - o_{\text{keep}}), \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function. We apply a threshold τ to obtain the routing indicator

$$\widehat{\gamma} = \mathbb{I}[p \geq \tau] \in \{0, 1\}, \tag{4}$$

where $\mathbb{I}[\cdot]$ is the indicator function. Predicted clean inputs bypass purification unchanged.

Degradation-aware conditioning. When purification is triggered ($\widehat{\gamma} = 1$), we extract the final hidden states $h \in \mathbb{R}^{N \times d}$ from $\mathcal{R}_\phi$ under a fixed editing instruction. These features condition $\mathcal{C}_\theta$.

3.4 3D-ORIENTED ONE-STEP PURIFIER

The diffusion purifier $\mathcal{C}_\theta$ purifies degraded inputs in a single forward pass with spatial losses that encourage preservation of object identity and geometric and textural cues. Let $\mathcal{E}$ and $\mathcal{D}$ denote the encoder and decoder of its frozen image autoencoder. We initialize $\mathcal{C}_\theta$ from the released FireRed-Image-Edit-1.1 (Qiao et al., 2026) checkpoint, whose architecture and flow-matching formulation follow the FireRed-Image-Edit-1.0 technical report (Qiao et al., 2026), and condition it on the observation latent $\mathcal{E}(x^o)$ and the continuous degradation-aware features h produced by the router.

One-step flow-matching. Standard flow-matching models (Qiao et al., 2026) define a linear probability path between the target latent $z_0 = \mathcal{E}(x^c)$ and Gaussian noise $\varepsilon \sim \mathcal{N}(0, I)$, given by

$$z_t = (1-t)\, z_0 + t\, \varepsilon, \quad t \in [0,1], \tag{5}$$

The velocity network $v_\theta(z_t, \mathcal{E}(x^o), h, t)$ is trained to regress $\varepsilon - z_0$. Standard inference integrates from $t = 1$ to $t = 0$ over multiple network evaluations. We train and evaluate $\mathcal{C}_\theta$ at $t = 1$, where $z_1 = \varepsilon$, to reduce inference latency. One prediction yields the latent estimate and decoded image

$$\widehat{z}_0 = \varepsilon - v_\theta\big(\varepsilon,\, \mathcal{E}(x^o),\, h,\, 1\big), \qquad \widehat{x}^c = \mathcal{D}(\widehat{z}_0), \tag{6}$$

where $\mathcal{D}$ is the frozen VAE decoder. Training and inference both use Equation 6, with one evaluation of v_θ followed by VAE decoding. At this endpoint, the noisy latent contains only Gaussian noise. The observation latent and router features provide the image-specific conditioning. The velocity network is supervised using the clean latent and the decoded image, without iterative sampling.

3D-oriented spatial constraints. For a training triple $(x_i^o, x_i^c, m_i) \in \mathcal{D}_{\text{train}}$, let $z_{0,i} = \mathcal{E}(x_i^c)$, and let $\widehat{z}_{0,i}$ and $\widehat{x}_i^c$ be obtained from Equation 6. Standard pixel reconstruction losses treat all regions uniformly. Generic image editing methods overwrite uncorrupted regions and blur high-frequency shading cues that 3D generators rely on. The region-weighted reconstruction and edge losses guide purification by supervising the subject, emphasizing degradation-affected regions, and encouraging preservation of edges. We obtain a soft degradation map r_i by smoothing the pairwise residual between x_i^o and x_i^c. For an error tensor a and a nonnegative spatial weight w broadcast to the shape of a, we define the normalized weighted mean

$$\overline{\ell}(a, w) = \frac{\sum_j w_{(j)} |a_{(j)}|}{\max\Big(\sum_j w_{(j)},\, 1\Big)}, \tag{7}$$

where the parenthesized subscript (j) indexes channel and spatial entries within each tensor after broadcasting, while i indexes training samples. With $\delta_i = \widehat{x}_i^c - x_i^c$, the implemented losses are

$$\mathcal{L}_{\text{region}} = \overline{\ell}(\delta_i, m_i) + \overline{\ell}(\delta_i, r_i), \qquad \mathcal{L}_{\text{edge}} = \tfrac{1}{2}\left[\overline{\ell}(D_u \delta_i, m_i^u) + \overline{\ell}(D_v \delta_i, m_i^v)\right], \tag{8}$$

The two terms in $\mathcal{L}_{\text{region}}$ are normalized independently: the first supervises the full subject through m_i, while the second focuses on degradation-affected regions through r_i. The edge term matches finite differences along the horizontal (u) and vertical (v) image axes, weighted by adjacent subject-mask values. These spatial weights guide purification. Selectivity refers to the image-level routing decision. Section A specifies the degradation-map pooling and adjacent-mask edge weights.

Total training objective. Combining these spatial constraints with the backbone objectives yields the full purifier loss, given by

$$\mathcal{L} = \lambda_{\text{lat}}\mathcal{L}_{\text{lat}} + \lambda_1 \mathcal{L}_1 + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}} + \lambda_{\text{region}}\mathcal{L}_{\text{region}} + \lambda_{\text{edge}}\mathcal{L}_{\text{edge}}, \tag{9}$$

where $\mathcal{L}_{\text{lat}} = \text{MSE}(\widehat{z}_{0,i}, z_{0,i})$ supervises latent reconstruction at the endpoint, $\mathcal{L}_1 = \text{mean}\,|\widehat{x}_i^c - x_i^c|$ is the full-image photometric loss, and $\mathcal{L}_{\text{lpips}} = \text{LPIPS}(\widehat{x}_i^c, x_i^c)$ enforces perceptual agreement. The latent term supervises the clean representation, while photometric and perceptual terms supervise the decoded image alongside spatial and gradient constraints. Loss weights appear in Section A.

Clean references and foreground masks are used only during training. Inference uses the observed image and router features. Spatial losses encourage preservation of geometric and textural cues without gradients through the frozen 3D generator, whose output quality is evaluated in Section 4.4.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmark and evaluation protocols. Our benchmark contains 1,082 held-out DTC cases (Dong et al., 2025) with 3D ground truth and clean references sharing the degraded observations' camera poses. We compare 3D outputs from degraded, edited, and clean-reference images against ground truth, treating clean references as oracle controls. Our purifier and all editing baselines process every degraded input. Routing decisions are evaluated separately and do not affect these results.

Baselines. We compare against 7 instruction-guided image editing methods, namely FLUX.1-Kontext (Labs et al., 2025), FireRed-Image-Edit (1.1) (Qiao et al., 2026), Mage-Flow-Edit (Zhang et al., 2026), OmniGen2 (Wu et al., 2026), Qwen-Image-Edit (Wu et al., 2025), Step1X-Edit (Liu et al., 2025), and GPT Image-2.5 (Flare) (OpenAI, 2026). We evaluate downstream 3D quality with Pixal3D (Li et al., 2026) and TRELLIS.2 (Xiang et al., 2026).

Evaluation metrics. For *degradation-aware routing*, we report PURIFY-class recall, overall accuracy, and H, defined as their harmonic mean, all expressed as percentages (Table 1). For *2D image fidelity*, we report foreground LPIPS, PSNR, SSIM, Sobel gradient MAE, and foreground mask IoU against the clean reference. For *3D metrics*, we report all-surface Mesh Distance (MD) (Xiang et al., 2026) and F_1^{all} ($F_1@10^{-8}$), outer-surface Chamfer Distance (CD) (Fan et al., 2017) and F_1^{outer} ($F_1@10^{-6}$) after ICP alignment to the ground-truth asset, PSNR-N (Wang et al., 2004) on rendered normal maps, and LPIPS (Zhang et al., 2018) on four-view shaded and albedo renderings.

4.2 DEGRADATION-AWARE ROUTING RESULTS

Table 1 reports routing results on $1,082$ degraded and 934 clean synthetic inputs and on 96 in-the-wild images. Because real photographs lack explicit routing labels, a GPT-5 agent assigns proxy PURIFY/KEEP labels by comparing multi-view renderings and no-reference metrics for 3D assets generated from the unedited and purified images by Pixal3D and TRELLIS.2. The router should direct degraded inputs to the purifier and leave clean inputs unchanged, as uncorrected degradations can introduce geometric artifacts (Table 3). **On the synthetic benchmark**, the router attains 88.45% recall and 90.77% accuracy, yielding the highest H of 89.59%, which exceeds the strongest baseline, MANIQA (Yang et al., 2022), by 5.55 percentage points. NIQE (Mittal et al., 2013), mPLUG-Owl2 (Ye et al., 2023), and DepictQA-Wild (You et al., 2024) exceed 98% recall, but their accuracies remain within one percentage point of the 53.7% obtained by outputting PURIFY for all inputs, resulting in H values of only 69.19–70.23%. MANIQA attains 89.04% accuracy but only 79.57% recall. Our router improves both, indicating that its higher H reflects better detection of inputs needing purification alongside more accurate routing overall. Zero-shot Qwen2.5-VL, which shares the backbone of our router, reaches $\text{H} = 78.64\%$. Training on $\mathcal{D}_{\text{route}}$ raises recall by 4.81 percentage points and accuracy by 16.56 points, improving H by 10.95 points. **On the in-the-wild benchmark**, the router ranks second in H (70.56%), with 89.29% recall and 58.33% accuracy against the proxy labels. DepictQA-Wild achieves the highest H (73.68%), but its 100% recall and 58.33% accuracy match the strategy of always predicting PURIFY. Thus, a high H alone does not establish effective preservation of KEEP inputs, and our router does not outperform this trivial strategy under the real-image proxy protocol. The high recall and lower accuracy are consistent with over-routing of inputs labeled KEEP to purification. This gap may reflect the use of clean renders for the KEEP class in $\mathcal{D}_{\text{route}}$. Whether adding real photographs to this class improves routing remains to be tested.

4.3 2D PURIFICATION RESULTS

Table 2 reports fidelity to clean references. All seven generic editors lower PSNR and SSIM and raise gradient MAE by 17.8%–125.9% relative to unedited inputs. Only FireRed (Qiao et al., 2026) lowers LPIPS, by 0.0004. Degradation removal thus comes with reduced fidelity to clean references. OmniGen2 (Wu et al., 2026) also reduces silhouette IoU. Gate3DGen alone improves all five metrics, lowering LPIPS by 39.1% and gradient MAE by 16.4% while raising IoU to 0.9758. The gains over the FireRed backbone support the benefit of our purification training (Section 3.4).

Table 1: **Degradation-aware routing on synthetic and real images (%).** Recall is measured for the PURIFY class, Accuracy covers both classes, and $\mathrm{H} = 2\,\mathrm{Recall}\,\mathrm{Accuracy}/(\mathrm{Recall}+\mathrm{Accuracy})$. GPT-5 supplies the real-image proxy labels and is excluded from evaluation against them (–).

	Benchmark			In-the-wild		
Method	Recall	Accuracy	H↑	Recall	Accuracy	H↑
mPLUG-Owl2	100.00	54.12	70.23	0.00	41.67	0.00
DepictQA-Wild	99.54	53.42	69.53	100.00	58.33	**73.68**
LIQE-MIX	51.66	65.48	57.75	17.86	48.96	26.17
NIQE	98.52	53.32	69.19	75.00	54.17	62.91
Qwen2.5-VL	83.64	74.21	78.64	14.29	43.75	21.54
MANIQA	79.57	89.04	84.04	12.50	45.83	19.64
GPT-5	72.27	77.58	74.83	–	–	–
Ours	88.45	90.77	**89.59**	89.29	58.33	<u>70.56</u>

Table 2: **Synthetic 2D purification.** Metrics use clean references. Input is unedited, Grad. is Sobel gradient MAE, and IoU is foreground mask overlap. * denotes a commercial closed-source model.

Method	LPIPS↓	PSNR↑	SSIM↑	Grad.↓	IoU↑
Degraded (Input)	0.2262	20.254	0.8323	0.0584	0.9537
FLUX.1	0.2577	17.992	<u>0.7907</u>	<u>0.0688</u>	0.9373
OmniGen2	0.4166	13.491	0.6478	0.1123	0.6238
Step1X	0.2426	18.828	0.7761	0.0781	0.8971
Qwen-Edit	0.2389	18.710	0.7812	0.0727	0.9537
Mage-Flow	0.2418	<u>19.175</u>	0.7755	0.0835	0.9530
GPT Image-2.5 Flare*	0.2710	16.571	0.7020	0.1319	0.9189
FireRed	<u>0.2258</u>	15.868	0.7677	0.0814	<u>0.9604</u>
Ours	**0.1378**	**22.514**	**0.8708**	**0.0488**	**0.9758**

4.4 DOWNSTREAM 3D GENERATION RESULTS

On TRELLIS.2 (Xiang et al., 2026) (Table 3), Gate3DGen lowers MD and CD by 61.5% relative to the degraded input, raises F_1^{all} and F_1^{outer} by 3.8% and 6.4%, and lowers shaded and albedo LPIPS by 10.9% and 10.6%, closing 83.1% of the MD gap to the clean-reference oracle. The baselines reveal a trade-off, as six of the seven image editing methods reduce MD yet all seven lower both F_1 scores below the unedited input. Large geometric artifacts, such as an occluder fused into the object (Figure 3), can increase MD, whereas F_1 at a small distance threshold is sensitive to local geometric deviations. Generic image editing can therefore reduce large artifacts while introducing local geometric deviations. Only Gate3DGen improves MD and both F_1 scores among the compared editors, consistent with the spatial constraints in Equation 8. Albedo LPIPS also improves, consistent with removing input illumination effects before they are incorporated into the generated material, while the residual gap to the oracle is largest for F_1^{all} (11.96 against 12.46), as geometry hidden by occlusion is only partially recoverable from a single view (Section E).

Figure 3 shows Pixal3D reconstructions. Unedited inputs fuse the hand into the doll. FireRed retains the occluder or recolors the teapot, while GPT Image-2.5 Flare removes degradations but re-synthesizes objects (red boxes). Gate3DGen purifies inputs and retains the objects (green boxes).

4.5 IN-THE-WILD RESULTS

We evaluate 96 mobile and web photographs without ground-truth meshes using VLM assessment, human preferences, and FD-DINOv2 (Stein et al., 2023). Qwen3.8-27B serves as the VLM judge (Chen et al., 2024), scoring Identity (object preservation), Purification (degradation removal),

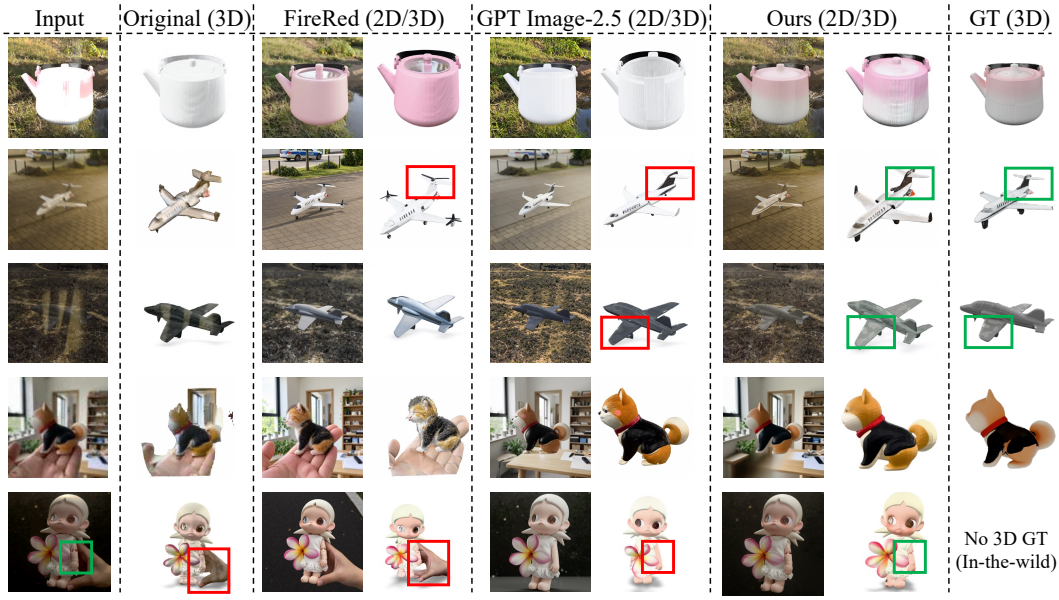


Figure 3: **Comparison of images (2D) and Pixal3D outputs (3D).** Original (3D) uses unedited inputs. Ours (2D) uses purified images. Boxes mark structure and appearance changes.

Table 3: **Downstream 3D metrics on TRELLIS.2.** Input is degraded, and Oracle uses clean references. MD, CD, and F_1^{all} are $\times 10^{-3}$, and F_1^{outer} is $\times 10^{-1}$. PSNR-N evaluates normals. LPIPS compares four-view shaded and albedo renderings. * denotes a commercial closed-source model.

Method	MD↓	F_1^{all}↑	CD↓	F_1^{outer}↑	PSNR-N↑	LPIPS↓	Albedo LPIPS↓
Clean reference (Oracle)	0.517	12.46	0.517	1.1586	16.36	0.1548	0.1575
Degraded (Input)	1.984	11.52	1.985	1.0531	15.26	0.1857	0.1916
FLUX.1	1.528	11.06	1.528	1.0238	15.35	0.1864	0.1935
OmniGen2	3.485	6.88	3.485	0.6407	13.11	0.2362	0.2411
Step1X	1.228	10.48	1.229	0.9637	15.25	0.1803	0.1861
Qwen-Image-Edit	1.047	10.81	1.047	1.0123	15.69	0.1780	0.1859
Mage-Flow	1.386	<u>11.42</u>	1.386	<u>1.0422</u>	15.39	0.1797	0.1855
GPT Image-2.5 Flare*	1.179	10.39	1.179	0.9720	15.29	0.1792	0.1859
FireRed	<u>0.835</u>	10.58	<u>0.835</u>	0.9889	<u>15.82</u>	<u>0.1701</u>	<u>0.1781</u>
Ours	**0.764**	**11.96**	**0.765**	**1.1205**	**15.98**	**0.1654**	**0.1714**

and Quality (asset quality independent of the input). Identity and Purification average 2D and 3D scores, while Quality averages 3D scores over four views and both frozen generators.

For FD-DINOv2, we fit Gaussians per generator to L2-normalized DINOv2 ViT-L/14-reg CLS features from the 96 photographs and 384 shaded renders (four per asset). Images are resized to 224×224 and ImageNet-normalized. Rendered views are separate samples without pooling.

In Table 4, Gate3DGen leads in VLM Identity (8.62) and Purification (8.00). FireRed and GPT Image-2.5 Flare score higher on Quality (6.92 versus 6.58), but lower on Identity (8.00 and 6.95, respectively). The VLM scores support object preservation and degradation removal.

The human study comprises 341 valid ratings from pairwise comparisons with each competing image editing method. Preferences for Gate3DGen exceed 50% on all three criteria against every baseline. Rates are lowest against GPT Image-2.5 Flare (58.5%, 57.1%, and 57.5% for Identity, Purification, and Quality). Human Quality preferences differ from the VLM ranking.

Table 4: **In-the-wild comparisons against 2D editing baselines.** FD-DINOv2 uses multi-view renders per generator. VLM scores span 0–10. Purification measures degradation removal in both VLM evaluation and the user study. Pref% reports preferences for Gate3DGen over each baseline, with values above 50% favoring ours. * denotes a commercial closed-source model.

	FD-DINOv2↓		VLM Scores↑			User Study (Pref%)↑		
Method	Pixal3D	TRELLIS.2	Identity	Purification	Quality	Identity	Purification	Quality
FLUX.1	0.795	0.790	7.31	6.22	6.24	77.5%	81.0%	72.1%
OmniGen2	0.893	0.902	5.43	4.51	4.72	86.0%	93.2%	92.3%
Step1X	0.788	0.789	7.91	7.09	5.98	81.8%	83.3%	86.8%
Qwen-Image-Edit	0.792	0.810	8.09	6.99	<u>6.60</u>	91.7%	94.6%	82.1%
Mage-Flow	0.739	0.769	<u>8.50</u>	<u>7.55</u>	6.22	84.2%	88.1%	80.0%
GPT Image-2.5 Flare*	<u>0.733</u>	**0.722**	6.95	6.31	**6.92**	58.5%	57.1%	57.5%
FireRed	0.771	0.796	8.00	7.08	**6.92**	66.7%	63.6%	63.6%
Ours	**0.716**	<u>0.757</u>	**8.62**	**8.00**	6.58	–	–	–

Table 5: **Purifier ablations on our synthetic benchmark.** We report 2D image fidelity and downstream 3D metrics from TRELLIS.2. Checkmarks indicate the use of router features (h) and spatial losses (region-weighted reconstruction and edge). MD, CD, and F_1^{all} are $\times 10^{-3}$, and F_1^{outer} is $\times 10^{-1}$. Shaded and albedo LPIPS are computed on the corresponding four-view renderings.

Method	h	Spatial losses	2D LPIPS↓	2D PSNR↑	MD↓	F_1^{all}↑	CD↓	F_1^{outer}↑	Shaded LPIPS↓	Albedo LPIPS↓
A			**0.1351**	22.292	0.820	**12.06**	0.820	1.1012	0.1794	0.1781
B	✓		<u>0.1363</u>	**22.572**	**0.755**	<u>11.96</u>	**0.755**	**1.1210**	<u>0.1658</u>	<u>0.1716</u>
Full	✓	✓	0.1378	<u>22.514</u>	<u>0.764</u>	<u>11.96</u>	<u>0.765</u>	<u>1.1205</u>	**0.1654**	**0.1714**

Among image editing methods, Gate3DGen ranks first in FD-DINOv2 on Pixal3D and second on TRELLIS.2, but none improves over unedited inputs. FD-DINOv2 compares feature distributions of inputs and renders, without directly measuring degradation removal or per-object fidelity. Its rankings should therefore be interpreted alongside the VLM scores and human preferences.

4.6 Ablation Studies

3D-oriented one-step purifier. Table 5 compares router features h and spatial losses, with $\mathcal{L}_1$ retained in all variants. A removes both components, B retains only h, and Full uses both. A achieves the best 2D LPIPS (0.1351) and F_1^{all} (12.06), but the worst MD, CD, F_1^{outer}, and rendering LPIPS scores. Its rendered LPIPS (0.1794) is comparable to Mage-Flow and GPT Image-2.5 Flare (Table 3), despite at least 0.1 lower 2D LPIPS. Thus, image similarity alone is insufficient to predict 3D quality. Adding h lowers MD from 0.820 to 0.755 (7.9%) and rendered LPIPS from 0.1794 to 0.1658 (7.6%), while increasing 2D LPIPS by 0.0012. With the spatial losses, Full has slightly higher MD (0.764 versus 0.755) and lower 2D PSNR than Variant B, but achieves the lowest shaded and albedo LPIPS (0.1654 and 0.1714). Full also preserves object contours more faithfully than B, supporting the full configuration. Sections B and D report routing and Pixal3D results.

5 Conclusion

Trained on image pairs synthesized from 3D assets, Gate3DGen combines a degradation-aware router and a 3D-oriented one-step purifier before frozen 3D generators. On 1,082 cases, its purifier outperforms seven image editing methods in image fidelity and surface geometry for TRELLIS.2 and Pixal3D, closing up to 83.1% of the mesh-distance gap to clean-reference inputs. VLM scores and human preferences on real photographs support identity preservation and degradation correction. Ablations and real-image comparisons motivate evaluating purification through downstream surface geometry, rendered appearance, and object fidelity beyond image similarity or visual quality alone.

AI USE STATEMENT

Generative AI tools assisted manuscript drafting and editing, LaTeX formatting, and scripts for figure layout. The data pipeline uses AI models for anchor-view selection, degradation synthesis, and structural consistency checks (Section 3.2). Evaluation on real images also uses VLM scores, reported separately from human preferences (Section 4.5). The authors reviewed the manuscript, claims, and figures and take responsibility for the final content, including the AI-assisted text and figures.

ETHICS STATEMENT

Input purification can improve 3D generation but may also change object identity or introduce unsupported details in occluded regions (Section E). Such outputs can misrepresent the photographed object and should not be treated as measurements of its geometry. Source assets and web photographs require attention to usage permissions, attribution, and privacy. Public availability alone does not establish permission to redistribute them. Synthetic training data and pretrained models may introduce biases, and VLM evaluation may reflect model-specific preferences. The human preference study in Section 4.5 also requires appropriate participant consent and privacy protection.

REPRODUCIBILITY STATEMENT

Section 3 describes data construction and both modules. Section 4.1 specifies the baselines, metrics, and evaluation protocol, while Section A provides trainable-layer configurations, loss weights, and purifier training and inference settings. Evaluation on real images and additional Pixal3D results appear in Section 4.5 and Section D, respectively. Routing decisions are evaluated separately from purification, and clean-reference inputs are identified as oracle controls. We will release the code.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025.

Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. SF3D: Stable fast 3D mesh reconstruction with UV-unwrapping and illumination disentanglement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 16240–16250. Computer Vision Foundation / IEEE, 2025.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6562–6595. PMLR / OpenReview.net, 2024.

Junhyeong Cho, Kim Youwang, Hunmin Yang, and Tae-Hyun Oh. Robust 3d shape reconstruction in zero-shot from a single image in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 22786–22798. Computer Vision Foundation / IEEE, 2025.

Marcos V. Conde, Gregor Geigle, and Radu Timofte. InstructIR: High-quality image restoration following human instructions. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXXVI*, Lecture Notes in Computer Science, pp. 1–21. Springer, 2024.

Zhao Dong, Ka Chen, Zhaoyang Lv, Hong-Xing Yu, Yunzhi Zhang, Cheng Zhang, Yufeng Zhu, Stephen Tian, Zhengqin Li, Geordie Moffatt, Sean Christofferson, James Fort, Xiaqing Pan,

Mingfei Yan, Jiajun Wu, Carl Yuheng Ren, and Richard A. Newcombe. Digital twin catalog: A large-scale photorealistic 3d object digital twin dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 753–763. Computer Vision Foundation / IEEE, 2025.

Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 2463–2471. IEEE Computer Society, 2017.

Hang Guo, Tao Dai, Yuanchao Bai, Bin Chen, Xudong Ren, Zexuan Zhu, and Shu-Tao Xia. Parameter efficient adaptation for image restoration with heterogeneous mixture-of-experts. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: large reconstruction model for single image to 3D. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

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. *CoRR*, 2025.

Zeqiang Lai, Yunfei Zhao, Haolin Liu, Zibo Zhao, Qingxiang Lin, Huiwen Shi, Xianghui Yang, Mingxin Yang, Shuhui Yang, Yifei Feng, Sheng Zhang, Xin Huang, Di Luo, Fan Yang, Fang Yang, Lifu Wang, Sicong Liu, Yixuan Tang, Yulin Cai, Zebin He, Tian Liu, Yuhong Liu, Jie Jiang, Linus, Jingwei Huang, and Chunchao Guo. Hunyuan3D 2.5: Towards high-fidelity 3D assets generation with ultimate details. *CoRR*, 2025.

Dong-Yang Li, Wang Zhao, Yuxin Chen, Wenbo Hu, Meng-Hao Guo, Fang-Lue Zhang, Ying Shan, and Shi-Min Hu. Pixal3D: Pixel-aligned 3D generation from images. In Christopher P. Redmann, Justin Solomon, and Mirela Ben-Chen (eds.), *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference, SIGGRAPH Conference Papers 2026, Los Angeles, CA, USA, July 19-23, 2026*, pp. 8:1–8:12. ACM, 2026.

Xiang Li, Zirui Wang, Zixuan Huang, and James M. Rehg. Cue3D: Quantifying the role of image cues in single-image 3D generation. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025a.

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, and Yan-Pei Cao. TripoSG: High-fidelity 3D shape synthesis using large-scale rectified flow models. *CoRR*, 2025b.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. DiffBIR: Towards blind image restoration with generative diffusion prior. *CoRR*, abs/2308.15070, 2023.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *CoRR*, abs/2504.17761, 2025.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013.

OpenAI. Openai GPT-5 system card. *CoRR*, abs/2601.03267, 2026.

OpenAI. GPT Image-2.5 Flare model documentation. OpenAI API documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-image-2.5-flare.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman H. Khan, and Fahad Shahbaz Khan. PromptIR: Prompting for all-in-one image restoration. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Changhao Qiao, Chao Hui, Chen Li, Cunzheng Wang, Dejia Song, Jiale Zhang, Jing Li, Qiang Xiang, Runqi Wang, Shuang Sun, Wei Zhu, Xu Tang, Yao Hu, Yibo Chen, Yuhao Huang, Yuxuan Duan, Zhiyi Chen, and Ziyuan Guo. Firered-image-edit-1.0 technical report. *CoRR*, abs/2602.13344, 2026.

Yansong Qu, Shaohui Dai, Xinyang Li, Yuze Wang, You Shen, Shengchuan Zhang, and Liujuan Cao. Deocc-1-to-3: 3d de-occlusion from a single image via self-supervised multi-view diffusion. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 8677–8685. AAAI Press, 2026.

Qwen Team. Qwen3-VL technical report. *CoRR*, abs/2511.21631, 2025.

Yukai Shi, Weiyu Li, Zihao Wang, Hongyang Li, Xingyu Chen, Ping Tan, and Lei Zhang. Scenemaker: Open-set 3d scene generation with decoupled de-occlusion and pose estimation model. *CoRR*, abs/2512.10957, 2025.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. TripoSR: Fast 3D object reconstruction from a single image. *CoRR*, 2024.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *CoRR*, abs/2508.02324, 2025.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ziyi Xia, Ze Liu, Chaofan Li, Haoge Deng, Kun Luo, Bo Zhang, Jiajun Zhang, Dong Liu, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. OmniGen2: Towards instruction-aligned multimodal generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21964–21975, June 2026.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D latents for scalable and versatile 3D generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 21469–21480. Computer Vision Foundation / IEEE, 2025.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. Native and compact structured latents for 3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14419–14429, 2026.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. InstantMesh: Efficient 3D mesh generation from a single image with sparse-view large reconstruction models. *CoRR*, 2024.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pp. 1190–1199. IEEE, 2022.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration. *CoRR*, abs/2311.04257, 2023.

Zhiyuan You, Jinjin Gu, Zheyuan Li, Xin Cai, Kaiwen Zhu, Tianfan Xue, and Chao Dong. Descriptive image quality assessment in the wild. *CoRR*, abs/2405.18842, 2024.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 25669–25680. IEEE, 2024.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 586–595. Computer Vision Foundation / IEEE Computer Society, 2018.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 14071–14081. IEEE, 2023.

Xinjie Zhang, Peng Zhang, Shicheng Zheng, Jinghao Guo, Zhaoyang Jia, Yifei Shen, Xun Guo, Yuxuan Luo, Jiahao Li, Wenxuan Xie, Fanyi Pu, Xiaoyi Zhang, Kaichen Zhang, Zongyu Guo, Tianci Bi, Dongnan Gui, Zhening Liu, Zimo Wen, Zihan Zheng, Senqiao Yang, Xiao Li, Jinglu Wang, Bin Li, and Yan Lu. Mage-Flow: An efficient native-resolution foundation model for image generation and editing. *CoRR*, 2026.

Yibo Zhang, Li Zhang, Rui Ma, and Nan Cao. TexVerse: A universe of 3d objects with high-resolution textures. *CoRR*, abs/2508.10868, 2025.

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, Huiwen Shi, Sicong Liu, Junta Wu, Yihang Lian, Fan Yang, Ruining Tang, Zebin He, Xinzhou Wang, Jian Liu, Xuhui Zuo, Zhuo Chen, Biwen Lei, Haohan Weng, Jing Xu, Yiling Zhu, Xinhai Liu, Lixin Xu, Changrong Hu, Tianyu Huang, Lifu Wang, Jihong Zhang, Meng Chen, Liang Dong, Yiwen Jia, Yulin Cai, Jiaao Yu, Yixuan Tang, Hao Zhang, Zheng Ye, Peng He, Runzhou Wu, Chao Zhang, Yonghao Tan, Jie Xiao, Yangyu Tao, Jianchen Zhu, Jinbao Xue, Kai Liu, Chongqing Zhao, Xinming Wu, Zhichao Hu, Lei Qin, Jianbing Peng, Zhan Li, Minghui Chen, Xipeng Zhang, Lin Niu, Paige Wang, Yingkai Wang, Haozhao Kuang, Zhongyi Fan, Xu Zheng, Weihao Zhuang, YingPing He, Tian Liu, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, Jingwei Huang, and Chunchao Guo. Hunyuan3D 2.0: Scaling diffusion models for high resolution textured 3D assets generation. *CoRR*, abs/2501.12202, 2025.

Gate3DGen: A Pre-processing Model that Selectively Purifies Real-World Images for 3D Generation

Supplementary Materials

A METHOD DETAILS

Under a fixed prompt, the degradation-aware router is trained to output the PURIFY/KEEP decision described in Section 3.3. These decisions correspond to the implementation tokens `Purify` and `Keep`, respectively. It updates language-transformer layers 24–27 and the final language normalization. The language-model head and visual backbone are frozen except for rank-16 LoRA adapters in vision blocks 28–31. Training uses token-level supervision with higher weights on the decision tokens and the end-of-sequence token, and with higher sample weights on rare degradation families and on subtle degradations. The discrete routing indicator $\widehat{\gamma}_i$ follows Equation 4, with the threshold τ frozen on the development set. The language-only router variant omits the vision LoRA adapters.

For the purifier $\mathcal{C}_\theta$, image-transformer blocks 1–28 remain frozen and blocks 29–49 are updated together with the cross-attention projection layers receiving h. Training uses the flow-matching endpoint at $t = 1$ in Equation 5, estimates $\widehat{z}_0$, and reconstructs $\widehat{x}^c = \mathcal{D}(\widehat{z}_0)$ by Equation 6. The subject mask m_i is the 8-bit soft alpha predicted by RMBG-2.0 from x_i^c. It is normalized to $[0, 1]$ and receives the same resize and random crop as the paired images. The degradation map r_i averages $|x_i^o - x_i^c|/2$ over RGB channels and applies 7×7 average pooling with stride 1 and padding 3. In Equation 7, the single-channel weight is broadcast over RGB channels, and each weighted sum is divided by its own broadcast weight sum, clamped below at 1. For the edge term, D_u and D_v are one-pixel forward differences, with m_i^u and m_i^v given by the products of the two subject-mask values adjacent to each horizontal or vertical edge, respectively. The loss weights in Equation 9 are $\lambda_{\text{lat}} = 0.1$, $\lambda_1 = 1$, $\lambda_{\text{lpips}} = 0.5$, $\lambda_{\text{region}} = 1$, and $\lambda_{\text{edge}} = 1$. We train at 1024×1024 resolution for 15,000 updates on 32 GPUs (global batch size 32, BF16 precision, AdamW optimizer, cosine learning-rate schedule, peak learning rate 5×10^{-7}). At inference, image purification uses a single forward pass with guidance scale 1.0, an empty negative prompt, and a fixed random seed.

B ROUTER ABLATION

Table 6 compares the full router with an MLP on VAE latents and with a language-only variant that omits vision LoRA. The MLP achieves 53.42% recall. Removing vision LoRA raises recall to 97.04% but lowers accuracy to 75.30%, indicating a stronger bias toward PURIFY. The full router achieves 88.45% recall and 90.77% accuracy. Vision LoRA reduces this bias with high recall.

Table 6: **Router ablations on synthetic evaluation inputs.** Variants use an MLP on VAE latents or omit vision LoRA. Recall and accuracy are percentages. **Bold** marks column maxima.

Method	Recall↑	Accuracy↑
MLP on VAE latents	53.42	71.48
Without vision LoRA	**97.04**	75.30
Full router	88.45	**90.77**

C Relative Performance across 2D and 3D Metrics

Figure 4 provides a unified view of the 2D purification metrics in Table 2 and the downstream TRELLIS.2 metrics in Table 3. The observed degraded input is the zero reference and the matched clean input is the oracle reference at one; the direction of lower-is-better metrics is reversed so that larger radial values consistently indicate better performance. Among the representative editing baselines shown, Gate3DGen makes positive progress toward the clean oracle on every metric. In contrast, FireRed, GPT Image-2.5 Flare, and Mage-Flow exhibit negative changes on several 2D fidelity or fine-grained 3D geometry measures, illustrating the trade-off between visible degradation removal and preservation of reconstruction cues.

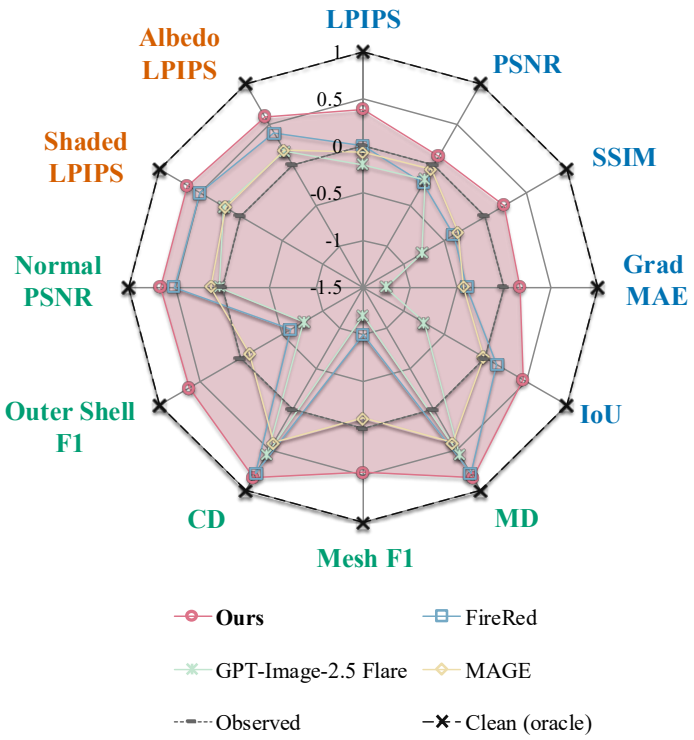


Figure 4: **Oracle-normalized comparison of 2D purification and downstream 3D generation.** The radar plot summarizes percentage progress from the observed degraded input (0) toward the matched clean-reference oracle (1) using the metrics reported in Tables 2 and 3; negative values indicate deterioration relative to the observed input. Blue labels denote 2D image metrics, green labels denote 3D geometry and normal metrics, and orange labels denote rendered-appearance metrics. Lower-is-better metrics are direction-adjusted so that farther outward is always better. GPT-Image-2 and MAGE in the legend correspond to GPT Image-2.5 Flare and Mage-Flow, respectively.

D ADDITIONAL PIXAL3D RESULTS

The main paper reports downstream 3D metrics on TRELLIS.2 (Table 3). Table 7 repeats the same evaluation on Pixal3D (Li et al., 2026), using the same $1,082$ cases and the same metric definitions. Gate3DGen obtains the best MD, F_1^{all}, CD, F_1^{outer}, and albedo LPIPS among image editing methods. Despite these gains in surface geometry, Gate3DGen does not achieve the best results on all rendering metrics on Pixal3D. FireRed obtains higher normal-map PSNR and lower shaded LPIPS, while our albedo LPIPS is only marginally lower (0.2773 versus 0.2775). Despite gains in surface geometry, improvements in rendered appearance are smaller on Pixal3D than on TRELLIS.2.

Table 7: **Downstream 3D metrics on Pixal3D.** Input and Oracle use degraded and clean-reference images, respectively. MD and CD are $\times 10^{-3}$. F_1^{all} is all-surface $F_1@10^{-8}$ ($\times 10^{-3}$). F_1^{outer} is outer-surface $F_1@10^{-6}$ ($\times 10^{-1}$). PSNR-N evaluates rendered normals. LPIPS and albedo LPIPS use four-view shaded and albedo renderings, respectively. **Bold** and underline mark the best and second-best editing results, excluding Input and Oracle. * denotes a commercial closed-source model.

Method	MD↓	F_1^{all}↑	CD↓	F_1^{outer}↑	PSNR-N↑	LPIPS↓	Albedo LPIPS↓
Degraded (Input)	1.221	9.37	1.221	0.8821	10.31	0.2824	0.2805
Clean reference (Oracle)	0.753	10.75	0.753	1.0086	10.42	0.2760	0.2713
FLUX.1	1.214	9.06	1.214	0.8582	10.26	0.2916	0.2906
Mage-Flow	1.259	8.88	1.259	0.8428	10.30	0.2883	0.2864
OmniGen2	3.204	6.00	3.204	0.5700	10.03	0.3051	0.3047
Step1X	1.411	8.79	1.412	0.8289	10.20	0.2911	0.2897
GPT Image-2.5 Flare*	1.467	8.29	1.468	0.7864	10.53	0.2825	0.2836
Qwen-Image-Edit	1.150	9.29	1.150	0.8757	10.31	0.2889	0.2883
FireRed	1.148	9.12	1.149	0.8639	**10.77**	**0.2759**	0.2775
Ours	**1.091**	**10.16**	**1.091**	**0.9588**	10.36	0.2796	**0.2773**

E LIMITATIONS AND FUTURE WORK

With foreground occlusions larger than 60% or severe truncation, single-view generative priors can lead the purifier to hallucinate symmetric geometry in occluded regions. Recovering clean appearance for refractive materials and inter-reflections from one view remains underconstrained without inverse rendering. The router can misclassify clean real photographs and trigger unnecessary purification (Section 4.2). We will explore multi-view and video-to-3D generation. Further explorations may focus on multi-view and video-to-3D generation.