CoRest: Multi-View Image Restoration with Geometric Guidance

Yu-Ju Tsai* Lee Hsin-Ying* Ming-Hsuan Yang
University of California, Merced

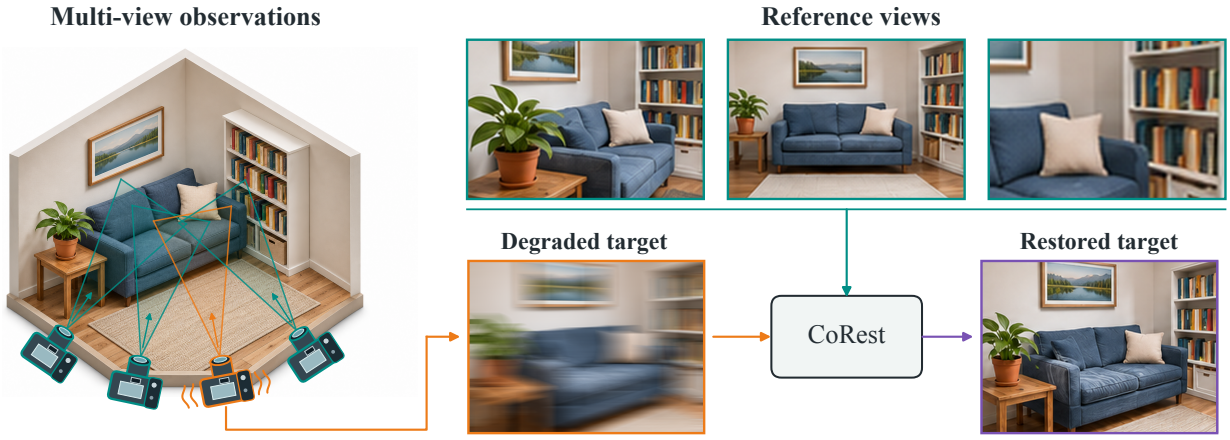

Figure 1. **Restoration with same-scene references.** During multi-view capture, camera shake may blur some photographs while other viewpoints retain useful details. CoRest uses these same-scene observations as references to restore a degraded target. Restored outputs can support subsequent targets, progressively improving the reference pool.

Abstract

Multi-view image collections often contain both sharp and degraded observations of the same scene. Restoring each image independently overlooks details visible in other views, while generative restoration may introduce plausible content that does not match the scene. Using complementary views is challenging because viewpoint changes, partial overlap, and varying image quality make useful reference content difficult to associate with the target. We propose CoRest, a reference-conditioned restoration method that adapts a pretrained video generative model to recover degraded images using same-scene observations. During training, a frozen VGGT encoder associates target and reference regions. We use these associations to retrieve appearance features from the restoration model's conditioning pathway, providing geometric guidance. At inference, CoRest restores targets using reference conditioning, without running VGGT. Restored views become references for subsequent targets, allowing the method to progressively update the available scene evidence. Experiments on DL3DV, ScanNet++, and ETH3D demonstrate strong perceptual restoration, and the restored images also improve consistency with depth predictions from clean images and increase feature-matching support, showing their utility beyond visual appearance.

1. Introduction

A collection of photographs often records the same scene with uneven image quality. Camera motion may blur some views while leaving nearby views sharp. Restoring each photograph independently [1, 16, 27, 38] ignores a useful source of evidence: details obscured in the target may still be visible in another observation. This evidence is especially valuable for generative restoration. A model can produce plausible detail that differs from the observed scene, and such changes can also disrupt the correspondences used by downstream vision systems. Therefore, we study the restoration of a degraded target using reference images from other viewpoints of the same scene, as shown in Fig. 1.

Using multiple views introduces its own challenge. References differ from the target in viewpoint and visibility, and some may themselves be degraded. A useful feature may appear at a different location in another view or be absent altogether. Recent approaches demonstrate the potential of

*Equal contribution

joint multi-view restoration: SIR-Diff [20] restores sparse image sets, while GARD [12] restores geometry-aware representations for image and geometry prediction. A central challenge is learning which reference regions provide useful evidence for a degraded target. Simply supplying more images does not make these associations explicit. We address this challenge by using geometric associations to guide how a restoration model learns from same-scene references.

To this end, we propose *CoRest*, a reference-conditioned restoration method that adapts a pretrained video generative model to restore images using other views of the same scene [8, 28]. Video generative models learn to relate visual content across frames, providing a useful starting point for combining information across viewpoints. We present the degraded target and selected references as a sequence of scene views, with the target designated for restoration and the references providing scene context. The model is conditioned on these observed images to recover target details using evidence available in the references. We train lightweight low-rank adaptation (LoRA) [5] modules to learn this restoration task while keeping the pretrained model weights frozen.

During training, we use a frozen VGGT encoder [30] to associate target regions with relevant reference regions. We use these associations to retrieve reference appearance features from the model's image-conditioning pathway. A learned adapter converts the retrieved features into an update to the target's image-conditioning features. Geometry therefore guides where to obtain reference evidence, while the evidence itself remains in the representation used by the restoration model.

During inference, *CoRest* retains the learned LoRA parameters and reference conditioning, without running the VGGT retrieval branch. The adapted model uses the supplied reference images directly to restore the target. For scene restoration, we select references using image matches and process degraded targets sequentially, replacing each degraded image in the reference pool with its restored output. Later targets can thus draw on both original observations and previously restored views.

We mainly evaluate motion deblurring on DL3DV [17], ScanNet++ [35], and ETH3D [24], reporting three degraded-view fractions separately. *CoRest* achieves strong perceptual restoration across these datasets, and evaluations on DL3DV and ETH3D show that its outputs also improve depth consistency and feature-matching support over degraded inputs. Brightness and mixed-degradation experiments in the supplementary material further demonstrate the framework's applicability beyond motion blur.

Our contributions are:

- We propose *CoRest*, a multi-view reference-based restoration framework that adapts a video generative prior using same-scene references.
- We introduce geometric guidance for LoRA training, enabling reference-conditioned restoration without using VGGT retrieval at inference.
- We evaluate *CoRest* through extensive experiments and ablation studies, demonstrating state-of-the-art quantitative and qualitative performance and adaptability to different degradation types.

2. Related Work

Multi-View and Reference-Based Image Restoration. Image restoration combines evidence in the input with learned priors about image appearance. Diffusion-based methods provide strong generative priors: DDRM [10] addresses linear inverse problems through posterior sampling, DPS [2] extends diffusion posterior sampling to more general noisy inverse problems, and DiffBIR [16] combines degradation removal with generative reconstruction. Such priors help recover plausible detail. Additional observations can further constrain which detail belongs to the target scene.

Reference-based restoration explicitly transfers information from another image. In super-resolution, TTSR [34] uses attention to transfer relevant textures, while C2-Matching [7] learns correspondences robust to geometric transformations and resolution differences. Video and multi-view restoration exploit related evidence through alignment and joint processing. RVRT [13] combines recurrent propagation with guided deformable attention, and RC-Net [19] couples low-light enhancement with recurrent cross-view alignment and fusion. More recently, SIR-Diff [20] jointly restores sparse image sets, and GARD [12] restores geometry-aware representations for geometry prediction and RGB decoding. These approaches motivate using complementary observations while accounting for their misalignment and varying quality. *CoRest* adopts a target-and-reference formulation and studies geometric retrieval as guidance for adapting a video generative prior.

Geometric Guidance for Generative Models. Geometry can organize the generation of multiple views. MV-Dream [25] combines camera embeddings with cross-view attention, SyncDreamer [18] couples denoising states through 3D-aware feature attention, and CAT3D [4] conditions on observed images and camera rays to synthesize views for reconstruction. In these methods, geometric structure helps coordinate the images produced during generation.

Learned representations provide another route for incorporating geometric information. DUSt3R [31] predicts pointmaps from uncalibrated image pairs, and VGGT [30] predicts geometric attributes from image collections. Conversely, DIFT [26] extracts correspondences from pretrained diffusion features, showing that generative representations can carry information about relationships between images. External representations can also guide what a generator learns. REPA [36] aligns denoising features with pretrained image representations, VideoREPA [41] transfers

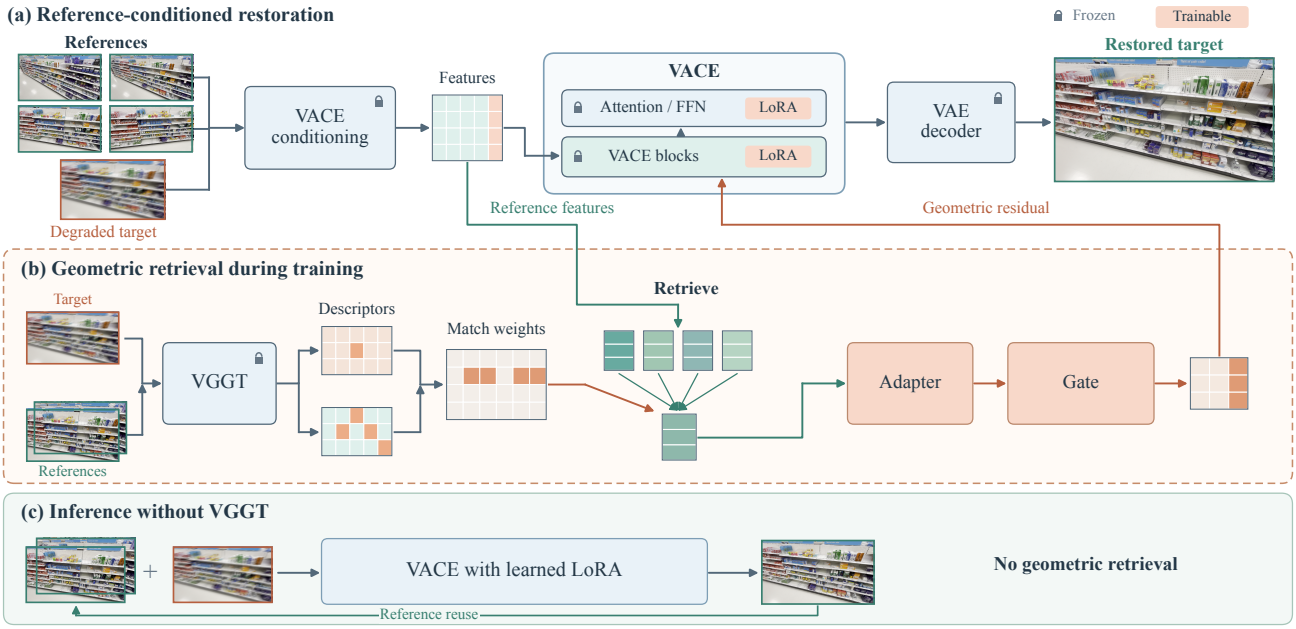

Figure 2. **CoRest overview.** (a) References and a degraded target condition VACE with LoRA. (b) During training, frozen VGGT descriptors select four tokens across all references for each target token. Their VACE features are aggregated, adapted, and weighted by confidence and a timestep gate before addition to each target control hint. Training uses flow matching and an auxiliary feature loss. (c) Inference retains LoRA and reference conditioning, omits VGGT, the adapter, and the gate, and reuses restored views.

spatiotemporal token relations, and Geometry Forcing [33] aligns video diffusion features with geometric representations. *CoRest* uses VGGT descriptors to associate target and reference regions, but retrieves appearance features from the restoration model's conditioning pathway rather than directly aligning the generator to VGGT features. We evaluate whether the resulting restoration benefit persists after explicit retrieval is removed.

3. *CoRest*

CoRest adapts a video generative model to restore a degraded target using same-scene references, as shown in Fig. 2. Geometric retrieval guides adaptation during training; inference retains the learned LoRA and reference conditioning without the retrieval branch.

3.1. Reference-Conditioned Restoration

Given a degraded target $\widetilde{I}$ and R ordered same-scene reference images $\mathcal{R} = (I_1, \ldots, I_R)$, we predict a restored image $\widehat{I}$. References may differ in viewpoint and quality; the clean target I supplies training supervision.

We use VACE [8], an image-conditioned video generation framework built on Wan [28]. Its conditioning pathway encodes observed images and supplies feature updates, called *control hints*, to the generative backbone. We arrange the inputs as $[I_1, \ldots, I_R, \widetilde{I}]$, with target slot $T = R + 1$. This sequence indexes viewpoints rather than consecutive video frames. A binary conditioning mask is one over the entire target view and zero over every reference view, identifying which view to restore while retaining the degraded image as conditioning. Each view is encoded independently by the variational autoencoder (VAE), avoiding temporal compression across views. VACE encodes the masked and complementary image components separately and combines their latents with the mask. Its conditioning patch embedding produces a spatial token grid. The model processes all views jointly, and we retain the decoded target output $\widehat{I}$.

We adapt the pretrained model for reference-conditioned restoration through low-rank adaptation (LoRA) [5], learning weight updates while keeping the pretrained parameters frozen. The learned updates adapt the video generative prior to restoration, while reference conditioning provides scene-specific evidence during both training and inference.

3.2. Geometrically Guided Adaptation

Our training-time branch jointly encodes the observed references $\mathcal{R}$ and degraded target $\widetilde{I}$ with frozen VGGT [30]. VGGT descriptors identify corresponding regions, and we retrieve their VACE appearance features to guide restoration. Thus, geometry determines the associations while the retrieved content stays in the restoration model's feature space. The associations use learned VGGT descriptors, without camera poses or explicit depth maps. VGGT receives only the observed inputs, never the clean target.

Cross-view retrieval. Let $r \in \{1, \dots, R\}$ index references and $v \in \{1, \dots, R, T\}$ index all views. For view v, f_v and g_v are VGGT feature maps after within-view and cross-view attention at an intermediate layer. We bilinearly resample their difference to the VACE token grid using $\mathcal{B}$, then apply per-token ℓ_2 normalization. At spatial position i, the resulting descriptor is $d_{v,i} = \mathrm{Normalize}([\mathcal{B}(g_v - f_v)]_i)$. For target position i and reference position j, their cosine similarity is $s_{i,rj} = d_{T,i}^\top d_{r,j}$. Let $\mathcal{N}_K(i)$ contain the K reference index pairs (r, j) with highest similarity across all references. For $(r, j) \in \mathcal{N}_K(i)$, we define normalized weights $a_{i,rj}$ with temperature $\tau > 0$ and aggregate reference features $C_{r,j}$ from the frozen VACE conditioning patch embedding into V_i:

$$\begin{aligned} a_{i,rj} &= \frac{\exp(s_{i,rj}/\tau)}{\sum_{(r',j') \in \mathcal{N}_K(i)} \exp(s_{i,r'j'}/\tau)}, \\ V_i &= \sum_{(r,j) \in \mathcal{N}_K(i)} a_{i,rj} C_{r,j}. \end{aligned} \tag{1}$$

Only VGGT descriptors are resampled; VACE appearance features remain on their native grid. K counts retrieved tokens, not reference images.

Injecting reference evidence. An adapter A, comprising layer normalization and a multilayer perceptron (MLP) with two linear layers, maps V_i to the control-hint feature space. At diffusion timestep t, a separate MLP defines the scalar gate $q(t) = \tanh(\mathrm{MLP}(t/1000))$. Training applies inverted dropout with probability p: a scalar $b \sim \mathrm{Bernoulli}(1 - p)$ is sampled per example and shared across target tokens. Let c_i denote match confidence, with similarity threshold θ and temperature $\tau_c > 0$. The target correction $\Delta_i(t)$ is

$$\begin{aligned} c_i &= \mathrm{sigmoid}\left(\frac{\max_{r,j} s_{i,rj} - \theta}{\tau_c}\right), \\ \Delta_i(t) &= \frac{b}{1-p}\, q(t) c_i A(V_i). \end{aligned} \tag{2}$$

The maximum ranges over all reference tokens. We place $\Delta_i(t)$ at target positions and zeros at reference positions to form the sequence-shaped residual $\Delta(t)$. Let H_ℓ denote the ordinary VACE control hint supplied to backbone block ℓ. We modify each hint before its addition to the backbone:

$$H_\ell \leftarrow H_\ell + \Delta(t). \tag{3}$$

The same residual augments all control hints only at target positions. Confidence attenuates weak matches, while the timestep gate controls the correction across noise levels. Dropping retrieval exposes the model to training examples without this correction while keeping ordinary reference conditioning active.

3.3. Training Objective

Let ϕ collect the trainable LoRA, adapter, and gate parameters; all pretrained parameters remain frozen. Let z_0 be the VAE latent sequence of clean training views, including target I, and let ϵ be standard Gaussian noise of the same shape. At scheduler noise level σ_t, we construct $z_t = (1 - \sigma_t) z_0 + \sigma_t \epsilon$. The model predicts velocity $v_\phi(z_t, t; \mathcal{R}, \widetilde{I})$, conditioned on the observed images and the training-time retrieval residual. Subscript T selects the target slot. Let $w(t)$ be the scheduler weight, and let $\mathbb{E}$ average over training samples, noise, timesteps, and retrieval dropout. With MSE averaging squared error over target latent elements, flow matching minimizes

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\Big[w(t)\,\mathrm{MSE}\Big(v_{\phi,T}(z_t, t; \mathcal{R}, \widetilde{I}), (\epsilon - z_0)_T\Big)\Big]. \tag{4}$$

This loss jointly trains LoRA, the adapter, and the gate, allowing the restoration model to learn while receiving geometrically retrieved evidence. Clean latents define supervision, not observed conditioning inputs.

An auxiliary loss trains the retrieved correction to move degraded-target features toward clean-target features. Let C_i^{deg} and C_i^{clean} be target patch features obtained by conditioning on $\widetilde{I}$ and I, respectively, with the same frozen VACE embedding. For this auxiliary loss, these features and confidences c_i are constants; gradients flow through $\Delta_i(t)$ to the adapter and gate, but not to LoRA. Using Smooth L1 loss $\rho_{0.1}$ with beta 0.1, averaged over feature channels, we define

$$\mathcal{L}_{\mathrm{ret}} = \frac{\sum_i c_i\, \rho_{0.1}\Big(C_i^{\mathrm{deg}} + \Delta_i(t) - C_i^{\mathrm{clean}}\Big)}{\max(\sum_i c_i, 10^{-6})}. \tag{5}$$

Sums cover all target tokens across the minibatch, with batch indices omitted. This loss requires no additional spatial mask because all target tokens are supervised; the VACE conditioning mask remains unchanged. The loss supervises the actual gated residual, including dropout. The total objective is $\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + 0.1\mathcal{L}_{\mathrm{ret}}$; hyperparameters are given in the supplementary material.

3.4. Inference and Scene Restoration

Inference retains LoRA and reference conditioning, omitting VGGT, the adapter, and the gate. Given a scene and a supplied list of degraded views, we use OmniGlue [6] to select reference images and restore each target once. Match counts score target–reference pairs, with reduced scores between two unrestored degraded views. We choose the remaining target with the largest sum of positive scores and its R highest-scoring references, ordered by increasing score. Its restored output replaces the degraded input in the reference pool, and we recompute its pairwise scores before selecting the next target. The process ends when all identified degraded views have been restored.

Table 1. **Motion-blur restoration on three datasets.** Scores average degraded targets only; fractions denote their proportion among scene views, not blur strength. Colors indicate first and second place among eight methods per dataset and fraction. Degraded input is an unranked reference. *CoRest* uses no VGGT at inference. * denotes our SIR-Diff reproduction. DreamSim is $100(1 - \text{distance})$.

Fraction	Method	DL3DV				ScanNet++				ETH3D			
		LPIPS↓	DreamSim↑	NIQE↓	MUSIQ↑	LPIPS↓	DreamSim↑	NIQE↓	MUSIQ↑	LPIPS↓	DreamSim↑	NIQE↓	MUSIQ↑
25%	Degraded input	0.541	68.33	6.067	19.00	0.403	71.67	5.967	19.64	0.458	74.21	5.588	24.95
	NAFNet	0.495	73.48	5.361	27.18	0.395	76.14	5.524	25.19	0.424	78.76	5.126	32.64
	MoCE-IR	0.504	72.63	5.636	24.99	0.390	75.47	5.632	24.39	0.438	76.90	5.421	29.69
	R2R	0.506	74.32	5.675	24.85	0.391	78.77	5.587	24.78	0.435	79.62	5.337	30.98
	VRT	0.526	71.55	5.111	24.15	0.407	67.93	4.907	21.46	0.466	71.97	4.768	28.23
	BSSTNet	0.518	73.94	5.482	24.65	0.410	73.35	5.269	22.18	0.463	74.44	5.144	28.76
	SIR-Diff*	0.509	76.22	6.469	26.56	0.404	75.82	5.870	24.45	0.430	80.69	5.864	34.29
	GARD	0.479	85.47	3.273	57.23	0.448	76.58	3.545	57.70	0.414	86.18	3.246	61.38
	CoRest	0.347	95.45	3.150	71.18	0.445	83.03	4.124	69.84	0.282	96.61	3.726	72.21
50%	Degraded input	0.546	67.56	6.054	18.72	0.403	73.72	5.949	19.85	0.490	72.90	5.735	23.39
	NAFNet	0.505	72.14	5.607	26.71	0.391	78.28	5.500	26.33	0.456	77.24	5.377	33.08
	MoCE-IR	0.511	71.66	5.628	24.42	0.389	77.60	5.623	25.11	0.461	75.88	5.508	29.84
	R2R	0.513	73.28	5.689	24.49	0.391	80.20	5.577	25.76	0.464	78.24	5.555	29.66
	VRT	0.532	71.24	5.143	23.07	0.411	68.25	4.882	21.64	0.499	71.73	4.797	25.49
	BSSTNet	0.525	73.54	5.517	23.77	0.408	74.28	5.238	22.61	0.489	75.55	5.086	27.19
	SIR-Diff*	0.519	75.08	6.589	25.51	0.403	77.72	5.831	25.41	0.462	79.95	6.014	30.52
	GARD	0.486	84.22	3.237	56.87	0.450	77.12	3.381	58.21	0.436	85.68	3.129	61.38
	CoRest	0.369	94.53	3.177	71.20	0.449	81.66	4.142	70.98	0.329	93.33	3.859	72.00
75%	Degraded input	0.548	67.31	6.093	18.56	0.402	72.76	5.949	19.76	0.500	72.37	5.913	21.59
	NAFNet	0.505	72.20	5.599	26.54	0.389	77.58	5.484	26.42	0.462	77.12	5.968	32.47
	MoCE-IR	0.512	71.61	5.639	24.33	0.388	76.70	5.611	25.20	0.472	76.45	5.659	28.00
	R2R	0.514	73.27	5.699	24.27	0.390	79.24	5.575	25.54	0.469	78.69	5.563	28.68
	VRT	0.534	71.29	5.203	22.07	0.412	66.30	4.799	21.61	0.512	71.42	4.874	24.15
	BSSTNet	0.528	73.50	5.585	23.08	0.406	73.74	5.224	22.76	0.506	74.34	4.967	26.12
	SIR-Diff*	0.525	74.75	6.656	24.74	0.402	76.85	5.860	25.45	0.473	80.38	5.995	28.38
	GARD	0.492	83.75	3.184	56.71	0.452	75.79	3.412	57.71	0.452	85.36	3.112	60.92
	CoRest	0.397	92.91	3.279	70.86	0.449	80.11	4.251	70.37	0.362	92.71	3.966	71.27

4. Experiments

4.1. Experimental Setup

Implementation and training. We adapt Wan 2.1 1.3B with VACE conditioning [8, 28], using rank-16 LoRA [5] in the self-attention and feed-forward layers of both the backbone and VACE blocks. Training uses Hypersim [23] and TartanGround [22], with four references and one synthetically blurred target per sample. Each reference is independently degraded with probability 0.25. We jointly train LoRA, the retrieval adapter, and the timestep gate for 1,000 steps with learning rate 10^{-4} and effective batch size 32, keeping the other pretrained parameters frozen. Restoration inputs have a 480-pixel short side, with dimensions aligned to the model grid. *CoRest* uses reference conditioning and learned LoRA at inference, without VGGT. Further implementation details are provided in the supplementary material.

Datasets and baselines. We evaluate motion blur on 50 DL3DV scenes [17], 49 ScanNet++ scenes [35], and seven ETH3D scenes [24]. The three conditions degrade 25%, 50%, or 75% of the views; these percentages describe the prevalence of degraded views, not blur strength. All methods are scored on the same fixed degraded targets, excluding unaffected views, separately for each dataset and fraction. We compare NAFNet [1], MoCE-IR [37], and R2R [29] as single-image methods, and VRT [14], BSSTNet [39], SIR-Diff [20], and GARD [12] as methods using multiple images. SIR-Diff* denotes our reproduction. Input policies differ: SIR-Diff processes consecutive groups of four views, GARD processes half-scenes of 8 or 12 views, and *CoRest* selects four references from the full scene and reuses restored views. Target counts and additional baseline settings are provided in the supplementary material.

Restoration metrics. We report LPIPS [40], DreamSim [3], NIQE [21], and MUSIQ [11]. LPIPS compares predictions resized to the unchanged native ground-truth resolution. DreamSim uses its model preprocessing and is expressed as $100(1 - \text{distance})$. NIQE and MUSIQ use native outputs and their default preprocessing. Ablations report PSNR, SSIM [32], and LPIPS. Restoration scores are target means within each condition. Detailed metric configurations and preprocessing are provided in the supplementary material.

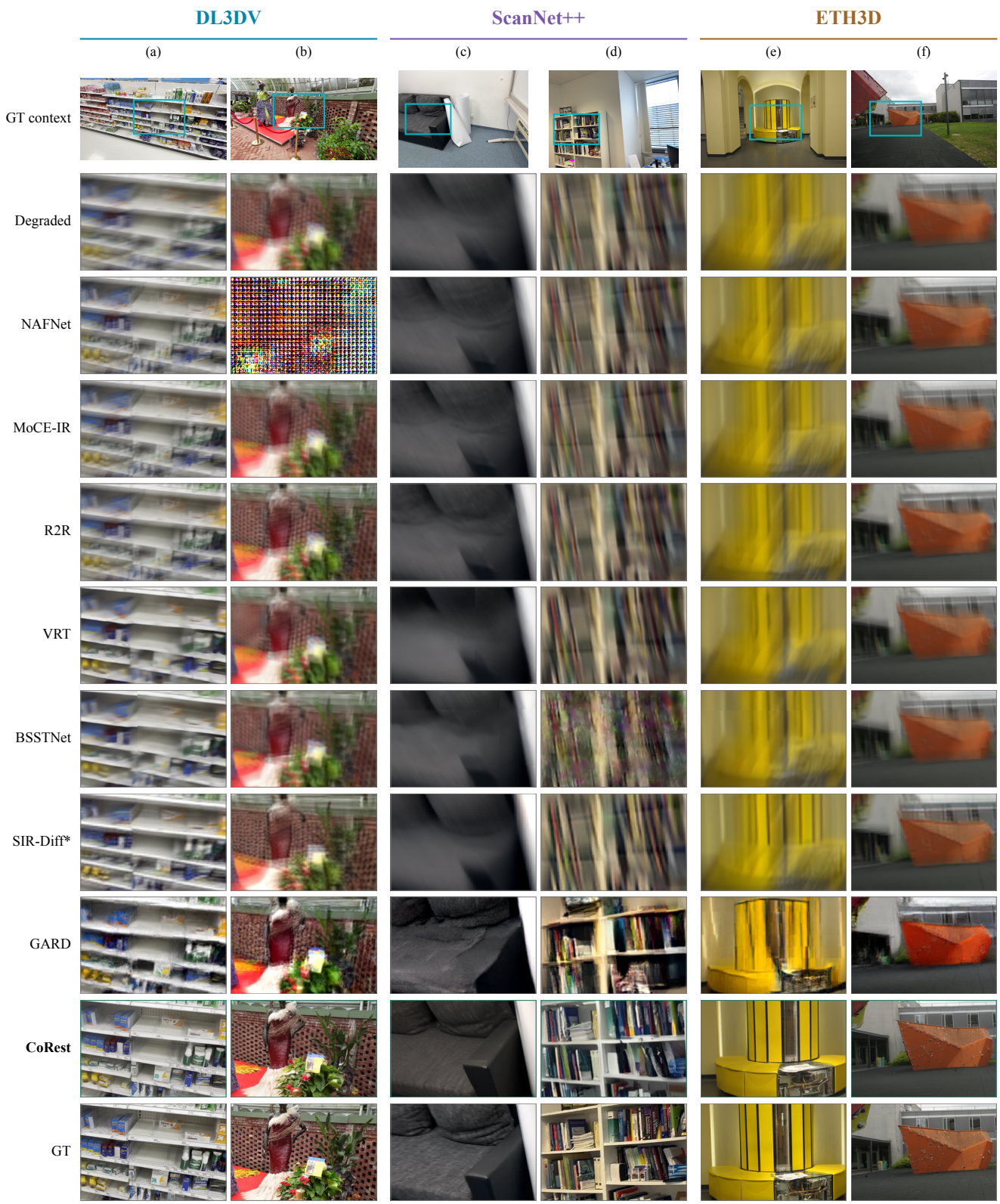

Figure 3. **Qualitative comparison with 50% motion-blurred views.** Each dataset has two target columns. Ground-truth (GT) views mark shared crops; predictions are resized bicubically to native GT resolution. * denotes our SIR-Diff reproduction.

Table 2. **Downstream depth consistency at 50% degraded views.** For motion blur, DA3 predictions are aligned by one positive scale per scene and compared with clean-input predictions, not measured depth. Results on DL3DV and ETH3D use equal scene weighting. nRMSE is divided by the scene median reference depth. Colors indicate first and second place per dataset. * denotes our SIR-Diff reproduction. Complete results are provided in the supplementary material.

Method	DL3DV			ETH3D		
	AbsRel↓	nRMSE↓	δ_1 ↑	AbsRel↓	nRMSE↓	δ_1 ↑
Degraded input	0.122	0.505	0.827	0.114	0.347	0.842
NAFNet	0.122	0.493	0.827	0.118	0.340	0.838
MoCE-IR	0.118	0.488	0.832	0.110	0.338	0.845
R2R	0.112	0.474	0.840	0.101	0.317	0.862
VRT	0.124	0.523	0.820	0.145	0.397	0.783
BSSTNet	0.118	0.497	0.833	0.119	0.345	0.840
SIR-Diff*	0.113	0.467	0.846	0.106	0.313	0.849
GARD	0.086	0.284	0.913	0.087	0.256	0.907
CoRest	0.071	0.236	0.936	0.075	0.211	0.917

4.2. Image Restoration

In Table 1, *CoRest* achieves the highest DreamSim and MUSIQ among all compared methods at every degraded-view fraction, and the lowest LPIPS on DL3DV and ETH3D at all three fractions. At 50% degraded views, *CoRest* reduces LPIPS from 0.486 to 0.369 on DL3DV and from 0.436 to 0.329 on ETH3D compared with GARD, the strongest LPIPS baseline on these two datasets. At the same fraction, MUSIQ improves over GARD from 56.87 to 71.20 on DL3DV, 58.21 to 70.98 on ScanNet++, and 61.38 to 72.00 on ETH3D. Brightness and mixed-degradation results, obtained with checkpoints trained for their respective degradation settings, are provided in the supplementary material.

Visual comparison. Fig. 3 compares two scenes from each dataset at 50% degraded views using identical crops across methods. *CoRest* recovers distinct shelf boundaries, book spines, and architectural edges, while the sofa example (c) shows clearly defined cushion and armrest contours. Additional visual comparisons and crop-selection details are provided in the supplementary material.

4.3. Downstream Tasks

Depth consistency. We run Depth Anything 3 (DA3-GIANT-1.1) [15] on complete sampled scenes from DL3DV and ETH3D, replacing blurred targets with restored images and preserving the clean context. We measure consistency with predictions from the original clean scene, rather than accuracy against measured depth ground truth. After fitting one positive scale per scene, condition, and method, we measure absolute relative error (AbsRel), normalized RMSE (nRMSE), and δ_1 on the restored targets, averaging scenes equally.

Table 3. **Downstream feature matching at 50% degraded views.** Mean OmniGlue counts on fixed motion-blur target pairs from DL3DV and ETH3D. Matches counts correspondences at the native threshold of 0.001; Conf. counts matches with confidence ≥ 0.01. Both are higher-is-better counts, not verified inliers. Colors indicate first and second place per dataset, excluding clean input. * denotes our SIR-Diff reproduction. Complete results are provided in the supplementary material.

Method	DL3DV		ETH3D	
	Matches↑	Conf.↑	Matches↑	Conf.↑
Clean input (reference)	282.1	227.8	151.0	116.0
Degraded input	59.38	37.66	24.86	9.286
NAFNet	118.6	82.96	43.57	22.29
MoCE-IR	98.82	69.94	29.00	11.43
R2R	101.1	71.36	38.57	23.00
VRT	101.4	69.86	41.29	19.14
BSSTNet	111.7	76.44	38.86	19.14
SIR-Diff*	87.48	62.96	41.71	24.71
GARD	184.9	132.7	94.86	60.29
CoRest	262.2	206.8	131.7	92.43

CoRest improves all three depth metrics over degraded inputs and every restoration baseline on both datasets at all three fractions. At 50% degraded views (Table 2), *CoRest* obtains AbsRel of 0.071 on DL3DV and 0.075 on ETH3D, compared with 0.113 and 0.106 for SIR-Diff. Complete results at all three fractions and evaluation details are provided in the supplementary material.

Feature matching. We evaluate OmniGlue [6] on one fixed pair of degraded targets per scene and fraction, shared across all methods. Outputs are resized to the native ground-truth grid and then downscaled to a maximum long side of 640 pixels. We report mean correspondence counts at the native threshold and at confidence ≥ 0.01, retaining zero-match pairs. Clean images provide a reference and are excluded from the rankings.

Table 3 reports the 50% condition, where *CoRest* leads all baselines in both counts on both datasets. At 50%, *CoRest* yields 262.2 and 131.7 native matches on DL3DV and ETH3D, versus 87.48 and 41.71 for SIR-Diff. Across all fractions, *CoRest* leads all baselines on DL3DV and in confidence-filtered counts on ETH3D; GARD has more native matches on ETH3D at 25%. These counts measure matching support, not ground-truth-verified correspondence accuracy. Complete results at all three fractions and pair-sampling details are provided in the supplementary material.

4.4. Ablation Study

We conduct motion-blur ablations on DL3DV at 50% degraded views, reporting PSNR, SSIM, and LPIPS.

Table 4. **Training and inference ablations on DL3DV at 50% degraded views.** For motion blur, unguided LoRA adapts self-attention alone (Attention-only) or also feed-forward layers (Attention+FFN). *CoRest* adds geometric training to Attention+FFN, without inference-time VGGT; *CoRest* + VGGT activates retrieval in the same checkpoint. Colors indicate first and second place. Complete results are provided in the supplementary material.

Configuration	PSNR↑	SSIM↑	LPIPS↓
Attention-only	18.37	0.514	0.391
Attention+FFN	18.87	0.527	0.373
CoRest + VGGT	19.02	0.531	0.368
CoRest	19.04	0.533	0.369

Table 5. **Number of reference images.** *CoRest* on DL3DV motion blur at 50% degraded views, varying the inference reference count with one checkpoint trained using four references. Colors indicate first and second place.

References	PSNR↑	SSIM↑	LPIPS↓
1	18.58	0.525	0.413
2	18.98	0.534	0.381
3	19.02	0.533	0.373
5	19.01	0.531	0.368
6	18.98	0.529	0.369
4 (ours)	19.04	0.533	0.369

LoRA modules and geometric training. Table 4 compares Attention-only, Attention+FFN, and *CoRest* as three separately trained configurations, all evaluated without VGGT at inference. Attention-only and Attention+FFN are trained without the retrieval branch or its auxiliary loss; *CoRest* uses Attention+FFN LoRA with geometric guidance during training. Extending LoRA to the feed-forward layers reduces LPIPS from 0.391 to 0.373; geometric training further reduces it to 0.369 while improving PSNR and SSIM. These gains persist after the retrieval branch is removed. Results at all three degraded-view fractions are provided in the supplementary material.

Adding VGGT at inference. Table 4 also compares *CoRest* with *CoRest* + VGGT using the same checkpoint and reference-selection policy. At 50% degraded views, activating retrieval changes PSNR by less than 0.022 dB, SSIM by less than 0.0013, and LPIPS by less than 0.0006. Together with the gains over unguided adaptation, these small changes suggest that the model learns to use cross-view information during training without relying on explicit geometric retrieval at inference. Full results on DL3DV, ScanNet++, and ETH3D at all three fractions are provided in the supplementary material.

Reference count. Table 5 varies the number of references at inference using the same checkpoint trained with four references. Moving from one to four references reduces LPIPS from 0.413 to 0.369 and improves PSNR from 18.58 to 19.04 dB. Although the best scores occur at different reference counts, four references capture most of the improvement while keeping the input sequence short.

Table 6. **Reference selection, order, and reuse.** *CoRest* on DL3DV motion blur at 50% degraded views. All rows use the same checkpoint without VGGT. Reuse marks whether restored outputs are eligible as references (✓: yes; ×: no; –: not applicable to wrong-scene references). Colors indicate first and second place. Complete results are provided in the supplementary material.

Reference input	Reuse	PSNR↑	SSIM↑	LPIPS↓
Wrong scene	–	17.08	0.467	0.565
Random selection	✓	18.39	0.507	0.429
Shuffled order	✓	19.01	0.531	0.370
Matching-based selection	×	18.93	0.528	0.367
Matching-based selection (ours)	✓	19.04	0.533	0.369

Reference content, order, and reuse. In Table 6, LPIPS rises from 0.369 with matching-based selection to 0.429 with random selection and 0.565 with references from another scene. Shuffling the selected references changes LPIPS only to 0.370, indicating that reference relevance matters more than their order. All variants use the same trained checkpoint without inference-time VGGT. The no-reuse variant selects references from original inputs, retaining the original degraded versions of already processed views instead of their restored outputs. With matching-based selection, reusing restored outputs raises PSNR from 18.93 to 19.04 dB and SSIM from 0.528 to 0.533, but slightly increases LPIPS from 0.367 to 0.369.

Each policy changes the reference images supplied to VACE throughout sequential restoration, so its effect includes changes to the reference pool for later targets. Complete results at all three fractions and details of the no-reuse policy are provided in the supplementary material.

5. Conclusion

We presented *CoRest*, a reference-conditioned restoration method that recovers degraded images using complementary observations of the same scene. To make this evidence useful across viewpoint changes, we adapt a pretrained video generative model with geometric guidance: during training, VGGT associates target and reference regions, guiding the retrieval of reference appearance features for restoration. The learned LoRA and reference conditioning retain the benefit of this training, enabling inference without VGGT. Restored views can then support subsequent targets through sequential reuse. Experiments demonstrate leading perceptual quality across three datasets and improved depth consistency and feature-matching support. Brightness and mixed-

degradation experiments further show that the framework can adapt to different degradation types. Together, these results demonstrate the value of learning from geometric guidance to restore scene-specific details and produce images useful for downstream vision tasks.

LLM Usage Statement

We used ChatGPT (GPT-6) to assist with experimental code and workflows, organizing and analyzing results, and drafting and revising the manuscript. ChatGPT image-generation tools were also used to assist with figure preparation. The authors reviewed and verified all code, experimental results, citations, and manuscript content, and take full responsibility for the final work.

References

[1] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple Baselines for Image Restoration. In *ECCV*, 2022. 1, 5

[2] Hyungjin Chung, Jeongsol Kim, Michael Thompson McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *ICLR*, 2023. 2

[3] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. In *NeurIPS*, 2023. 5

[4] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole. CAT3D: Create Anything in 3D with Multi-View Diffusion Models. In *NeurIPS*, 2024. 2

[5] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*, 2022. 2, 3, 5

[6] Hanwen Jiang, Arjun Karpur, Bingyi Cao, Qixing Huang, and André Araújo. OmniGlue: Generalizable Feature Matching with Foundation Model Guidance. In *CVPR*, 2024. 4, 7, 1

[7] Yuming Jiang, Kelvin C. K. Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust Reference-Based Super-Resolution via C2-Matching. In *CVPR*, 2021. 2

[8] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-One Video Creation and Editing. In *ICCV*, 2025. 2, 3, 5

[9] Donggoo Jung, Daehyun Kim, Guanghui Wang, and Tae Hyun Kim. Exposure-slot: Exposure-centric Representations Learning with Slot-in-Slot Attention for Region-aware Exposure Correction. In *CVPR*, 2025. 4

[10] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising Diffusion Restoration Models. In *NeurIPS*, 2022. 2

[11] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale Image Quality Transformer. In *ICCV*, 2021. 5

[12] Jin Hyeon Kim, Jaeeun Lee, Claire Kim, Kyoungjin Oh, Paul Hyunbin Cho, Jaewon Min, Yeji Choi, Jihye Park, Hyunhee Park, Minkyu Park, and Seungryong Kim. Geometry-Aware Representation Denoising for Robust Multi-view 3D Reconstruction. *CoRR*, 2026. 2, 5

[13] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent Video Restoration Transformer with Guided Deformable Attention. In *NeurIPS*, 2022. 2

[14] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. VRT: A Video Restoration Transformer. *IEEE TIP*, 2024. 5

[15] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the Visual Space from Any Views. In *ICLR*, 2026. 7, 4

[16] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. DiffBIR: Toward Blind Image Restoration with Generative Diffusion Prior. In *ECCV*, 2024. 1, 2

[17] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A Large-Scale Scene Dataset for Deep Learning-based 3D Vision. In *CVPR*, 2024. 2, 5

[18] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. In *ICLR*, 2024. 2

[19] Hao Luo, Baoliang Chen, Lingyu Zhu, Peilin Chen, and Shiqi Wang. RCNet: Deep Recurrent Collaborative Network for Multi-View Low-Light Image Enhancement. *IEEE TMM*, 2025. 2

[20] Yucheng Mao, Boyang Wang, Nilesh Kulkarni, and Jeong Joon Park. SIR-DIFF: Sparse Image Sets Restoration with Multi-View Diffusion Model. In *CVPR*, 2025. 2, 5

[21] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "Completely Blind" Image Quality Analyzer. *IEEE SPL*, 2013. 5

[22] Manthan Patel, Fan Yang, Yuheng Qiu, Cesar Cadena, Sebastian A. Scherer, Marco Hutter, and Wenshan Wang. TartanGround: A Large-Scale Dataset for Ground Robot Perception and Navigation. In *IROS*, 2025. 5

[23] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Ángel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In *ICCV*, 2021. 5

[24] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A Multi-view Stereo Benchmark with High-Resolution Images and Multi-camera Videos. In *CVPR*, 2017. 2, 5

[25] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view Diffusion for 3D Generation. In *ICLR*, 2024. 2

[26] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent Correspondence from Image Diffusion. In *NeurIPS*, 2023. 2

[27] Yu-Ju Tsai, Yu-Lun Liu, Lu Qi, Kelvin C. K. Chan, and Ming-Hsuan Yang. Dual Associated Encoder for Face Restoration. In *ICLR*, 2024. 1

[28] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Xiaofeng Meng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models. *CoRR*, 2025. 2, 3, 5

[29] Chenxu Wang, Kai Zhang, and Jian Yang. Retrieve-to-Restore: Efficient All-in-One Image Restoration with a Retrieval-Based Degradation Bank. In *CVPR*, 2026. 5, 4

[30] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. VGGT: Visual Geometry Grounded Transformer. In *CVPR*, 2025. 2, 3

[31] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D Vision Made Easy. In *CVPR*, 2024. 2

[32] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 5

[33] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry Forcing: Marrying Video Diffusion and 3D Representation for Consistent World Modeling. In *ICLR*, 2026. 3

[34] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning Texture Transformer Network for Image Super-Resolution. In *CVPR*, 2020. 2

[35] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes. In *ICCV*, 2023. 2, 5

[36] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think. In *ICLR*, 2025. 2

[37] Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity Experts are Task-Discriminative Learners for Any Image Restoration. In *CVPR*, 2025. 5, 4

[38] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient Transformer for High-Resolution Image Restoration. In *CVPR*, 2022. 1

[39] Huicong Zhang, Haozhe Xie, and Hongxun Yao. Blur-Aware Spatio-Temporal Sparse Transformer for Video Deblurring. In *CVPR*, 2024. 5

[40] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 2018. 5

[41] Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. VideoREPA: Learning Physics for Video Generation through Relational Alignment with Foundation Models. In *NeurIPS*, 2025. 2