GeoCR: Learning a Generalist Cloud Removal Prior from Heterogeneous Observations

Anonymous authors
Paper under double-blind review

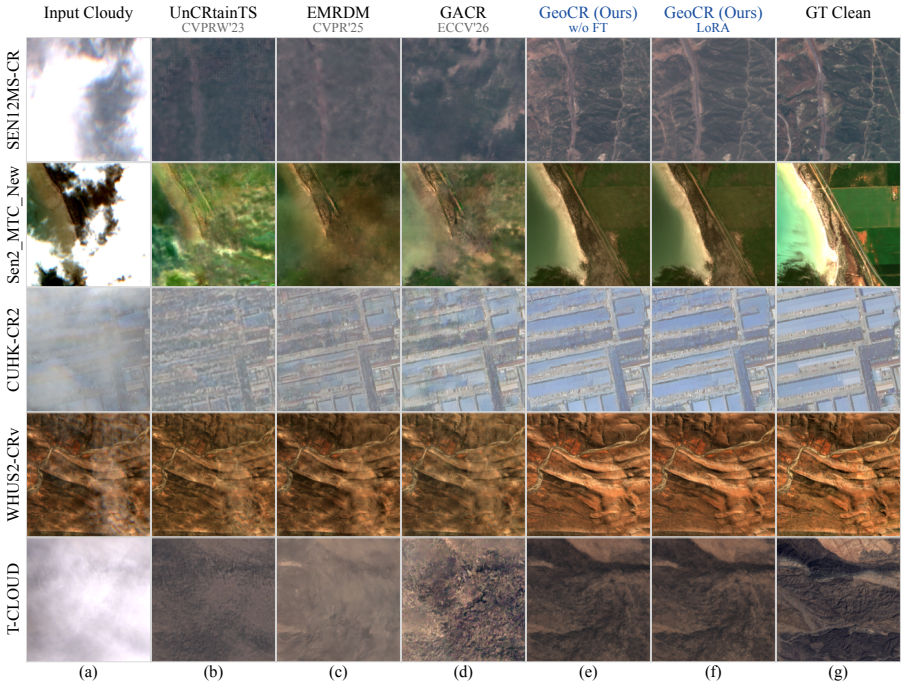

Figure 1: **Qualitative comparison on cloud-removal benchmarks.** (a) Input cloudy, (b) UnCRtainTS, (c) EMRDM, (d) GACR, (e) GeoCR without fine-tuning, (f) GeoCR with LoRA adaptation, and (g) ground-truth cloud-free image. GeoCR recovers coherent scene structures across diverse datasets without dataset-specific fine-tuning and with optional dataset-specific LoRA adaptation.

Abstract

Cloud removal (CR) methods are typically specialized to individual datasets and input configurations, limiting reuse across sensors, spectral bands, and observation settings. We introduce **GeoCR**, a generalist model that *firstly* unifies RGB-only-based CR and multispectral-based CR from single- or multi-temporal cloudy observations, with optional SAR guidance, within a single network. To accommodate different spectral and sensing domains, compact input and output stems extend a pretrained RGB autoencoder while keeping its encoder and decoder trunks frozen. This shared latent interface enables a single flow transformer to jointly model clean RGB and non-RGB latents, conditioned on separate cloudy-observation streams and optional SAR tokens. Through joint pretraining on the training splits of ten datasets comprising 883,331 cloud-free target images, GeoCR learns a *shared cloud removal prior* across these heterogeneous configurations. The same pretrained checkpoint supports direct inference without dataset-specific fine-tuning and efficient adaptation through low-rank adaptation (LoRA). We evaluate GeoCR against general image restoration and cloud removal methods on held-out test splits of the contributing datasets under full-band and RGB-only settings. GeoCR achieves the best FID and DISTS on full-band SEN12MS-CR and Sen2_MTC_New and RGB-only CUHK-CR2, *significantly* outperforming existing models and demonstrating the effectiveness of a reusable generative model across diverse cloud removal settings.

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

Optical satellite imagery provides rich spatial and spectral information for monitoring the Earth's surface. Clouds and haze obscure these observations and interrupt the temporal coverage needed for consistent analysis (Ebel et al., 2023). Cloud removal (CR) reconstructs cloud-free optical imagery using visible content and complementary observations. The available evidence varies substantially: some applications provide a single cloudy RGB image, whereas others offer multispectral time series and synthetic aperture radar (SAR) measurements. A reusable CR model must therefore reconstruct missing content while accommodating different spectral bands, temporal coverage, and sensing modalities.

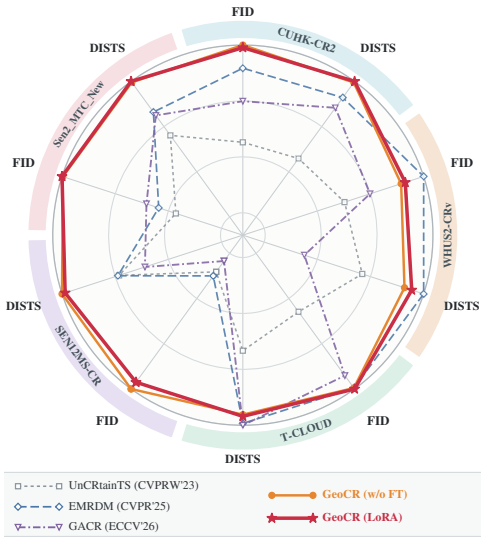

Figure 2: **Cross-dataset comparison of cloud removal (CR) methods.** We report FID and DISTS on five cloud removal benchmarks, independently normalized for each dataset–metric pair as $100 \times$ best/value.

Recent methods have advanced CR through attention-based temporal fusion (Ebel et al., 2023) and diffusion-based restoration (Zou et al., 2024; Liu et al., 2025). However, they are typically trained separately for individual datasets and input configurations, limiting reuse across observation settings. Large datasets such as AllClear (Zhou et al., 2024) provide opportunities for broader training, but combining heterogeneous sources requires a model that accommodates differences in both conditioning observations and output bands. This raises a central question: *Can one model learn a shared CR prior that supports RGB-only-based and multispectral-based reconstructions from single- or multi-temporal observations, with or without SAR?*

To the best of our knowledge, we address *for the first time* this question with **GeoCR**, a generalist model that brings these observation and reconstruction settings into a single generative framework. Here, generalist refers to one pretrained checkpoint supporting multiple datasets and observation configurations. We construct a unified corpus from the training splits of ten CR datasets, comprising 883,331 cloud-free target images. Spanning Sentinel-2, Landsat-8, and high-resolution optical imagery, this corpus supports joint learning across different spectral, temporal, and SAR configurations. GeoCR learns a *shared cloud removal prior* from these sources for direct inference and optional dataset-specific adaptation. This unification requires a common representation for heterogeneous measurements. Although optical bands differ in their spectral responses, they can share spatial structure, including scene layout and object boundaries; SAR provides complementary structural evidence through a different sensing mechanism. These properties motivate reusing pretrained spatial features while adapting the interfaces to each measurement domain. Our *unified embedding framework* connects non-RGB bands and SAR observations to a pretrained RGB autoencoder through compact, configuration-specific input and output stems. The encoder and decoder trunks remain frozen, and the original RGB path is preserved. Training only the stems that require adaptation provides a common latent interface while retaining the pretrained codec.

Building on this interface, we repurpose a pretrained image flow transformer for conditional CR through flow matching (Lipman et al., 2022). The generator jointly models clean RGB and available non-RGB latents using the supplied cloudy observations and optional SAR. RGB and non-RGB latents are combined within each optical observation, while different temporal observations and SAR retain separate conditioning streams. This design lets one generator combine complementary evidence across varying input configurations. After joint pretraining, the same checkpoint supports direct inference without dataset-specific fine-tuning (w/o FT) and parameter-efficient adaptation through LoRA (Hu et al., 2021).

We evaluate GeoCR against general image restoration and CR methods on held-out test splits of the contributing datasets, covering five primary benchmarks under full-band and RGB-only settings.

Figure 2 summarizes the comparison, with GeoCR (w/o FT) using the same checkpoint across all five benchmarks. GeoCR achieves the lowest FID and DISTS among the compared methods on full-band SEN12MS-CR and Sen2_MTC_New and RGB-only CUHK-CR2. These results demonstrate that a shared pretrained model can provide competitive distributional and perceptual quality across diverse CR configurations without requiring dataset-specific optimization for every setting.

Our contributions are summarized as follows:

- We introduce **GeoCR**, a generalist model that unifies *firstly* RGB-only-based CR and multispectral-based CR across single- and multi-temporal observations with optional SAR guidance.
- We develop a unified embedding and conditioning framework that extends frozen RGB encoder–decoder trunks through compact stems and combines observation-specific conditioning streams within a single flow transformer.
- We consolidate ten CR datasets comprising 883,331 cloud-free target images for joint pretraining. Extensive evaluations under full-band and RGB-only settings demonstrate state-of-the-art FID and DISTS on three of five primary benchmarks, *significantly* outperforming existing CR models.

2 Related Work

General image translation and restoration. Image translation and restoration provide complementary approaches to recovering images from degraded observations. Pix2pix (Isola et al., 2017) learns paired mappings through conditional adversarial training, while pix2pixHD (Wang et al., 2018) extends it to high-resolution synthesis with coarse-to-fine generation and multiscale discrimination. BBDM (Li et al., 2023) formulates image translation as a Brownian bridge between source and target domains. HI-Diff (Chen et al., 2023) generates compact latent priors with diffusion and integrates them hierarchically into a regression-based deblurring network. These methods provide general-purpose baselines for evaluating adversarial, bridge-based, and diffusion-assisted approaches to cloud removal.

Cloud removal (CR). CR methods exploit temporal redundancy, complementary sensor observations, and generative priors. UnCRtainTS (Ebel et al., 2023) aggregates optical and SAR observations through temporal attention while estimating reconstruction uncertainty. DiffCR (Zou et al., 2024) develops an efficient conditional diffusion framework for CR. IDF-CR (Wang et al., 2024) combines pixel-space restoration with latent diffusion refinement. EMRDM (Liu et al., 2025) establishes mean-reverting dynamics between cloudy and clear images for single- and multi-temporal restoration. ThiefCloud (Zhao et al., 2025) learns a cloud-thickness prior for thin-cloud removal, while GACR (Wang et al., 2026) combines observation-anchored residual flow with alignment to pretrained visual representations. Related efforts explore unified restoration, multisensor tokenization, and spatiotemporal modeling in remote sensing (Cui & Liu, 2026; Lehmann et al., 2026; Zhang et al., 2025). Our GeoCR focuses on jointly learning a reusable CR prior across heterogeneous datasets and observation configurations. Its shared latent interface and observation-specific conditioning streams support RGB-only-based and multispectral-based reconstructions with varying temporal coverage and SAR availability, using one pretrained checkpoint for direct inference and optional LoRA adaptation.

Cloud removal datasets. CR datasets cover diverse sensing conditions and reconstruction settings. RICE (Lin et al., 2019) provides paired imagery for thin- and thick-cloud removal, while CUHK-CR (Sui et al., 2024) offers high-resolution RGB–NIR observations. SEN12MS-CR (Ebel et al., 2020) pairs cloudy and clear multispectral optical imagery with SAR measurements. Sen2_MTC_Old (Sarukkai et al., 2020) and Sen2_MTC_New (Huang & Wu, 2022) provide multiple cloudy observations for temporal reconstruction. AllClear (Zhou et al., 2024) expands geographic coverage and multisensor time series, demonstrating the benefits of larger training sets and complementary observations. These resources differ in spectral coverage, radiometric processing, spatial resolution, temporal sampling, and SAR availability. Our GeoCR accommodates these differences through a common representation and conditioning framework, enabling joint pretraining across heterogeneous CR datasets.

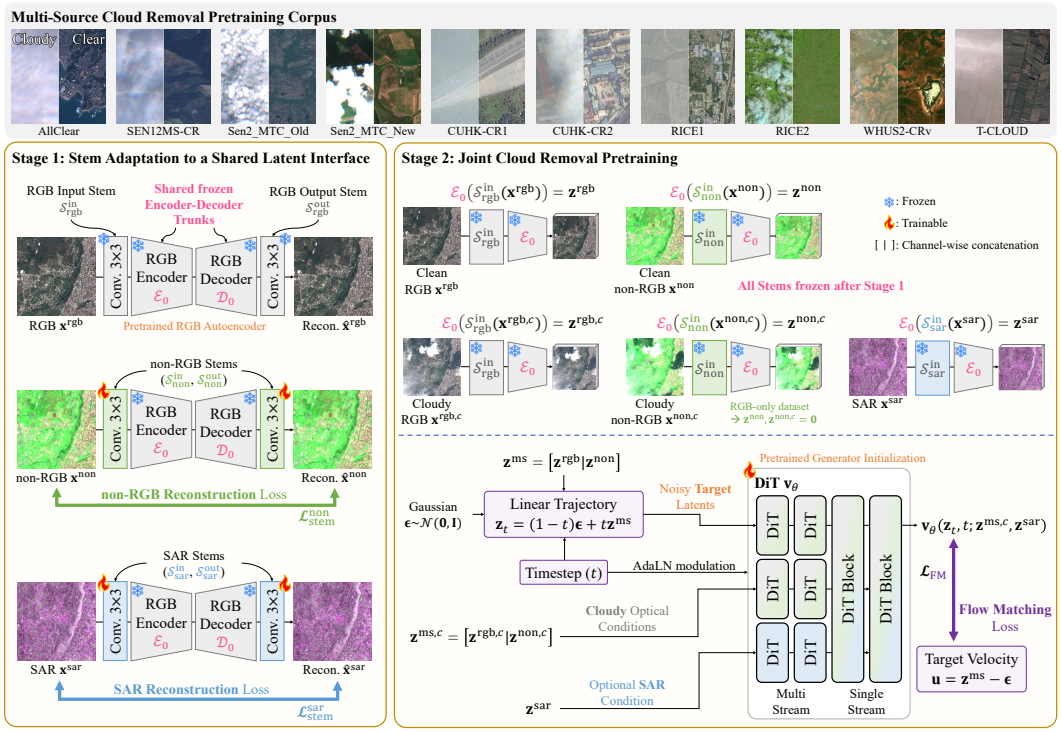

Figure 3: **GeoCR framework.** A shared latent interface enables joint cloud removal across heterogeneous spectral, temporal, and SAR configurations.

3 GeoCR

GeoCR learns a shared CR prior through two pretraining stages followed by direct inference or dataset-specific adaptation. Figure 3 illustrates the framework. Stage 1 trains compact input and output stems for ten-band non-RGB and SAR observations while keeping the shared pretrained encoder–decoder trunks and original RGB path frozen. Stage 2 learns a shared cloud removal prior through conditional flow matching on the heterogeneous corpus, jointly generating clean RGB and non-RGB latents. Each cloudy optical observation and the optional SAR input retain separate conditioning streams. Finally, the same pretrained checkpoint is used directly (without fine-tuning) or adapted through LoRA (Hu et al., 2021).

Problem setting. Each training sample contains a cloud-free optical target $\mathbf{x} = (\mathbf{x}^{\mathrm{rgb}}, \mathbf{x}^{\mathrm{non}})$ and K cloudy optical observations $\{(\mathbf{x}_k^{\mathrm{rgb,c}}, \mathbf{x}_k^{\mathrm{non,c}})\}_{k=1}^{K}$, optionally accompanied by SAR imagery $\mathbf{x}^{\mathrm{sar}}$. Here, the superscripts 'c' and 'non', respectively, denote cloudy observations and non-RGB spectral bands that may be absent from the target or conditioning observations. We use d to index the spectral and radiometric configuration that selects the corresponding non-RGB stem. For ten-band non-RGB inputs, we distinguish top-of-atmosphere (TOA) reflectance and bottom-of-atmosphere (BOA) imagery.

3.1 Stage 1: Stem Adaptation to a Shared Latent Interface

Shared pretrained encoder and decoder. As shown in Figure 3, different optical bands can share scene layout and spatial boundaries despite differences in their spectral responses. This motivates sharing pretrained spatial features while adapting the interfaces to different measurement domains. Let $\mathcal{E}_0$ and $\mathcal{D}_0$ denote the shared encoder and decoder trunks of a pretrained RGB autoencoder (Labs, 2025), excluding their input and output projections. For a representation m, we define

$$\mathcal{E}_m = \mathcal{E}_0 \circ \mathcal{S}_m^{\mathrm{in}}, \qquad \mathcal{D}_m = \mathcal{S}_m^{\mathrm{out}} \circ \mathcal{D}_0, \tag{1}$$

where $m \in \{\mathrm{rgb}, (\mathrm{non}, d), \mathrm{sar}\}$. The input and output stems, $\mathcal{S}_m^{\mathrm{in}}$ and $\mathcal{S}_m^{\mathrm{out}}$, each of which consists of a single convolutional layer. The RGB stems are the inherited pretrained projections; additional stems are initialized from these projections. All routes share the same frozen trunks, $\mathcal{E}_0$ and $\mathcal{D}_0$.

Reconstruction-based stem adaptation. For input $\mathbf{x}^m$, the reconstruction and stem objective are defined as

$$\widehat{\mathbf{x}}^m = \mathcal{D}_m(\mathcal{E}_m(\mathbf{x}^m)), \qquad \mathcal{L}_{\mathrm{stem}}^m = \mathbb{E}_{\mathbf{x}^m}\left[\|\widehat{\mathbf{x}}^m - \mathbf{x}^m\|_1\right], \tag{2}$$

where the ℓ_1 reconstruction error is measured in the preprocessed input domain. Only the input and output stems selected for adaptation are optimized while $\mathcal{E}_0$ and $\mathcal{D}_0$ remain fixed. We train stems for the ten-band non-RGB and dual-polarization SAR configurations, while retaining the original RGB path without stem training. Each route reconstructs its own observation, so this stage learns a representation interface without requiring a cloudy-to-clear mapping or an RGB reconstruction target for non-RGB measurements. Fixing the trunks constrains adaptation to their pretrained feature interfaces while allowing the stems to accommodate different channel layouts and measurement statistics. All stems and trunks remain frozen after Stage 1.

3.2 Stage 2: Joint Cloud Removal Pretraining

Optical and SAR latents. We encode cloud-free targets and cloudy conditions using the corresponding frozen routes having the following latents in a shared RGB latent space:

$$\begin{aligned} \mathbf{z}^{\mathrm{rgb}} &= \mathcal{E}_{\mathrm{rgb}}(\mathbf{x}^{\mathrm{rgb}}), & \mathbf{z}^{\mathrm{non}} &= \mathcal{E}_{\mathrm{non},d}(\mathbf{x}^{\mathrm{non}}), \\ \mathbf{z}_k^{\mathrm{rgb,c}} &= \mathcal{E}_{\mathrm{rgb}}(\mathbf{x}_k^{\mathrm{rgb,c}}), & \mathbf{z}_k^{\mathrm{non,c}} &= \mathcal{E}_{\mathrm{non},d}(\mathbf{x}_k^{\mathrm{non,c}}). \end{aligned} \tag{3}$$

Each available optical component (rgb or non) is mapped to a latent with the same channel dimension and spatial grid. We concatenate the RGB and non-RGB latents along the channel dimension to form joint multispectral (MS) representations of the target and each cloudy observation as:

$$\mathbf{z}^{\mathrm{ms}} = [\mathbf{z}^{\mathrm{rgb}} \mid \mathbf{z}^{\mathrm{non}}], \qquad \mathbf{z}_k^{\mathrm{ms,c}} = [\mathbf{z}_k^{\mathrm{rgb,c}} \mid \mathbf{z}_k^{\mathrm{non,c}}], \tag{4}$$

where $[\cdot \mid \cdot]$ denotes channel-wise concatenation and 'ms' stands for multispectral, combining RGB and non-RGB information. An unavailable non-RGB component is replaced by a zero latent of the corresponding shape. This retains a common input layout for RGB-only and multispectral samples. When available, SAR is encoded as $\mathbf{z}^{\mathrm{sar}} = \mathcal{E}_{\mathrm{sar}}(\mathbf{x}^{\mathrm{sar}})$. We denote the complete set of available optical and SAR conditions by $\mathcal{C}$.

Repurposing a pretrained generator. We initialize GeoCR from a pretrained image flow transformer (Labs, 2025) and condition it on spatial observations instead of language. The noisy target, cloudy optical conditions, and optional SAR condition are projected into spatial tokens. Each of the K cloudy observations forms a separate conditioning stream, yielding $1 + K$ streams including the target, with one additional stream when SAR is available. Multi-stream blocks use separate parameter sets for target, optical-conditioning, and SAR tokens, with parameters shared across the K optical-conditioning streams. All streams interact through joint attention. The subsequent single-stream blocks jointly process all tokens with shared parameters. Only target tokens are passed to the final velocity head; the optical and SAR tokens provide conditioning information.

Conditional latent flow matching. For each target latent $\mathbf{z}^{\mathrm{ms}}$, we sample an independent Gaussian source $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a time $t \sim \mathcal{U}(0, 1)$. The linear path $\mathbf{z}_t$ and target velocity $\mathbf{u}$ are

$$\mathbf{z}_t = (1 - t)\boldsymbol{\epsilon} + t\mathbf{z}^{\mathrm{ms}}, \qquad \mathbf{u} = \frac{\partial \mathbf{z}_t}{\partial t} = \mathbf{z}^{\mathrm{ms}} - \boldsymbol{\epsilon}. \tag{5}$$

The generator $\mathbf{v}_\theta$ predicts this velocity conditioned on the available observations:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left[\|\mathbf{v}_\theta(\mathbf{z}_t, t; \mathcal{C}) - \mathbf{u}\|_2^2\right]. \tag{6}$$

Only the generator is optimized in this stage. The expectation spans the heterogeneous training corpus and its available observation configurations, so a common velocity model $\mathbf{v}_\theta$ learns RGB and multispectral restoration from single- and multi-temporal inputs with or without SAR guidance.

Table 1: **Overview of the GeoCR pretraining corpus.** Counts refer to cloud-free target images in the training splits. Spectral bands describe the targets unless otherwise noted; cloudy conditions indicate the number of optical input frames. Tile sizes are in pixels. L1C: Level-1C; TOA/BOA: top-/bottom-of-atmosphere reflectance; RGB: red, green, and blue; NIR: near-infrared; SAR: synthetic aperture radar; GSD: ground sampling distance. B8 and B10 denote Sentinel-2 band identifiers.

Source	Imagery source	Spectral bands	Delivered tile	GSD (m)	Cloudy cond.	SAR cond.	Clean targets
AllClear (Zhou et al., 2024)	Sentinel-2 + Sentinel-1	13 (L1C, TOA)	256^2	10	1–3	✓	662,022
SEN12MS-CR (Ebel et al., 2020)	Sentinel-2 + Sentinel-1	13 (L1C, TOA)	256^2	10	1	✓	107,143
WHUS2-CRv (Li et al., 2022)	Sentinel-2	13 (BOA; B10: L1C, TOA)	$384^2/192^2/64^2$	10	1	✗	18,816
Sen2_MTC_Old (Sarukkai et al., 2020)	Sentinel-2	3 (RGB; +NIR for cloudy inputs)	256^2	10	3	✗	88,874
Sen2_MTC_New (Huang & Wu, 2022)	Sentinel-2	4 (RGB+B8, BOA)	256^2	10	3	✗	2,380
T-CLOUD (Ding et al., 2022)	Landsat-8	3 (RGB, 8-bit)	256^2	30	1	✗	2,234
RICE2 (Lin et al., 2019)	Landsat-8	3 (RGB, 8-bit)	512^2	30	1	✗	553
CUHK-CR1 (Sui et al., 2024)	Jilin-1 KF01B	4 (RGB+NIR, 8-bit)	512^2	0.5	1	✗	508
CUHK-CR2 (Sui et al., 2024)	Jilin-1 KF01B	4 (RGB+NIR, 8-bit)	512^2	0.5	1	✗	426
RICE1 (Lin et al., 2019)	Google Earth	3 (RGB, 8-bit)	512^2	–	1	✗	375

Table 2: **Quantitative comparison on CUHK-CR2.** Its RGB evaluation setting supports both RGB-only image translation/restoration methods and specialized cloud removal methods, enabling comparison across all ten baselines. **Bold** and underlined values indicate the best and second-best results, respectively. Additional results are provided in the *Appendix*.

Method	Venue	FID↓	DISTS↓	KID↓	DINO↑	LPIPS↓	SSIM↑	PSNR↑
General image translation/restoration methods								
pix2pix (Isola et al., 2017)	CVPR'17	194.8	0.241	0.1201	0.538	0.319	0.452	19.46
pix2pixHD (Wang et al., 2018)	CVPR'18	257.1	0.262	0.2130	0.379	0.304	0.401	19.34
BBDM (Li et al., 2023)	CVPR'23	514.6	0.542	0.5827	0.131	0.737	0.326	17.68
HI-Diff (Chen et al., 2023)	NeurIPS'23	165.7	0.192	0.0769	0.664	0.253	0.638	23.54
Cloud removal methods								
UnCRtainTS (Ebel et al., 2023)	CVPRW'23	165.6	0.265	0.0761	0.529	0.345	0.586	22.12
DiffCR (Zou et al., 2024)	TGRS'24	245.2	0.285	0.1874	0.615	0.353	0.582	22.85
IDF-CR (Wang et al., 2024)	TGRS'24	167.0	0.214	0.0873	0.645	0.270	<u>0.641</u>	23.18
ThiefCloud (Zhao et al., 2025)	TCSVT'25	136.3	0.226	0.0593	0.668	0.232	0.638	**23.88**
EMRDM (Liu et al., 2025)	CVPR'25	104.4	0.167	0.0259	0.727	0.208	**0.654**	<u>23.61</u>
GACR (Wang et al., 2026)	ECCV'26	125.0	0.178	0.0496	0.739	0.217	0.620	23.45
Ours								
GeoCR (w/o FT)	–	**93.6**	<u>0.153</u>	<u>0.0200</u>	**0.784**	**0.194**	0.607	23.21
GeoCR (LoRA)	–	<u>94.7</u>	**0.152**	**0.0194**	<u>0.779</u>	<u>0.196</u>	0.601	23.11

3.3 Inference and Adaptation

For each dataset, we use the shared Stage 2 generator while keeping all autoencoder trunks and stems frozen. We consider two modes: (i) direct inference without dataset-specific optimization, denoted as **GeoCR (w/o FT)**, and (ii) low-rank adaptation (LoRA) (Hu et al., 2021), denoted as **GeoCR (LoRA)**. Only the adapters are optimized on the target dataset's training split using the Stage 2 flow-matching objective. In both modes, inference integrates the learned conditional flow from Gaussian noise and decodes the resulting RGB and, when required, non-RGB latents through the corresponding frozen decoder routes.

4 Experiments

4.1 Pretraining and Evaluation Data

Multi-source pretraining corpus. We construct a unified corpus from the official training splits of ten cloud removal datasets, comprising 883,331 cloud-free optical targets. Table 1 summarizes their sensing platforms, spectral bands, radiometric representations, and observation availability. The corpus combines Sentinel-2, Landsat-8, and high-resolution imagery, including RGB and multispectral targets, single- and multi-temporal cloudy inputs, and optional SAR guidance. These configurations jointly contribute to the shared GeoCR prior through the common latent interface.

Evaluation benchmarks. We build on the released training and test partitions, and apply the split corrections and duplicate exclusions documented in *Appendix*. Evaluation uses the resulting held-

Table 3: **Quantitative comparison on Sen2_MTC_New.** (a) Full-band setting. (b) RGB setting.

(a) Full-band (RGB + NIR)

Method	FID↓	DISTS↓	KID↓	DINO↑	LPIPS↓	SSIM↑	PSNR↑
UnCRtainTS	113.4	0.258	0.0597	0.471	0.428	0.588	17.15
DiffCR	95.9	0.297	0.0388	0.499	0.312	0.599	19.26
ThiefCloud	143.6	0.333	0.0767	0.310	0.508	0.481	14.92
EMRDM	96.6	0.218	0.0459	0.533	0.314	0.640	18.20
GACR	87.4	0.223	0.0358	0.535	0.320	0.626	18.90
GeoCR (w/o FT)	52.9	0.182	0.0032	0.642	0.221	**0.661**	**20.77**
GeoCR (LoRA)	**52.7**	**0.181**	**0.0031**	**0.643**	**0.220**	0.660	20.76

(b) RGB-only

Method	FID↓	DISTS↓	KID↓	DINO↑	LPIPS↓	SSIM↑	PSNR↑
pix2pix	161.6	0.305	0.1188	0.224	0.480	0.541	20.16
pix2pixHD	153.9	0.292	0.1121	0.274	0.365	0.646	24.00
BBDM	179.4	0.392	0.1243	0.243	0.504	0.652	24.48
HI-Diff	125.8	0.302	0.0695	0.390	0.398	**0.742**	**25.66**
GeoCR (w/o FT)	79.1	**0.252**	**0.0223**	0.489	**0.313**	0.706	25.41
GeoCR (LoRA)	**78.4**	0.265	0.0313	**0.520**	0.333	0.712	25.07

Table 4: **Quantitative comparison on SEN12MS-CR.** (a) Full-band setting. (b) RGB setting.

(a) Full-band (13 spectral bands)

Method	FID↓	DISTS↓	KID↓	DINO↑	LPIPS↓	SSIM↑	PSNR↑
UnCRtainTS	80.5	0.276	0.0497	0.440	0.313	**0.884**	28.77
EMRDM	75.7	0.274	0.0475	0.467	0.304	0.873	28.70
GACR	96.9	0.331	0.0679	0.358	0.333	0.828	27.76
GeoCR (w/o FT)	**28.4**	**0.202**	**0.0091**	**0.571**	**0.215**	0.866	**29.29**
GeoCR (LoRA)	29.5	0.205	0.0098	0.569	0.216	0.866	29.19

(b) RGB-only

Method	FID↓	DISTS↓	KID↓	DINO↑	LPIPS↓	SSIM↑	PSNR↑
pix2pix	145.9	0.317	0.1000	0.281	0.465	0.632	21.64
pix2pixHD	70.8	**0.239**	0.0443	0.420	**0.277**	0.772	27.23
BBDM	96.1	0.284	0.0536	0.331	0.367	0.710	25.44
HI-Diff	62.7	0.284	0.0342	0.438	0.320	**0.837**	**28.28**
GeoCR (w/o FT)	**43.9**	0.244	**0.0246**	**0.532**	0.326	0.712	22.15
GeoCR (LoRA)	52.1	0.263	0.0375	0.529	0.308	0.781	24.59

out test samples. Our five primary benchmarks include SEN12MS-CR, Sen2_MTC_New, CUHK-CR2, WHUS2-CRv, and T-CLOUD. Full-band experiments retain the dataset-specific optical band configuration, while RGB-only experiments use RGB optical inputs and targets. Within each setting, all methods are evaluated on identical test samples and spatial extents. GeoCR (w/o FT) uses one jointly pretrained checkpoint across datasets; GeoCR (LoRA) adapts it using only the corresponding training split. Additional benchmark results, test-set sizes, and detailed input configurations are provided in the *Appendix*.

4.2 Implementation Details and Evaluation Protocol

Implementation details. All models are trained on 256×256 crops and NVIDIA B200 GPUs. Stage 1 trains 51.1K stem parameters around the frozen FLUX.2 (Labs, 2025) VAE for 400K updates (global batch size 64, two GPUs). We use AdamW with a constant learning rate of 10^{-4}, without warmup or gradient clipping. Stage 2 optimizes all 3.853B parameters of FLUX.2 4B for 500K updates (global batch size 64, four GPUs) in bf16, using AdamW with zero weight decay, a 2K-update linear warmup to 10^{-4}, and cosine decay to 10^{-6}. All conditions are replaced with a learned null token with probability 0.1; SAR is independently dropped with probability 0.3, and non-RGB conditioning is dropped with probability 0.05. Additionally, 5% of AllClear batches use SAR-only conditioning. Stages 1 and 2 take 47.0 and 195.3 hours, respectively. Downstream LoRA starts from the Stage 2 checkpoint and uses 10K updates with global batch size 32 on one GPU. With a learning rate of 10^{-4} and $r_{\mathrm{L}} = \alpha_{\mathrm{L}} = 16$, LoRA adapts all 80 transformer linear layers, updating 23.1M parameters (0.60%). For our GeoCR, inference uses only four Euler steps on a uniform time grid without classifier-free guidance and a fixed per-sample noise seed.

Baselines. We compare with ten methods spanning general image restoration and specialized cloud removal. General-purpose baselines comprise pix2pix (Isola et al., 2017), pix2pixHD (Wang et al., 2018), BBDM (Li et al., 2023), and HI-Diff (Chen et al., 2023). Cloud removal baselines include UnCRtainTS (Ebel et al., 2023), DiffCR (Zou et al., 2024), IDF-CR (Wang et al., 2024), ThiefCloud (Zhao et al., 2025), EMRDM (Liu et al., 2025), and GACR (Wang et al., 2026). Baselines are retrained separately on each dataset's training split using official implementations when available and assessed under a common evaluation pipeline. The reported baseline set follows the input compatibility of each setting; CUHK-CR2's RGB setting supports all ten methods. GeoCR's shared parent is pretrained on the pooled training splits.

Metrics. We use FID (Heusel et al., 2017) and DISTS (Ding et al., 2020) as the primary measures of distributional similarity and paired perceptual fidelity, respectively. We additionally report KID (Bińkowski et al., 2018), DINO feature similarity using a frozen DINOv3-SAT (Siméoni et al., 2025), LPIPS (Zhang et al., 2018), SSIM (Wang et al., 2004), and PSNR. Together, these metrics assess generated-image distributions, perceptual and feature-level correspondence, and reconstruction fidelity to the paired cloud-free reference. All feature-based metrics are computed on RGB views.

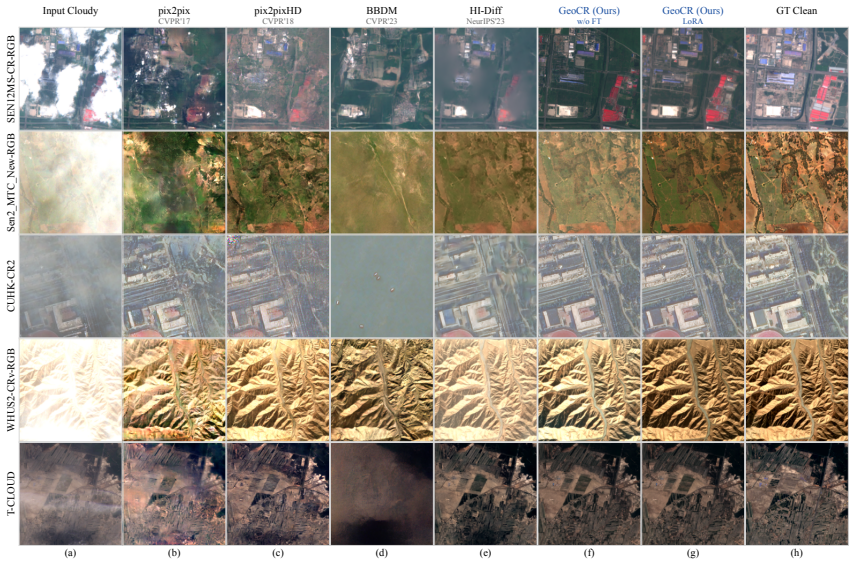

Figure 4: **Qualitative comparison with general restoration methods under RGB-only settings.**

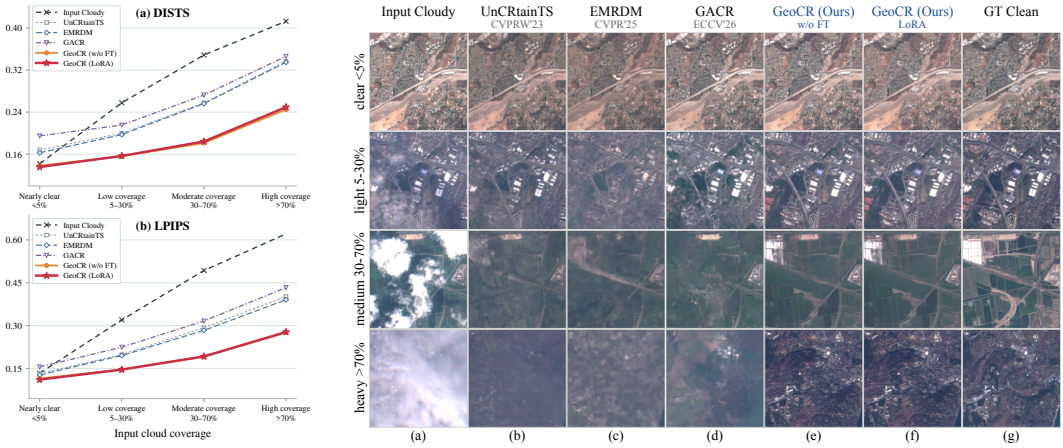

Figure 5: **Cloud removal across input cloud coverage on SEN12MS-CR.** Left: DISTS and LPIPS grouped by input cloud coverage estimated with `s2cloudless`; lower is better. Input Cloudy denotes the unrestored observation. Right: qualitative comparisons across the same coverage groups.

4.3 Comparison with the State of the Art

Quantitative comparison. Tables 2, 3, and 4 show that GeoCR achieves the best FID and DISTS on full-band Sen2_MTC_New and SEN12MS-CR and RGB-only CUHK-CR2. On full-band SEN12MS-CR, GeoCR (w/o FT) reduces FID by over 60% relative to the strongest competing baseline; GeoCR (LoRA) provides modest improvements on selected datasets. Both GeoCR (w/o FT) and GeoCR (LoRA) also lead FID on RGB-only Sen2_MTC_New and SEN12MS-CR, although pix2pixHD retains lower DISTS and LPIPS on the latter. We treat PSNR and SSIM as complementary measures: temporal appearance differences and cloud occlusion can make exact reference matching ambiguous, while minimizing squared pixel error under uncertainty can favor oversmoothed predictions (Blau & Michaeli, 2018). We therefore emphasize FID and DISTS for distributional quality and paired perceptual fidelity.

Qualitative comparison. Figures 1 and 4 show clearer terrain, field boundaries, and urban structures with less residual cloud contamination. These improvements are already evident in GeoCR (w/o FT), supporting the shared prior's ability to restore local structure across diverse observation settings. Additional examples are provided in the *Appendix* and *Supplementary Material*.

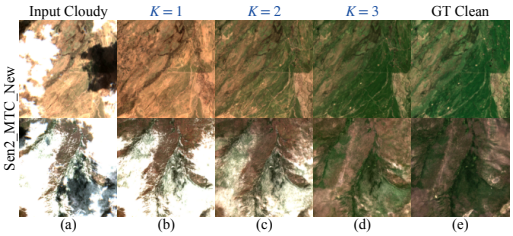

Figure 6: **Temporal conditioning on Sen2_MTC_New.** The same GeoCR (w/o FT) checkpoint uses K cloudy optical frames.

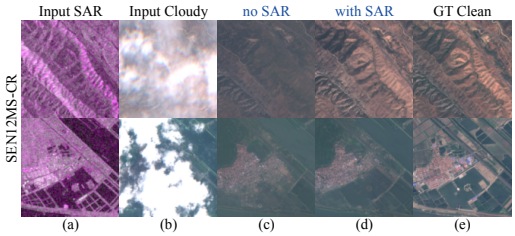

Figure 7: **SAR conditioning on SEN12MS-CR.** GeoCR (w/o FT) uses identical cloudy optical inputs with or without SAR guidance.

4.4 Ablation Studies and Analysis

Performance across cloud coverage. Figure 5 shows lower DISTS and LPIPS for both GeoCR variants than the compared baselines across SEN12MS-CR coverage groups defined solely from cloudy inputs. In nearly clear scenes, GeoCR improves on the unrestored input, while all compared baselines worsen DISTS. Its advantage persists under heavy cloud cover, indicating robust perceptual restoration across the evaluated coverage levels.

Effect of stem training. With the pretrained trunks frozen, stem training improves reconstruction PSNR by 21.24 dB for ten-band non-RGB inputs and 10.74 dB for SAR (Table 5). Gains across clear and cloudy observations and both SAR polarizations support adapting lightweight interfaces to different measurement domains. The RGB path already provides high reconstruction fidelity without adaptation. Similarly, the initialized single-band NIR stem achieves 42.06 dB on Sen2_MTC_New, 40.41 dB on CUHK-CR, and 49.51 dB on Sen2_MTC_Old without stem training, supporting its reuse without further optimization.

Table 5: **Effect of Stage 1 stem training.** Reconstruction PSNR (dB) on AllClear validation samples before and after stem training, with the pretrained encoder and decoder frozen. VV and VH denote SAR polarizations.

Recon. PSNR	**RGB** (w/o training)			**Non-RGB** (stem training)			**SAR** (stem training)		
	All	**Clear**	**Cloudy**	**All**	**Clear**	**Cloudy**	**All**	**VV**	**VH**
Before	45.78	46.77	45.11	18.84	20.60	18.20	18.40	18.57	18.24
After	45.78	46.77	45.11	40.08	41.41	38.34	29.14	28.54	30.21
Gain	+0.00	+0.00	+0.00	**+21.24**	**+20.81**	**+20.14**	**+10.74**	**+9.97**	**+11.98**

Effect of temporal and SAR conditioning. Tables 6 and 7 vary the available observations using the same GeoCR (w/o FT) checkpoint on fixed test subsets. Additional cloudy frames and SAR guidance improve reference fidelity, accompanied by reduced residual clouds and clearer terrain ridges, respectively (Figures 6 and 7). These results support using complementary evidence within one model without retraining for each input configuration.

Table 6: **Temporal conditioning on Sen2_MTC_New.** GeoCR uses the shared checkpoint without fine-tuning.

Method	**Cloudy**	**PSNR↑**	**SSIM↑**
GeoCR	1	16.390	0.4595
GeoCR	2	19.319	0.5963
GeoCR	3	**20.522**	**0.6568**

Table 7: **SAR conditioning on SEN12MS-CR.** GeoCR uses the shared checkpoint without fine-tuning.

Method	**SAR**	**PSNR↑**	**SSIM↑**
GeoCR	✗	28.596	0.8572
GeoCR	✓	**29.442**	**0.8703**

5 Conclusion

We presented GeoCR, a generalist model that learns a shared cloud removal prior through joint pretraining on ten datasets comprising 883,331 cloud-free target images. Its shared latent interface enables a single generator to accommodate diverse spectral bands, temporal observations, and optional SAR guidance. GeoCR achieves strong distributional and perceptual quality without dataset-specific fine-tuning and supports lightweight LoRA adaptation, by *significantly* outperforming existing CR methods. Conditioning ablations show that the same model benefits from additional temporal and SAR evidence without retraining. These results demonstrate the feasibility of learning a reusable cloud removal model across the evaluated heterogeneous observation settings.

AI USE STATEMENT

We used generative AI tools to assist with translation, English-language editing, manuscript organization, and review of experimental reporting and interpretation. The authors reviewed and revised the AI-assisted content, checking technical descriptions against the implementation and empirical statements against the reported results. The authors take full responsibility for the final submission, including its text, methodology, results, claims, and references.

ETHICS STATEMENT

This study uses existing cloud removal datasets containing optical imagery and, where available, SAR observations, subject to their original access, usage, and redistribution conditions. GeoCR reconstructs cloud-free imagery from incomplete observations; generated structures, colors, and spectral values may differ from the actual surface and should not be treated as verified measurements. Geospatial imagery also raises concerns about privacy, data governance, and misuse, including unauthorized monitoring. Performance on the evaluated datasets does not establish reliability across unseen regions, sensors, or acquisition conditions. Consequential applications, such as disaster assessment, require independent validation and corroboration with acquired observations.

REPRODUCIBILITY STATEMENT

Section 3 describes the GeoCR framework and learning objectives, and Section 4 reports the training settings, baselines, and evaluation protocol. Appendices A and B provide stem reconstruction diagnostics and architectural details. Appendices C and D present additional results, while Appendix E discusses failure cases and limitations. Dataset preprocessing and split checks are documented in Appendix F, and metric definitions and evaluation conventions in Appendix G. Code, pretrained checkpoints, and split-audit results will be released upon acceptance, subject to applicable licenses.

REFERENCES

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237. IEEE, 2018.

Zheng Chen, Yulun Zhang, Ding Liu, Jinjin Gu, Linghe Kong, Xin Yuan, et al. Hierarchical integration diffusion model for realistic image deblurring. *Advances in neural information processing systems*, 36:29114–29125, 2023.

Yongchuan Cui and Peng Liu. A unified foundation model for all-in-one multi-modal remote sensing image restoration and fusion with language prompting. *arXiv preprint arXiv:2604.05629*, 2026.

Haidong Ding, Yue Zi, and Fengying Xie. Uncertainty-based thin cloud removal network via conditional variational autoencoders. In *Asian Conference on Computer Vision*, pp. 52–68. Springer, 2022.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

Patrick Ebel, Andrea Meraner, Michael Schmitt, and Xiao Xiang Zhu. Multisensor data fusion for cloud removal in global and all-season sentinel-2 imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):5866–5878, 2020.

Patrick Ebel, Vivien Sainte Fare Garnot, Michael Schmitt, Jan Dirk Wegner, and Xiao Xiang Zhu. Uncrtaints: Uncertainty quantification for cloud removal in optical satellite time series. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2086–2096. IEEE, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Gi-Luen Huang and Pei-Yuan Wu. Ctgan: Cloud transformer generative adversarial network. In *2022 IEEE international conference on image processing (ICIP)*, pp. 511–515. IEEE, 2022.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5967–5976. Ieee, 2017.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

Nils Lehmann, Yi Wang, Zhitong Xiong, and Xiaoxiang Zhu. Eo-vae: Towards a multi-sensor tokenizer for earth observation data. *arXiv preprint arXiv:2602.12177*, 2026.

Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1952–1961. IEEE, 2023.

Jun Li, Yuejie Zhang, Qinghong Sheng, Zhaocong Wu, Bo Wang, Zhongwen Hu, Guanting Shen, Michael Schmitt, and Matthieu Molinier. Thin cloud removal fusing full spectral and spatial features for sentinel-2 imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:8759–8775, 2022.

Daoyu Lin, Guangluan Xu, Xiaoke Wang, Yang Wang, Xian Sun, and Kun Fu. A remote sensing image dataset for cloud removal. *arXiv preprint arXiv:1901.00600*, 2019.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Yi Liu, Wengen Li, Jihong Guan, Shuigeng Zhou, and Yichao Zhang. Effective cloud removal for remote sensing images by an improved mean-reverting denoising model with elucidated design space. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17851–17861. IEEE, 2025.

Vishnu Sarukkai, Anirudh Jain, Burak Uzkent, and Stefano Ermon. Cloud removal in satellite images using spatiotemporal generative networks. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1785–1794. IEEE, 2020.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jialu Sui, Yiyang Ma, Wenhan Yang, Xiaokang Zhang, Man-On Pun, and Jiaying Liu. Diffusion enhancement for cloud removal in ultra-resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

Meilin Wang, Yexing Song, Pengxu Wei, Xiaoyu Xian, Yukai Shi, and Liang Lin. Idf-cr: Iterative diffusion process for divide-and-conquer cloud removal in remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 8798–8807. Ieee, 2018.

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

Ziyao Wang, Maonan Wang, Yucheng He, Xianping Ma, Ziyi Wang, Hongyang Zhang, Yirong Cheng, and Man-on Pun. Interpretation-oriented cloud removal via observation-anchored residual flow with geo-contextual alignment. In *European Conference on Computer Vision*, pp. 248–265. Springer, 2026.

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/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Yuxiang Zhang, Shunlin Liang, Wenyuan Li, Han Ma, Jianglei Xu, Yichuan Ma, Jiangwei Xie, Wei Li, Mengmeng Zhang, Ran Tao, et al. Units: unified spatio-temporal generative model for remote sensing. *arXiv preprint arXiv:2512.04461*, 2025.

Anqi Zhao, Ruitao Feng, and Xinghua Li. Thiefcloud: A thickness fused thin cloud removal network for optical remote sensing image with self-supervised learnable cloud prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(12):11834–11848, 2025.

Hangyu Zhou, Chia-Hsiang Kao, Cheng Perng Phoo, Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Allclear: A comprehensive dataset and benchmark for cloud removal in satellite imagery. *Advances in Neural Information Processing Systems*, 37:53571–53597, 2024.

Xuechao Zou, Kai Li, Junliang Xing, Yu Zhang, Shiying Wang, Lei Jin, and Pin Tao. Diffcr: A fast conditional diffusion framework for cloud removal from optical satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

APPENDIX

This *Supplementary Material* provides further analysis of GeoCR's shared latent interface, implementation details, additional results, and evaluation protocols. Table 8 summarizes its organization.

Table 8: Overview of the *Supplementary Material*.

Section	Contents
Section A	Stem adaptation
Section B	Implementation details
Section C	Additional quantitative results
Section D	Additional qualitative results
Section E	Failure cases and limitations
Section F	Datasets and preprocessing
Section G	Evaluation protocols

A STEM ADAPTATION

Reusing the pretrained representation. Stage 1 adapts compact input and output stems while keeping the pretrained RGB encoder–decoder trunks frozen. Figure 8 compares reconstruction across RGB, single-band NIR, ten-band optical, and dual-polarization SAR inputs. The trained stems improve reconstruction of measurements whose channel layouts and statistics differ from RGB, while the original RGB route remains unchanged. The single-band NIR route retains its initialization and requires no stem training.

Reconstruction across configurations. Table 9 complements the held-out stem-training ablation in the main paper with per-source reconstruction diagnostics. The initialized NIR route achieves 42.06 dB on Sen2_MTC_New, 40.41 dB on CUHK-CR, and 49.51 dB on Sen2_MTC_Old. These diagnostics support retaining the initialized NIR interface while adapting stems for ten-band optical and SAR observations. They measure reconstruction of each input domain, rather than cloud removal or preservation of every spectral relationship.

B IMPLEMENTATION DETAILS

Autoencoder. All observation routes share the frozen encoder and decoder trunks of the pretrained FLUX.2 RGB autoencoder. Each input and output stem consists of a single convolutional layer. We retain the original RGB projections and the initialized single-band NIR stems, and train separate stems for ten-band TOA, ten-band BOA, and VV/VH SAR inputs using an ℓ_1 reconstruction loss. The added stems contain 53.5K parameters, of which 51.1K are trained. The encoder produces 32-channel posterior-mean latents at $1/8$ spatial resolution, which are packed into 128 channels at $1/16$ resolution and normalized using fixed BatchNorm statistics. Decoding reverses normalization and packing. All stems and trunks remain frozen after Stage 1.

Generator. GeoCR builds on FLUX.2 [klein] 4B Base (Labs, 2025), with five double-stream and twenty single-stream blocks, a hidden dimension of 3,072, and 24 attention heads. RGB and available non-RGB latents are concatenated along channels within each optical observation. The noisy target, each of the K cloudy observations, and optional SAR observations retain separate token sequences, yielding $1 + K + \mathbf{1}_{\mathrm{SAR}}$ token streams. The cloudy frames are therefore combined through attention rather than channel-wise concatenation across time. Double-stream blocks allow the target and conditioning tokens to interact through joint attention; subsequent single-stream blocks process their combined sequence with shared parameters. Only target tokens are passed to the final velocity head. GeoCR (w/o FT) uses the shared pretrained checkpoint, whereas GeoCR (LoRA) learns dataset-specific low-rank updates (Hu et al., 2021) while keeping the autoencoder fixed.

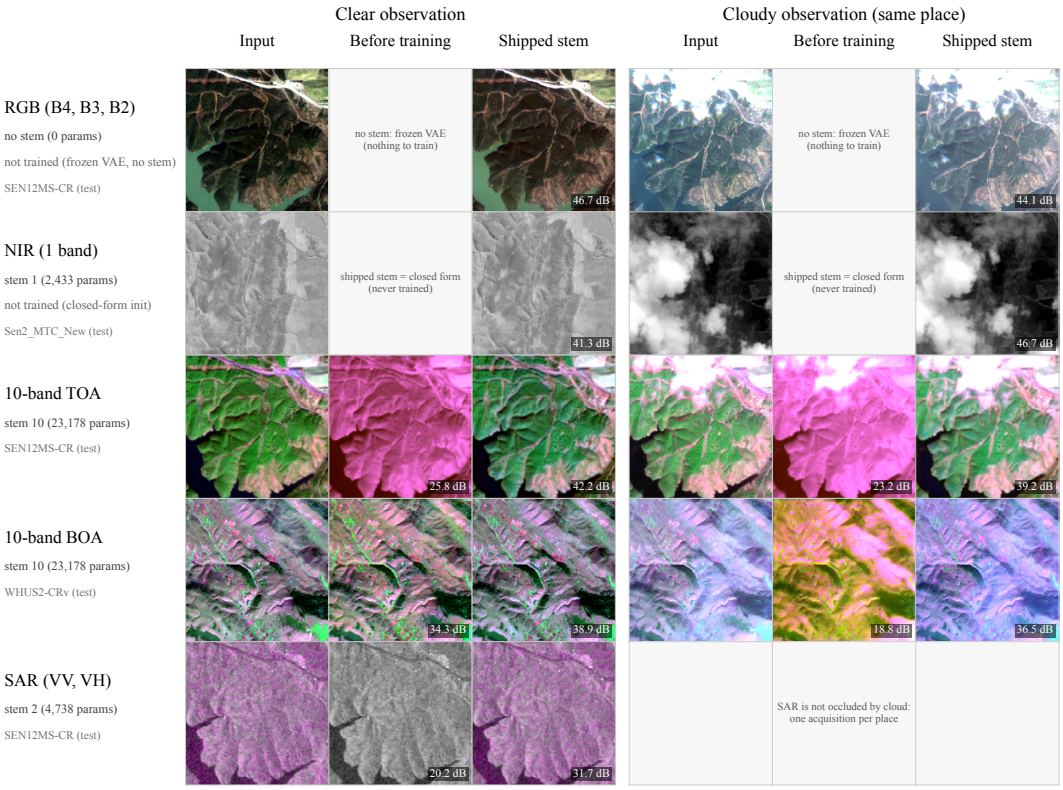

Figure 8: **Reconstruction through the shared autoencoder.** The pretrained RGB route and initialized NIR route remain fixed during Stage 1. Ten-band optical and SAR routes use learned stems with frozen encoder–decoder trunks. Ten-band visualizations use B12/B8/B5 false color, and SAR uses a VV/VH/VV composite.

Table 9: **Reconstruction by configuration.** PSNR (dB) is computed from pooled reconstruction error on 64 training images per source, without clipping. These diagnostics differ from the held-out before/after comparison in the main paper and do not measure cloud removal.

Route	Stem parameters	Trained	Reconstruction PSNR
10-band TOA	23,178	Yes	AllClear 38.23; SEN12MS-CR 37.69
10-band BOA	23,178	Yes	WHUS2-CRv 37.89
Single-band NIR	2,433	No	Sen2_MTC_New 42.06; CUHK-CR 40.41; Sen2_MTC_Old 49.51
SAR VV/VH	4,738	Yes	AllClear 28.51; SEN12MS-CR 31.19

C Additional Quantitative Results

Table 10 extends the main comparisons to T-CLOUD, CUHK-CR1, and WHUS2-CRv. The native and RGB-only settings follow the preprocessing and metric definitions in Sections F and G. Bold and underlined values indicate the best and second-best results.

Performance across observation settings. On T-CLOUD, specialized methods retain advantages in perceptual or pixel-level fidelity. On CUHK-CR1, GeoCR achieves competitive FID and the lowest KID, while EMRDM performs better on DISTS, LPIPS, SSIM, and PSNR. On WHUS2-CRv, GeoCR remains competitive in distributional and perceptual quality, although EMRDM leads the native setting and HI-Diff leads the RGB-only setting. LoRA provides selective improvements rather than consistent gains over direct inference. Together, these results show the reach of a shared cloud removal prior while identifying settings that remain challenging.

Table 10: **Additional quantitative comparisons.** Feature metrics use RGB views; PSNR/SSIM follow Table 12.

Method	FID↓	DISTS↓	KID↓	DINO↑	LPIPS↓	SSIM↑	PSNR↑
(a) T-CLOUD: RGB display imagery							
pix2pix (Isola et al., 2017)	107.6	0.254	0.0509	0.450	0.330	0.635	20.13
pix2pixHD (Wang et al., 2018)	60.8	0.169	0.0173	0.616	0.177	0.741	24.75
BBDM (Li et al., 2023)	186.6	0.422	0.1179	0.259	0.546	0.541	22.69
HI-Diff (Chen et al., 2023)	40.2	0.124	0.0070	0.726	0.114	**0.874**	**30.54**
UnCRtainTS (Ebel et al., 2023)	63.3	0.180	0.0188	0.640	0.178	0.809	26.28
DiffCR (Zou et al., 2024)	139.6	0.358	0.0732	0.376	0.482	0.402	20.67
IDF-CR (Wang et al., 2024)	84.4	0.220	0.0304	0.467	0.236	0.787	25.99
ThiefCloud (Zhao et al., 2025)	40.4	0.123	0.0058	0.726	**0.109**	0.861	29.36
EMRDM (Liu et al., 2025)	36.5	0.121	0.0045	0.759	0.110	0.867	28.25
GACR (Wang et al., 2026)	39.2	**0.120**	0.0058	0.747	0.113	0.858	29.56
GeoCR (w/o FT)	36.5	0.126	0.0024	**0.763**	0.122	0.785	27.15
GeoCR (LoRA)	**36.3**	0.125	**0.0022**	0.751	0.122	0.786	27.23
(b) CUHK-CR1: RGB+NIR setting							
UnCRtainTS	134.8	0.204	0.0283	0.712	0.311	0.679	23.82
DiffCR	237.6	0.291	0.1360	0.649	0.318	0.573	22.75
EMRDM	**77.8**	**0.115**	-0.0006	**0.845**	**0.146**	**0.760**	**25.64**
GACR	97.1	0.133	0.0099	0.825	0.160	0.726	25.02
GeoCR (w/o FT)	80.4	0.125	-0.0025	0.825	0.165	0.680	23.88
GeoCR (LoRA)	81.5	0.125	**-0.0027**	0.821	0.167	0.677	23.83
(c) WHUS2-CRv: native multispectral setting							
UnCRtainTS	26.9	0.142	0.0062	0.808	0.139	0.926	31.14
IDF-CR	40.0	0.179	0.0115	0.691	0.188	0.858	29.00
EMRDM	**16.9**	**0.101**	**0.0009**	**0.875**	**0.097**	**0.937**	**32.55**
GACR	22.6	0.230	0.0033	0.789	0.138	0.882	31.09
GeoCR (w/o FT)	18.9	0.111	0.0027	0.850	0.106	0.905	32.29
GeoCR (LoRA)	18.5	0.107	0.0025	0.858	0.103	0.903	32.15
(d) WHUS2-CRv: RGB-only setting							
pix2pix	95.1	0.238	0.0523	0.432	0.274	0.746	23.75
pix2pixHD	27.9	0.148	0.0057	0.751	0.145	0.828	26.77
BBDM	50.4	0.213	0.0171	0.413	0.291	0.639	25.19
HI-Diff	**17.8**	**0.107**	**0.0018**	**0.853**	**0.096**	**0.894**	**29.91**
GeoCR (w/o FT)	21.6	0.137	0.0032	0.786	0.133	0.833	27.16
GeoCR (LoRA)	20.7	0.138	0.0039	0.811	0.134	0.804	25.35

D Additional Qualitative Results

Perceptual quality and reference fidelity. Figure 9 compares the two GeoCR modes with strong dataset-specific baselines across different observation configurations. The examples illustrate recovery of scene layout and local appearance, as well as cases where a specialized baseline better matches the reference. LoRA produces relatively small changes in these examples and does not improve every image. The two SEN12MS-CR rows differ in both available observations and rendering, so their contrast does not isolate the contribution of SAR.

Additional examples by dataset. Figures 11–19, collected at the end of this supplement, provide four examples for each of the six native dataset settings and three RGB-only settings. Each figure compares GeoCR (w/o FT) and GeoCR (LoRA) with the available baselines and paired clear references. All visualizations show RGB views.

E Failure Cases and Limitations

Figure 10 examines cases in which the available observations or the learned prior are insufficient to reproduce the reference reliably.

Incomplete evidence. When every optical observation is heavily obscured, generated details may appear plausible while differing from the reference. SAR provides complementary structural evidence but does not determine a unique optical appearance. Fine structures under complete cloud therefore remain difficult to recover reliably, even when the scene layout is convincing.

Acquisition and radiometric differences. Surface changes between acquisitions can make the observed scene inconsistent with the target. The model may preserve the input state rather than reconstruct an unobserved state at another time. Photometric offsets also remain visible on T-CLOUD.

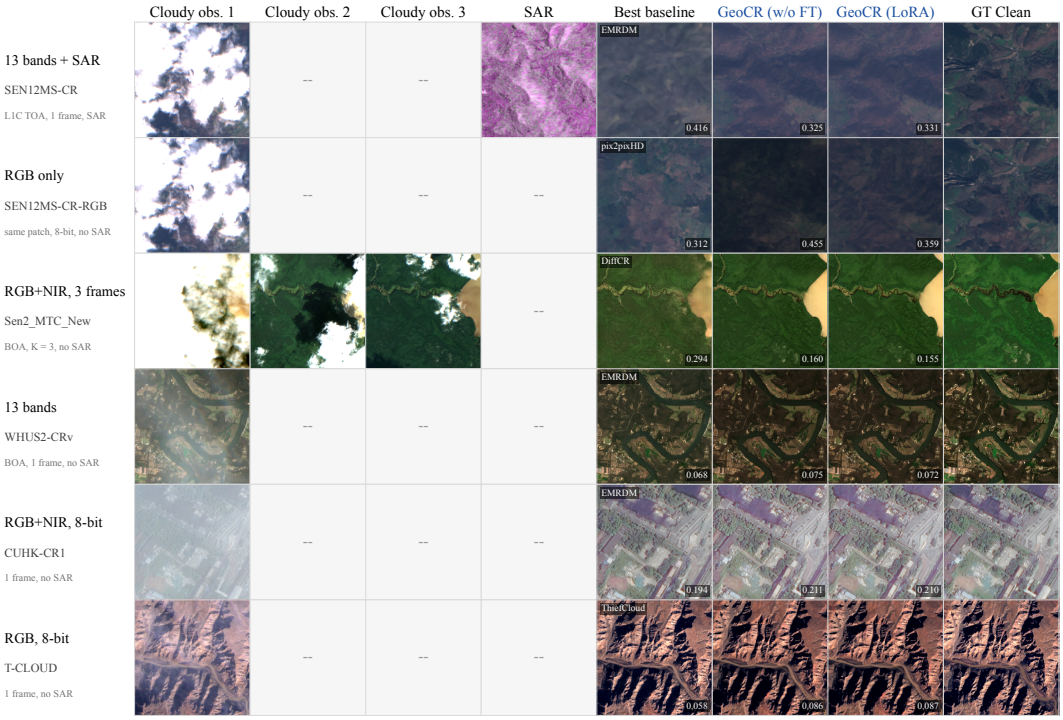

Figure 9: **Representative cloud removal results.** Rows show native and RGB-only SEN12MS-CR, Sen2_MTC_New, WHUS2-CRv, CUHK-CR1, and T-CLOUD. Baselines are selected by dataset-level LPIPS; numbers denote per-image LPIPS. The first two rows show the same scene.

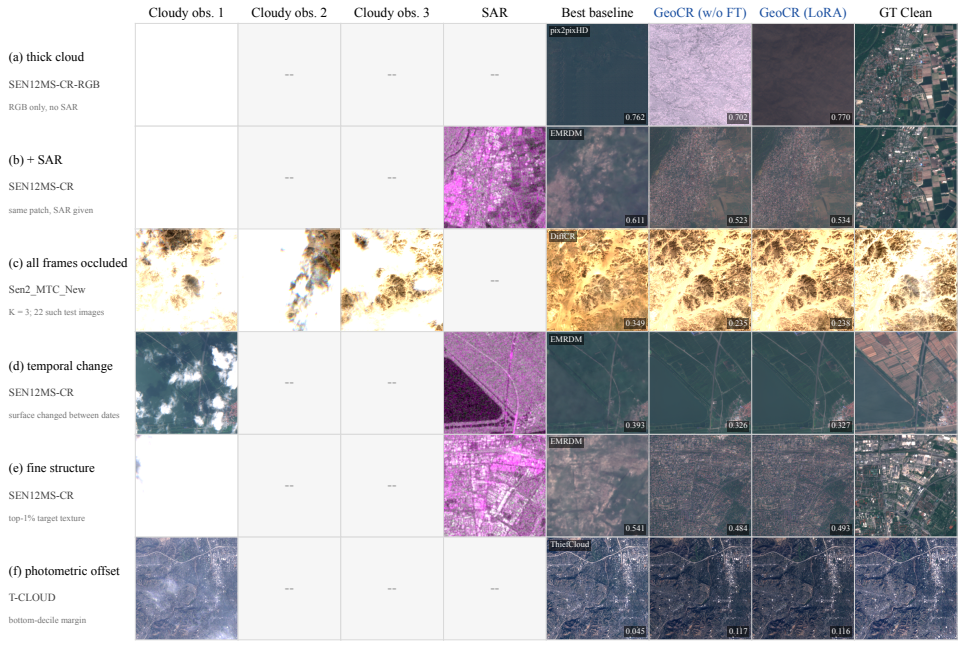

Figure 10: **Challenging cloud removal cases.** Examples cover severe occlusion, the same scene with multispectral and SAR inputs, occlusion across all temporal observations, temporal surface change, fine urban structure, and photometric mismatch. Cases are selected using occlusion, change, texture, or error criteria. Numbers indicate per-image LPIPS. Rows (a) and (b) change both optical bands and SAR availability.

Table 11: **Dataset representations and pretraining mixture.** Bands describe optical targets; K is the number of cloudy observations. Mix denotes the source sampling probability.

Source	Optical representation	SAR	K	Mix (%)	Training targets
AllClear (Zhou et al., 2024)	13 bands, L1C TOA	VV/VH	1–3	53.14	662,022
SEN12MS-CR (Ebel et al., 2020)	13 bands, L1C TOA	VV/VH	1	32.93	107,143
WHUS2-CRv (Li et al., 2022)	13 bands, BOA; B10 TOA	–	1	8.81	18,816
Sen2_MTC_Old (Sarukkai et al., 2020)	RGB; NIR in cloudy inputs only	–	3	2.00	88,874
Sen2_MTC_New (Huang & Wu, 2022)	RGB+B8, BOA	–	3	1.10	2,380
T-CLOUD (Ding et al., 2022)	RGB, 8-bit	–	1	1.10	2,234
RICE2 (Lin et al., 2019)	RGB, 8-bit	–	1	0.27	553
CUHK-CR1 (Sui et al., 2024)	RGB+NIR, 8-bit	–	1	0.25	508
CUHK-CR2 (Sui et al., 2024)	RGB+NIR, 8-bit	–	1	0.21	426
RICE1 (Lin et al., 2019)	RGB, 8-bit	–	1	0.19	375
Total				100.00	**883,331**

PSNR and SSIM measure agreement with the supplied reference, while perceptual metrics characterize complementary aspects of quality; neither establishes the physical accuracy of obscured content.

Scope of generalist behavior. GeoCR supports the evaluated spectral, temporal, and SAR configurations through one pretrained parent. All evaluated source datasets contribute training data, and transfer to an entirely unseen corpus is not assessed. Further evaluation is needed for unseen sensors, geographically independent transfer, and applications requiring calibrated spectral fidelity.

F Datasets and Preprocessing

Training corpus and evaluation scope. We combine the training splits of ten cloud removal datasets, totaling 883,331 cloud-free target images (Table 11). AllClear, Sen2_MTC_Old, and RICE1/2 contribute only pretraining data; the remaining six sources also provide evaluation benchmarks. Three additional RGB-only settings are derived from SEN12MS-CR, Sen2_MTC_New, and WHUS2-CRv. Evaluation uses held-out samples from the contributing sources.

Optical and SAR normalization. For Sentinel-2 imagery, we convert the stored digital numbers (DN) to reflectance as $\rho = \mathrm{DN}/10^4$ and normalize inputs as $2\,\mathrm{clip}(\rho, 0, 1) - 1$. Eight-bit imagery is normalized as $2u/255 - 1$, where u denotes the stored pixel value. For the separately rendered RGB-only datasets, we first compute $u = \mathrm{round}[255\,\mathrm{clip}(2\rho, 0, 1)]$ and apply the same 8-bit normalization. SAR measurements s in decibels (dB) are clipped to $[-30, 0]$ and mapped to $[-1, 1]$ as $2[\mathrm{clip}(s, -30, 0) + 30]/30 - 1$. TOA and BOA denote top- and bottom-of-atmosphere reflectance, respectively; L1C denotes the Sentinel-2 Level-1C product.

Spatial and temporal inputs. WHUS2-CRv bands are aligned to the 10 m grid by nearest-neighbor replication of the 20 m and 60 m bands. For AllClear, cloudy observations are sampled without replacement within ± 40 days of the target, excluding its acquisition date; SAR is the nearest available acquisition within this window. Eligible AllClear targets contain at most 10% cloud, 10% shadow, and 1% no-data coverage.

Split checks. We use exact region-of-interest (ROI) identifiers for SEN12MS-CR, correct scene assignments in WHUS2-CRv, and exclude 104 Sen2_MTC_Old samples that overlap Sen2_MTC_New test tiles. CUHK-CR1 evaluation excludes 32 train/validation duplicates and four within-test duplicates. RICE1/2 contribute only pretraining data because their supplied splits contain duplicate images. We will release the split-audit results documenting these exclusions.

Table 12: **Evaluation settings.** N_{test} denotes the size of each evaluated test split. CUHK-CR1 counts reflect the duplicate exclusions described in Section F.

Setting	PSNR/SSIM bands	N_{test}
SEN12MS-CR	13 bands	7,899
Sen2_MTC_New	RGB+B8	687
WHUS2-CRv	13 bands	3,746
T-CLOUD	RGB	588
CUHK-CR1	RGB+NIR	98
CUHK-CR2	RGB	111
SEN12MS-CR, RGB-only	RGB	7,899
Sen2_MTC_New, RGB-only	RGB	687
WHUS2-CRv, RGB-only	RGB	3,746

G Evaluation Protocols

Test sets and spectral scope. Table 12 lists the evaluated test splits and the bands used for PSNR and SSIM. FID, DISTS, KID, DINO similarity, and LPIPS are computed on RGB views, including in the native multispectral settings. Within each setting, competing methods are evaluated on common samples and spatial extents for each metric. FID and DISTS are the primary measures of distributional realism and paired perceptual fidelity, respectively. Lower values are better for FID, DISTS, KID, and LPIPS; higher values are better for DINO similarity, SSIM, and PSNR.

Distributional measures. FID (Heusel et al., 2017) compares Gaussian approximations to the generated and reference Inception-feature distributions. KID (Bińkowski et al., 2018) estimates squared maximum mean discrepancy using a polynomial kernel. Both use 2,048-dimensional Inception pool-3 features with 299×299 bilinear resizing. These measures compare image sets and do not directly evaluate correspondence to an individual cloudy input.

Perceptual and semantic measures. DISTS (Ding et al., 2020) compares structural and textural statistics using its released VGG-16 model. LPIPS (Zhang et al., 2018) measures distances between normalized deep features with learned perceptual weights; we use AlexNet v0.1. DINO similarity (Siméoni et al., 2025) averages cosine similarity between spatially aligned patch tokens from the frozen DINOv3-SAT ViT-L/16, excluding CLS and register tokens. All three compare each reconstruction with its paired clear reference.

Reference fidelity. SSIM (Wang et al., 2004) compares local luminance, contrast, and structure, while PSNR expresses pixelwise mean squared error relative to the squared intensity range on a logarithmic scale. We average PSNR over per-image scores. SSIM uses an 11×11 Gaussian window for native Sentinel-2 settings and a 7×7 uniform window for display/RGB settings. These measures quantify spatial and radiometric agreement with the reference and remain sensitive to misregistration and acquisition differences.

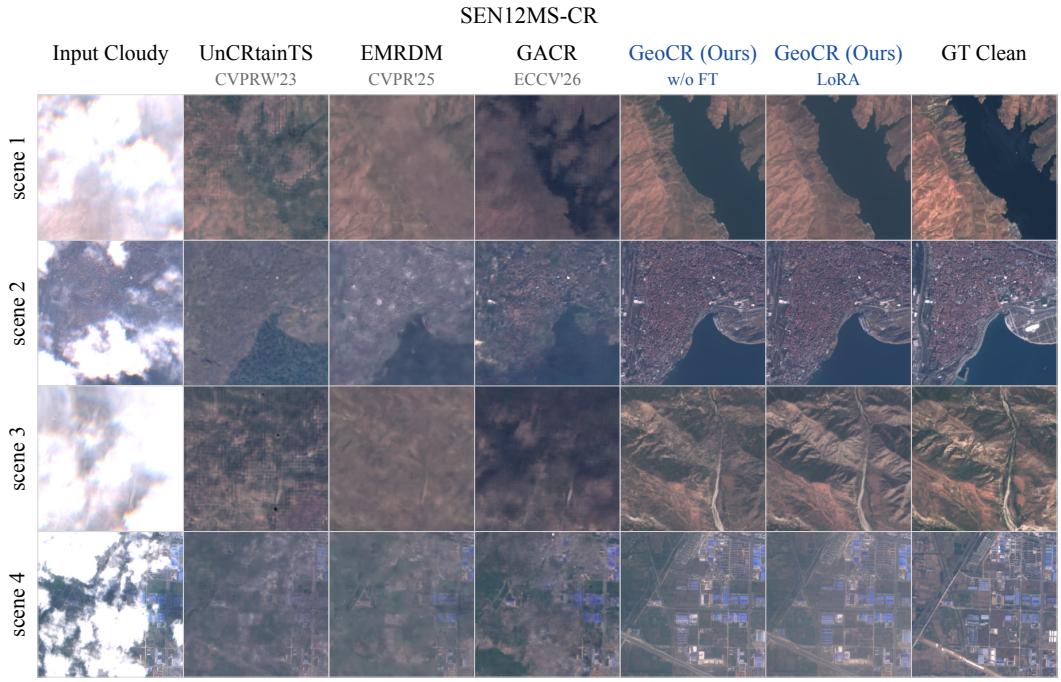

Figure 11: **Qualitative comparison on SEN12MS-CR (Ebel et al., 2020).** Four examples from the native multispectral setting, shown as RGB views with paired clear references. GeoCR is evaluated without fine-tuning (w/o FT) and with LoRA adaptation.

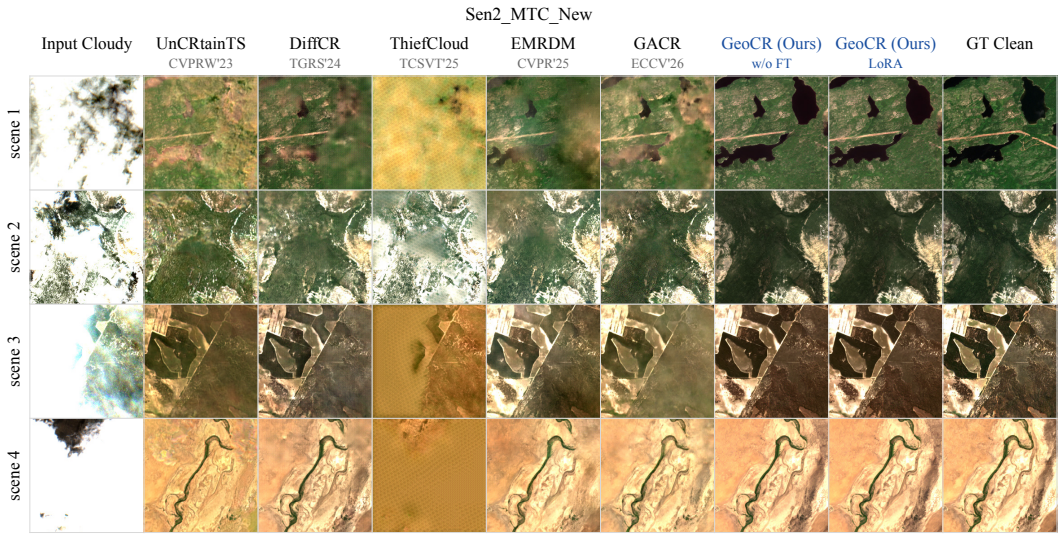

Figure 12: **Qualitative comparison on Sen2_MTC_New (Huang & Wu, 2022).** Four examples from the native multi-temporal setting, shown as RGB views with paired clear references. GeoCR is evaluated without fine-tuning (w/o FT) and with LoRA adaptation.

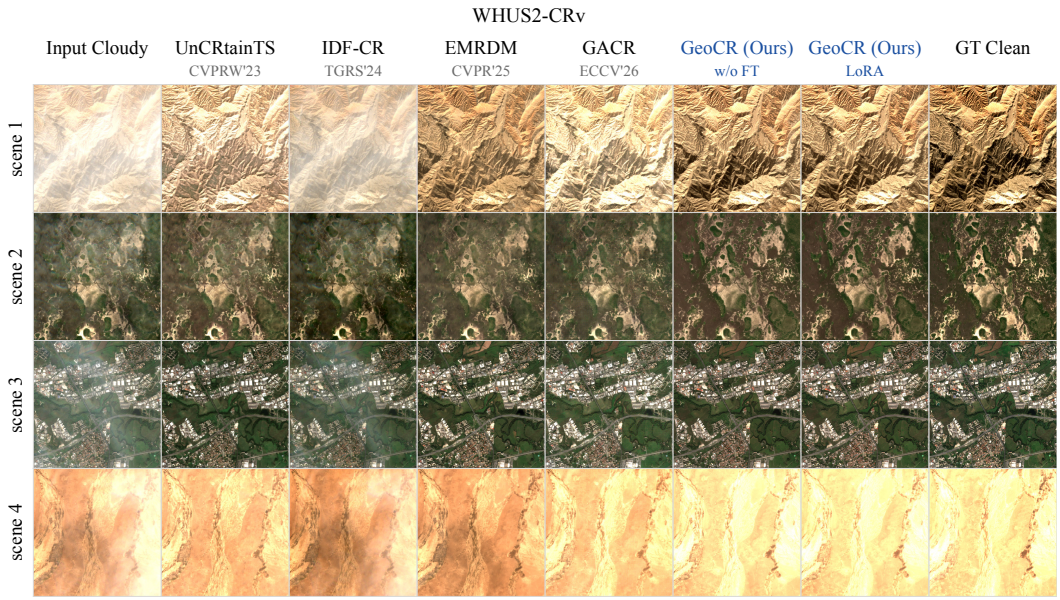

Figure 13: **Qualitative comparison on WHUS2-CRv (Li et al., 2022).** Four examples from the native multispectral setting, shown as RGB views with paired clear references. GeoCR is evaluated without fine-tuning (w/o FT) and with LoRA adaptation.

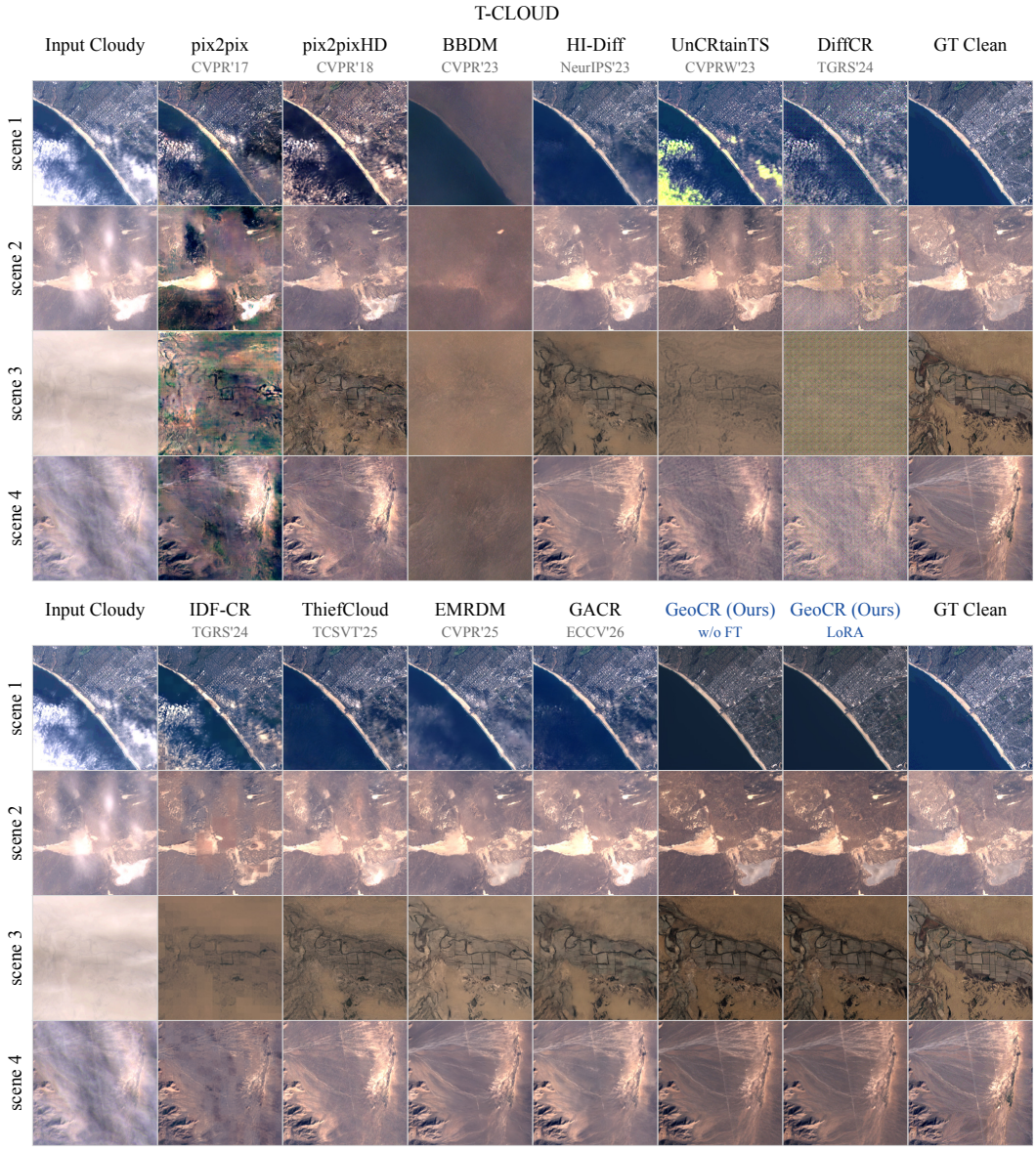

Figure 14: **Qualitative comparison on T-CLOUD (Ding et al., 2022).** The two panels show the same four scenes with different groups of competing methods. GeoCR is evaluated without fine-tuning (w/o FT) and with LoRA adaptation.

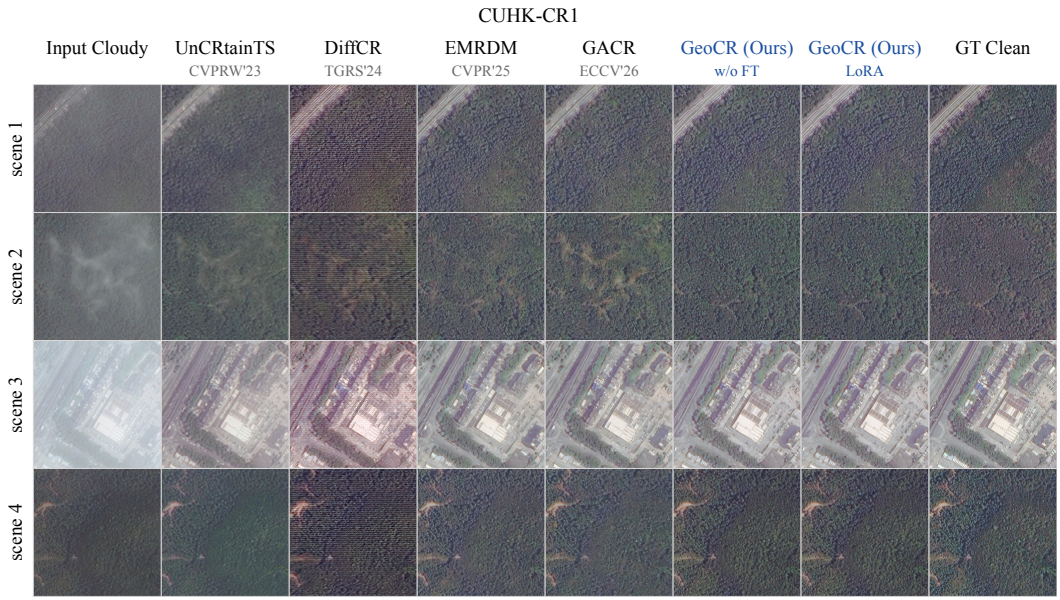

Figure 15: **Qualitative comparison on CUHK-CR1 (Sui et al., 2024).** Four examples from the RGB+NIR setting, shown as RGB views with paired clear references. GeoCR is evaluated without fine-tuning (w/o FT) and with LoRA adaptation.



Figure 16: **Qualitative comparison on CUHK-CR2 (Sui et al., 2024).** The two panels show the same four scenes with different groups of competing methods under the RGB evaluation setting. GeoCR is evaluated without fine-tuning (w/o FT) and with LoRA adaptation.

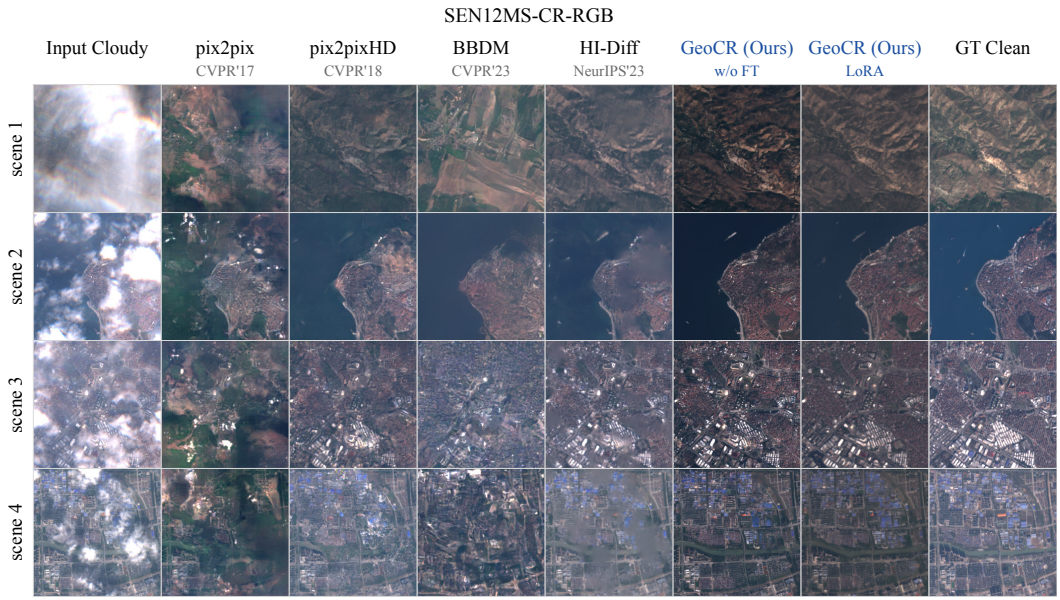

Figure 17: **Qualitative comparison on SEN12MS-CR (Ebel et al., 2020) (RGB-only).** Four examples from the RGB-only setting, with one cloudy RGB input and no SAR guidance. GeoCR is evaluated without fine-tuning (w/o FT) and with LoRA adaptation.

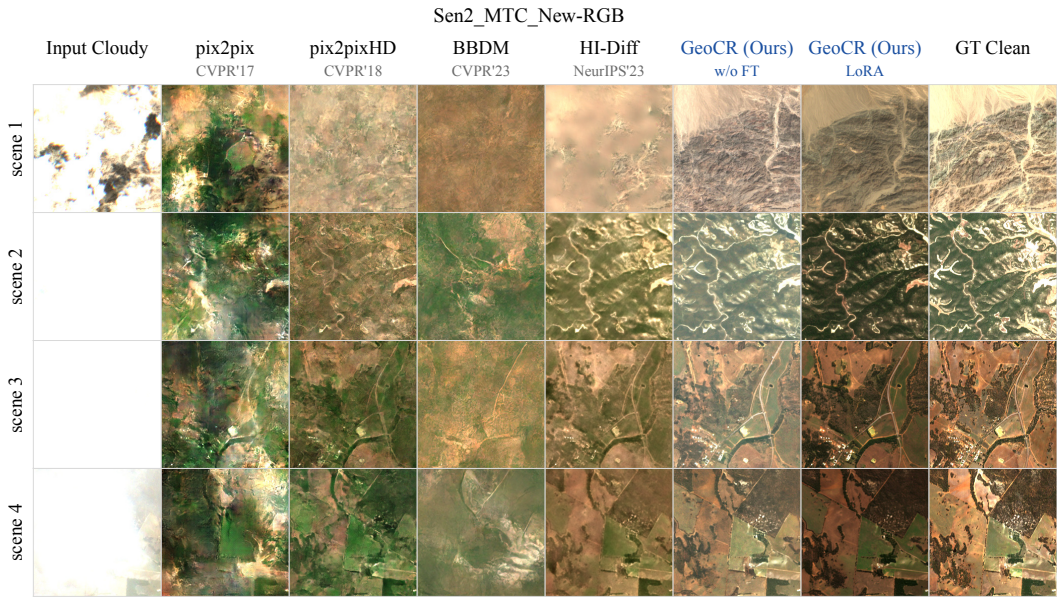

Figure 18: **Qualitative comparison on Sen2_MTC_New (Huang & Wu, 2022) (RGB-only).** Four examples from the RGB-only setting, with one cloudy RGB input and no SAR guidance. GeoCR is evaluated without fine-tuning (w/o FT) and with LoRA adaptation.

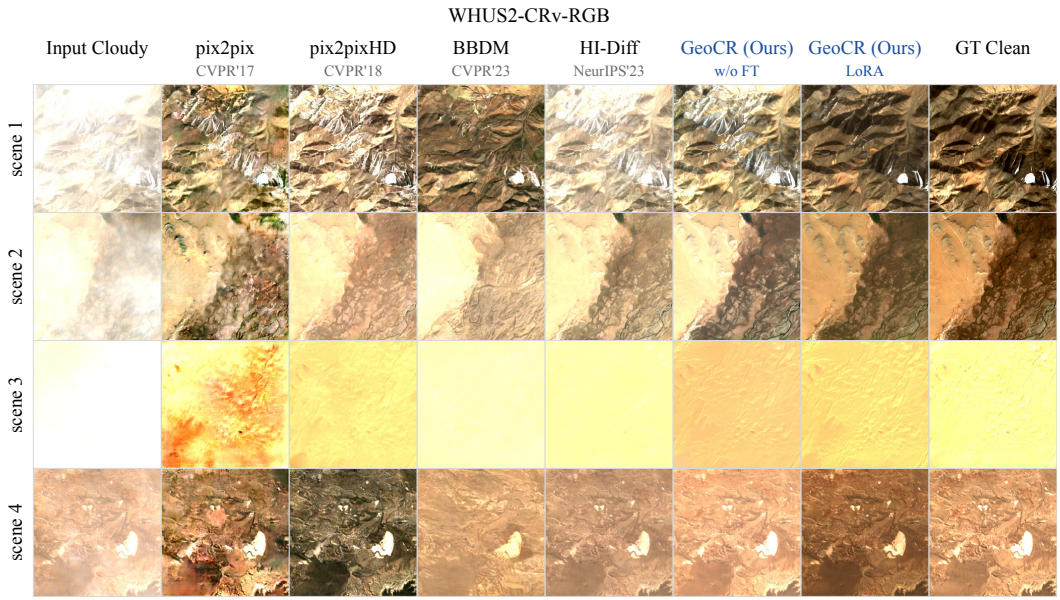

Figure 19: **Qualitative comparison on WHUS2-CRv (Li et al., 2022) (RGB-only).** Four examples from the RGB-only setting, with one cloudy RGB input and no SAR guidance. GeoCR is evaluated without fine-tuning (w/o FT) and with LoRA adaptation.