ConsistDreamer: 3D-Consistent 2D Diffusion for High-Fidelity Scene Editing

Jun-Kun Chen, Samuel Rota Bulò, Norman Müller, Lorenzo Porzi, Peter Kontschieder, Yu-Xiong Wang

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN · Meta · Project Page · CVPR SEATTLE, WA JUNE 17-21, 2024

Task and Motivation

- **Task**: Instruction-guided 3D scene editing

Original 3D Scene → Editing Instruction *"Turn him into an Elf"* → Edited 3D Scene

- **Widely-Adopted Solution**: Distill editing signals from 2D diffusion

Rendered Images → 2D Diffusion → Edited Images
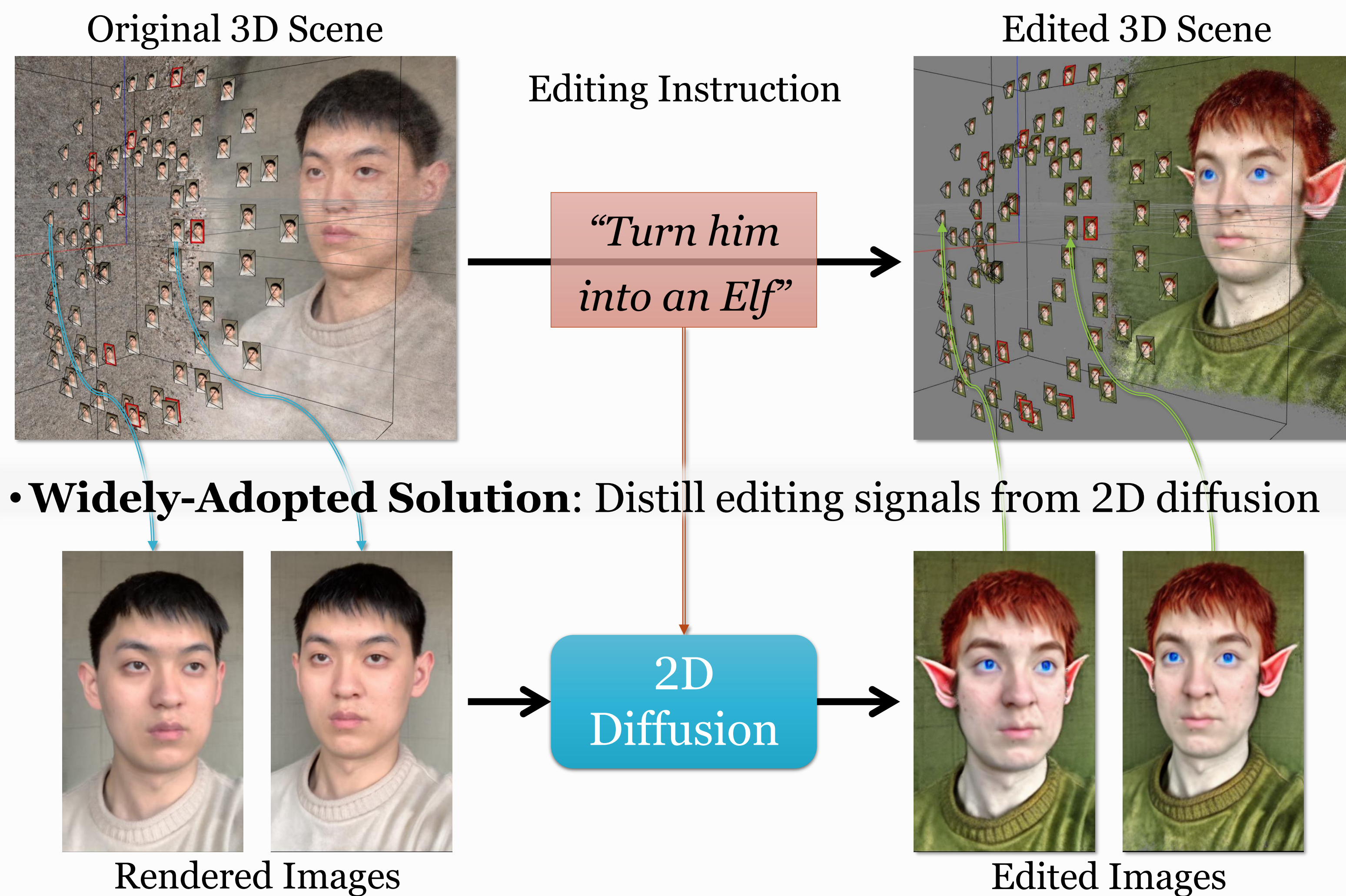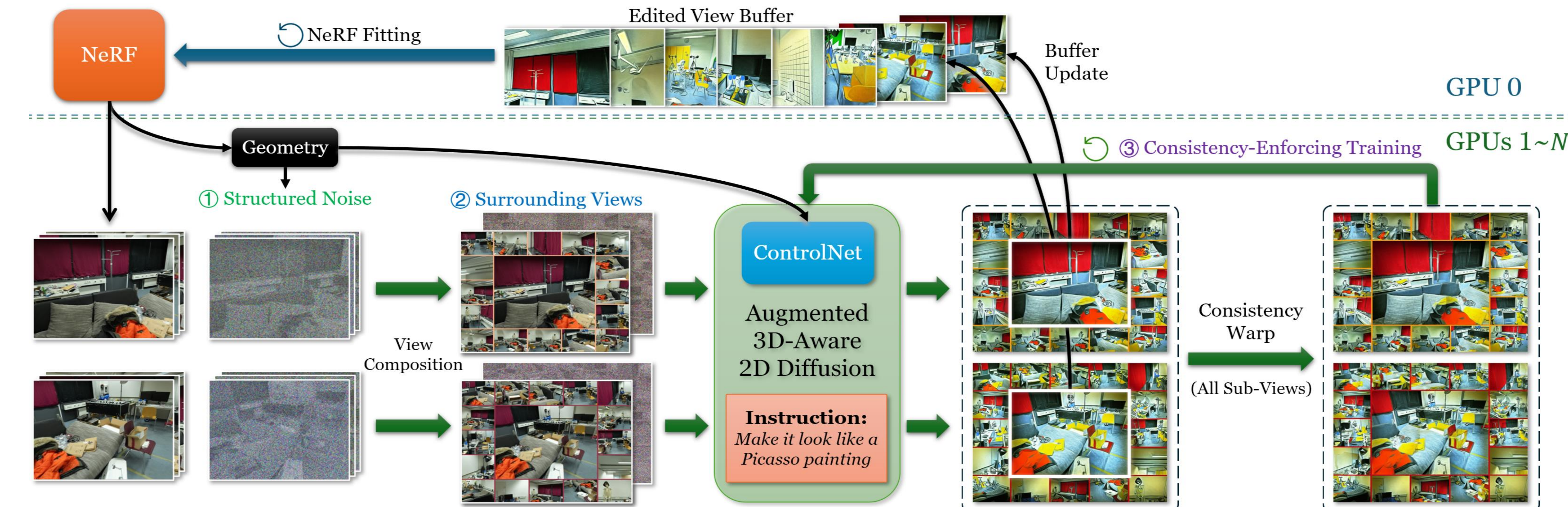
- **Fundamental Limitation**: Lack of 3D consistency
  - 2D model acting independently on different views produces inconsistent results
  - Converge towards an "average" of inconsistent results
  - Common failure case:
    Plaid pattern, *first solved* by our ConsistDreamer →

Original Scene · IN2N (Baseline) · **Ours**

Multi-GPU Pipeline
- GPU0: Iterative NeRF fitting
- GPUs 1~N: Diffusion generation and training
  ➤ Disentangle NeRF from diffusion model
  ➤ Improved efficiency: Edit N surrounding view images in parallel

③ **Consistency-Enforcing Training**
- Augmentation of the 2D diffusion: ControlNet for 3D-awareness
- Self-supervised loss: Construct the target ② surrounding view through warp-averaging all 13N views
- Consistent denoising procedure: Supervise intermediate denoising steps

NeRF — NeRF Fitting — Edited View Buffer — Buffer Update · GPU 0

Geometry · ① Structured Noise · ② Surrounding Views · ③ Consistency-Enforcing Training · GPUs 1~N

View Composition → ControlNet / Augmented 3D-Aware 2D Diffusion / **Instruction:** *Make it look like a Picasso painting* → Consistency Warp (All Sub-Views)
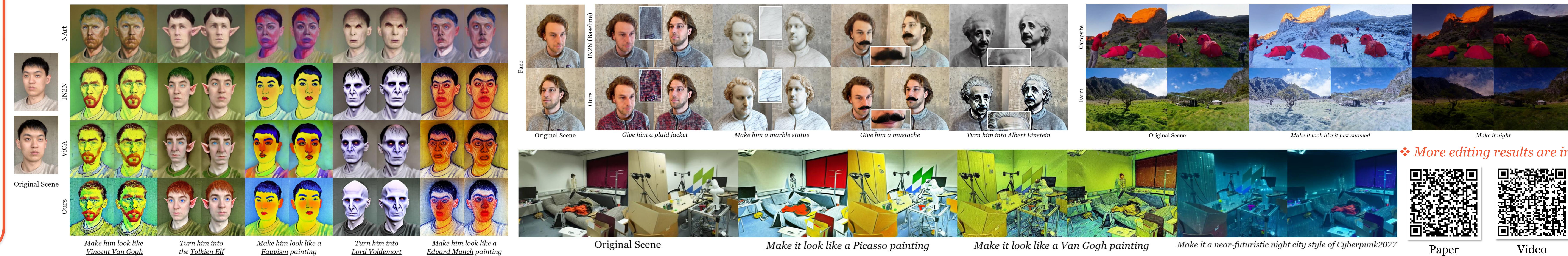
① **3D-Consistent Structured Noise**
- Render 3D-consistent noise maps for each view
  ➤ Enable a 3D-consistent denoising procedure

② **Surrounding Views**
- Composed by surrounding a central view with 12 small reference views
  ➤ Provide rich context to the diffusion model and facilitates ③ training

Our ConsistDreamer generates high-quality edited scenes with clear textures, bright colors, and high instruction fidelity, achieving state-of-the-art scene editing results


Make him look like *Vincent Van Gogh* · Turn him into *the Tolkien Elf* · Make him look like a *Fauvism painting* · Turn him into *Lord Voldemort* · Make him look like an *Edvard Munch painting*

Give him a plaid jacket · Make him a marble statue · Give him a mustache · Turn him into Albert Einstein

Make it look like it just snowed · Make it night

Original Scene · Make it look like a Picasso painting · Make it look like a Van Gogh painting · Make it a near-futuristic night city style of Cyberpunk2077

More editing results are in — Paper · Video

Key Insights: Three Synergistic Strategies

① Generate 3D-consistent structured noise as foundation of consistency
② Compose surrounding views as 3D-context-rich input
③ Enforce consistent denoising procedure through concurrent training

Quantitative Evaluation & Ablation Study

Our Novel Metric:
Distillation Fidelity Score (DFS) ↓
- Quantify the fidelity of distillation
- FID ↓ between:
  (1) *rendered images of edited scene*, and
  (2) *edited images by diffusion*

Ablation Study: Relative DFS ↓
Task A · Task B · Task C · Task D

Full (As 1) · No ① · No ③ · No ①③ · No ①②③
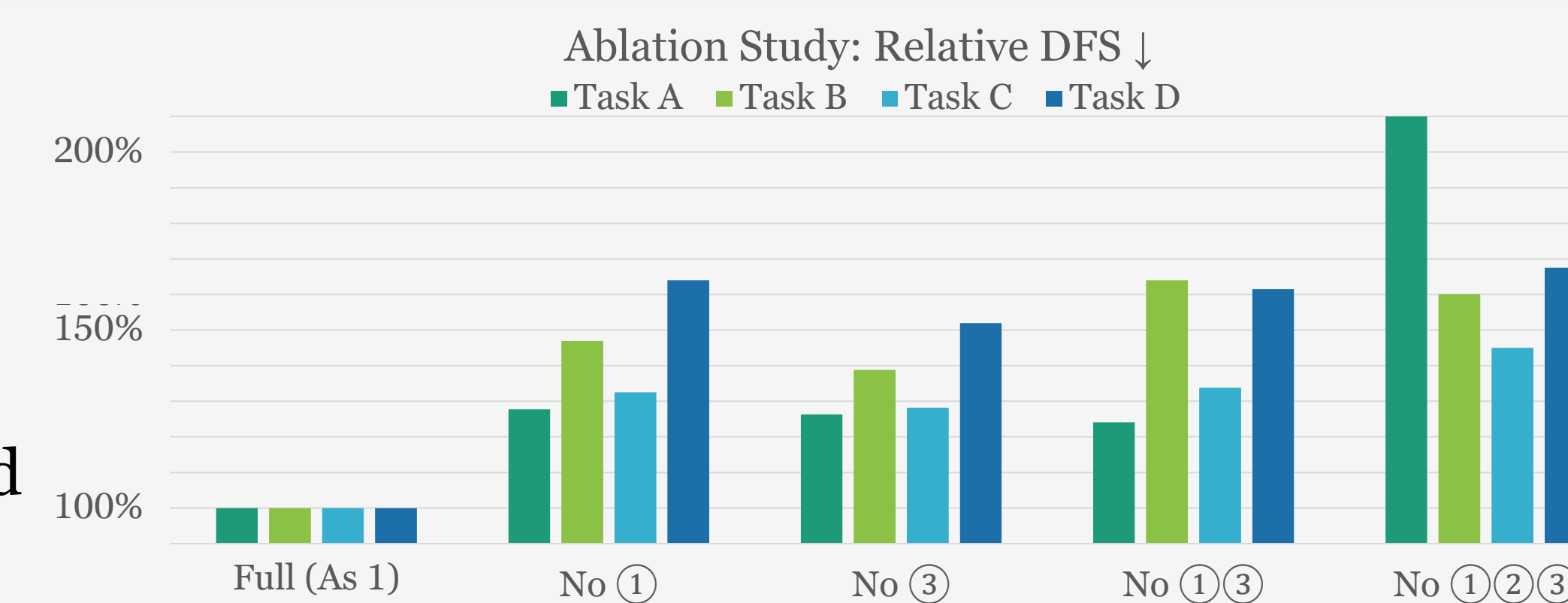
Take-Aways

- **ConsistDreamer** enables 3D-consistent instruction-guided scene editing based on 2D-diffusion-distillation
- **ConsistDreamer**'s three synergistic strategies lift 2D diffusion models to generate 3D-consistent images
- **ConsistDreamer** achieves state-of-the-art editing results across various scenes and editing tasks