NEURIPS 2026 · POSTER

CRISP: Fixing Flying Pixels in Latent LiDAR Generation via Diffusion Decoding

Andrea Ceron[1 2], Michael Schmidt[2], Alvaro Marcos-Ramiro[2], Sebastian Schmidt[1 2], Benjamin Busam[1]

[1] TU München [2] BMW AG

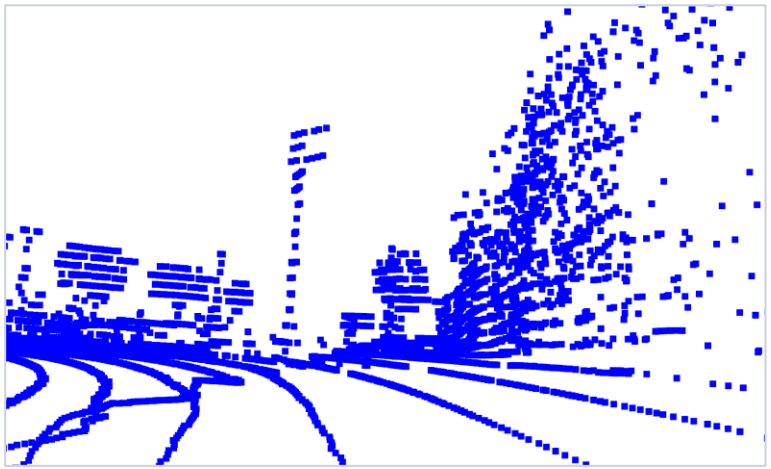


Paris Poster Session 6 · Fri 11 Dec 2026 · 14:30 CET

THE PROBLEM

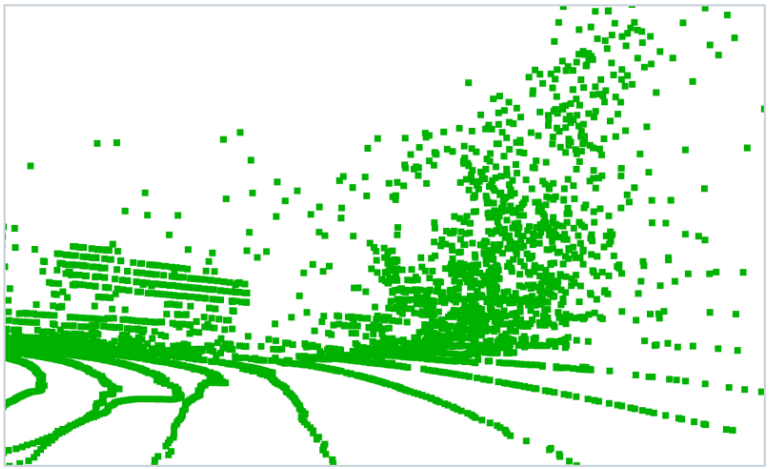
Convolutional VAE decoders make LiDAR points fly

Ground truth



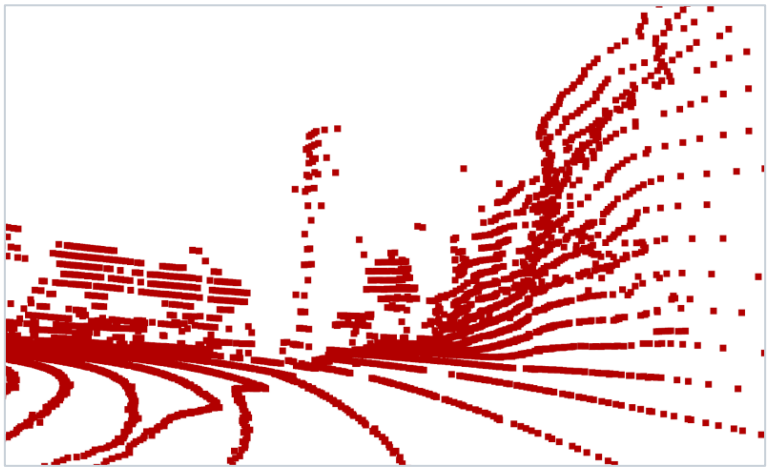
The real sweep: sharp edges, empty gaps

Baseline decoder



Points smeared into the gaps between surfaces

CRISP



Edges stay sharp, gaps stay empty

Why

Decoder blurs the depth at object edges; blurred depths back-project into the gap.
Not just video VAEs (SVD, Wan2.1): even LiDM's LiDAR-native decoder does it.

WHERE IT FAILS

Inherited decoders fail at depth edges

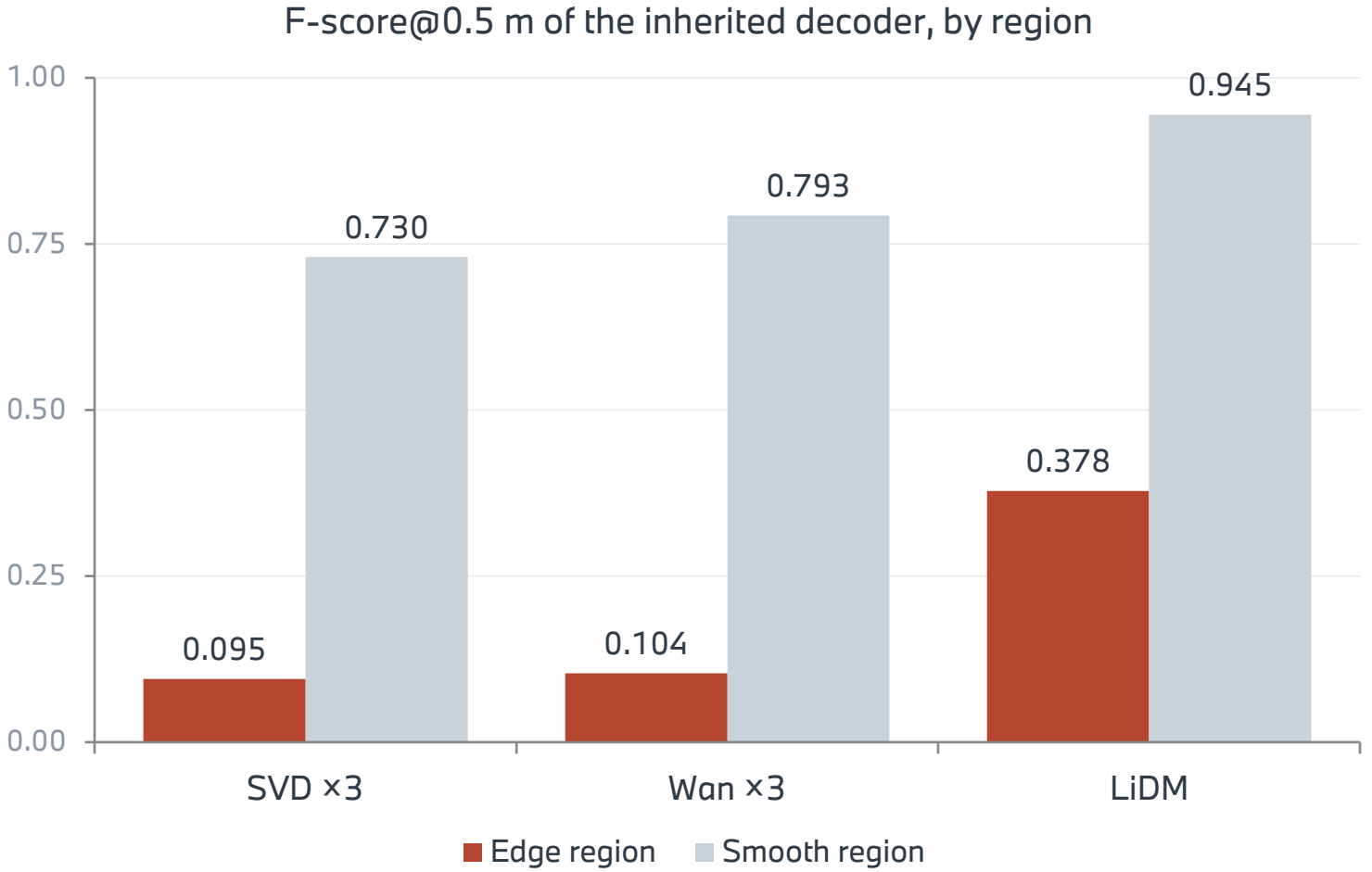
−78%

edge F-score relative to smooth surfaces, inherited decoders (mean of 3 frozen backbones)

How we split the cloud

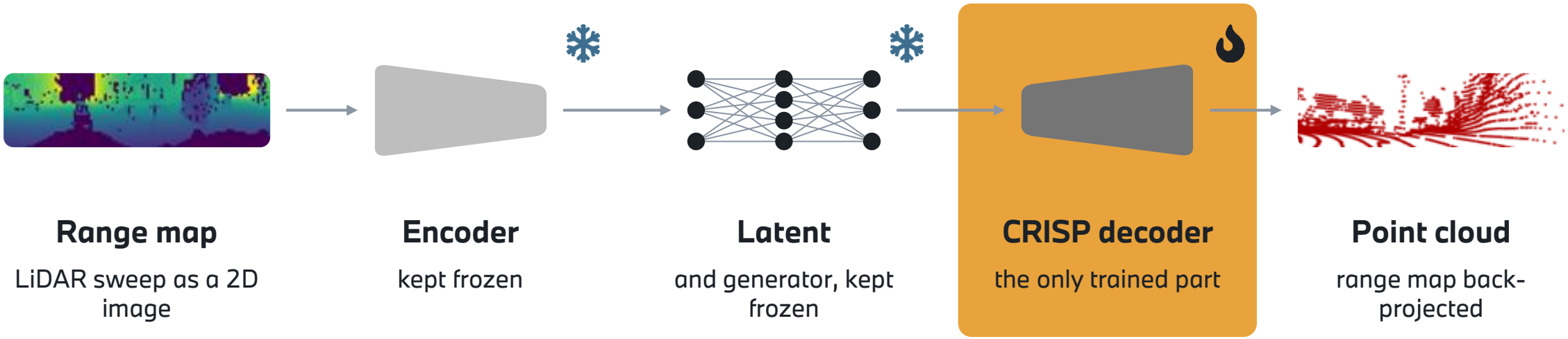
- **Edge** valid pixels at local depth discontinuities
- **Smooth** all other valid pixels

Each region is scored separately



THE IDEA

Swap only the decoder, keep the latent



Range map
LiDAR sweep as a 2D image

Encoder
kept frozen

Latent
and generator, kept frozen

CRISP decoder
the only trained part

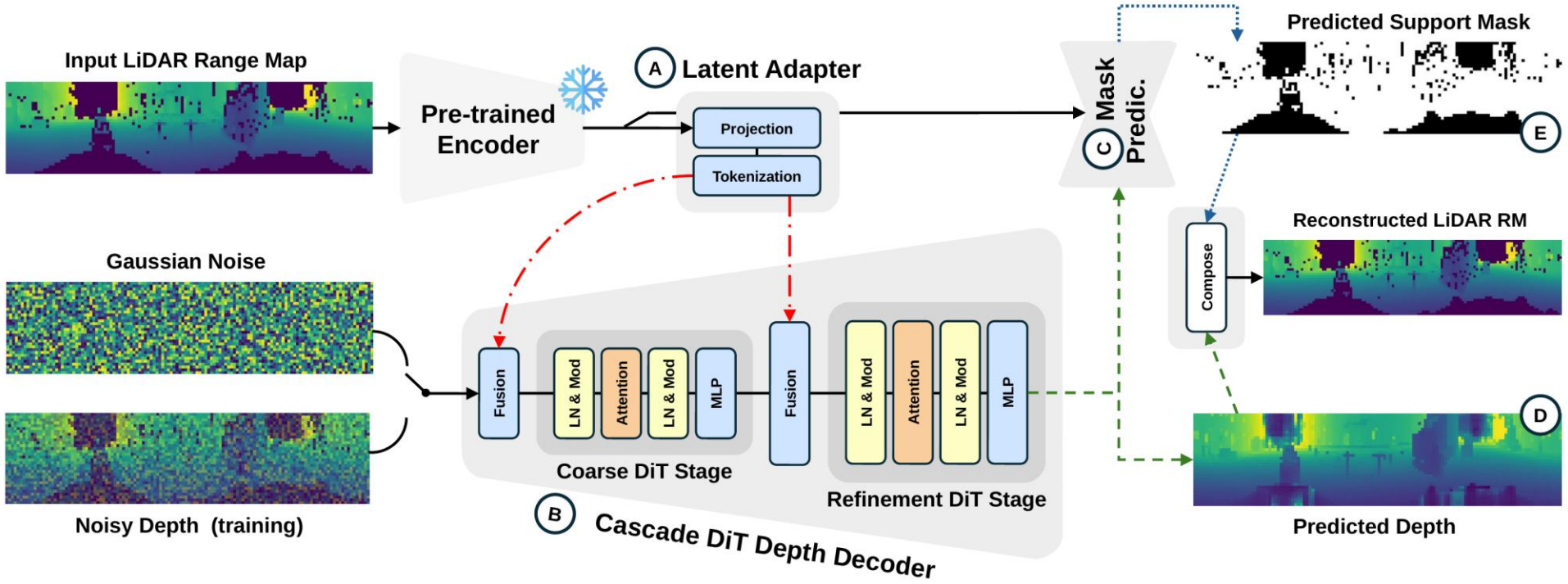
Point cloud
range map back-projected

One decoder design, plugged into four settings

- ✓ **SVD** — video VAE, nuScenes
- ✓ **Wan2.1** — video VAE, nuScenes
- ✓ **LiDM** — LiDAR-native, KITTI-360 / SemanticKITTI
- ✓ **LiDM world model** — zero-shot, generator untouched

METHOD

A pixel-space diffusion decoder, guided by the latent



A **Latent adapter**

Any latent becomes one token sequence

B **Pixel-space DiT**

5 denoising steps; latent fused at input and halfway

C **Support mask**

Predicts which beams return; blanks out the rest

D + E **Objectives**

Edge-weighted depth loss + boundary-aware mask loss

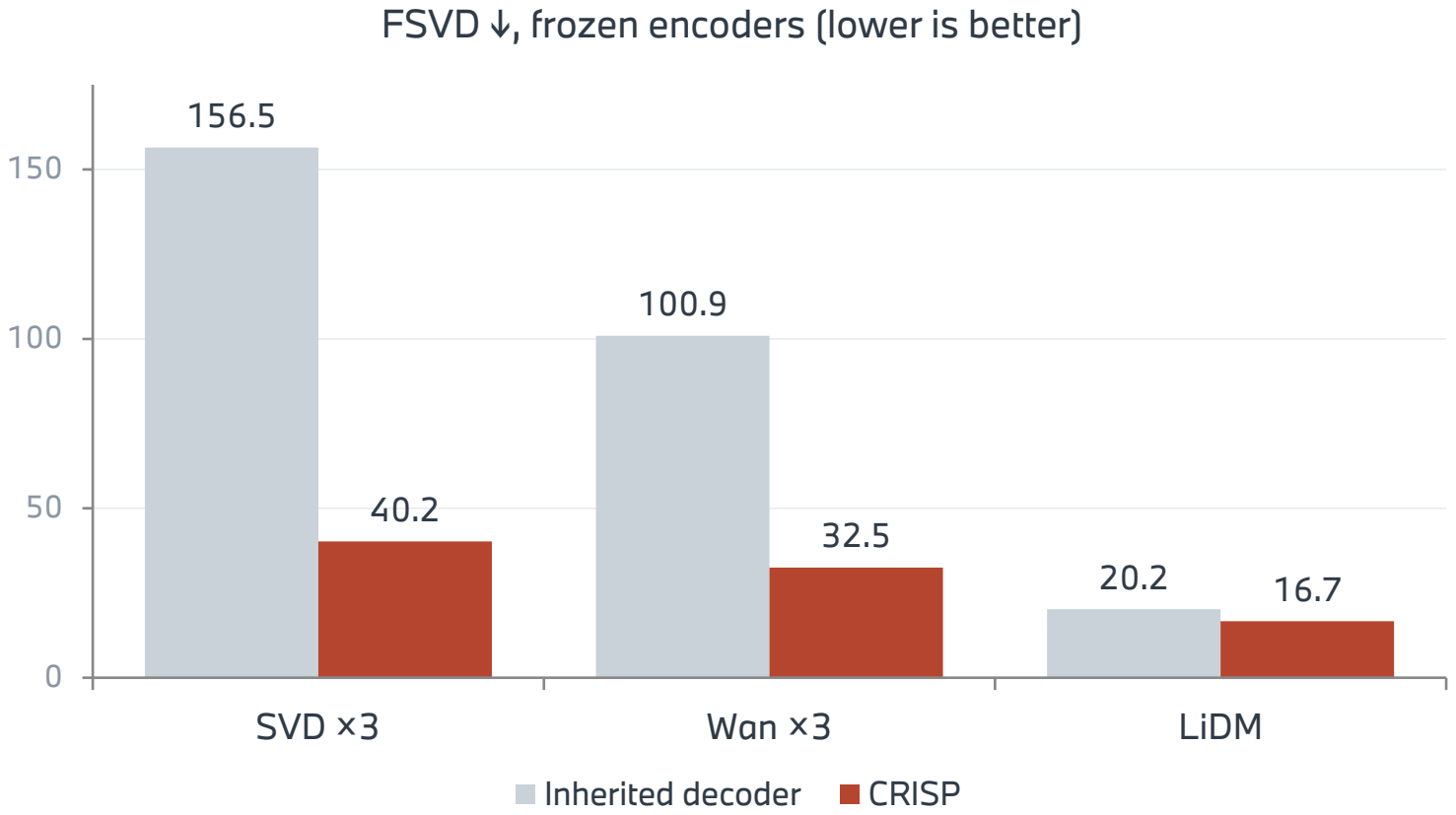
Largest gains where the modality gap is largest

−71% / −74%

FSVD / FPVD, video-VAE backbones (SVD, Wan2.1)

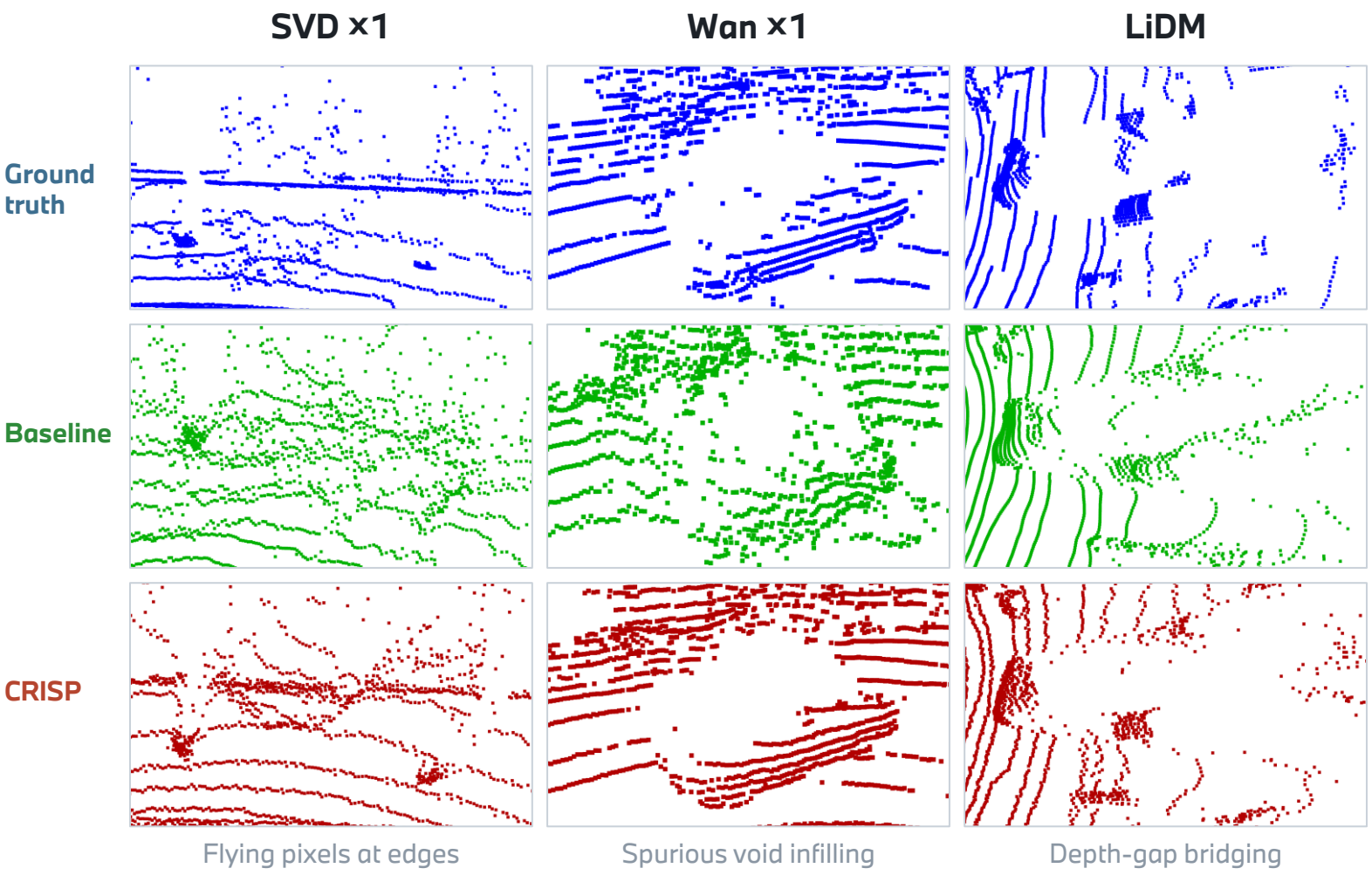
−71%

FRID, LiDAR-native LiDM (2.24 → 0.64)



One exception: LiDM FPVD 16.15 → 17.03. CRISP has ≈516M parameters (8–60× the decoders it replaces), yet LiDM gets the most capacity and gains the least.

The gains sit exactly at the edges



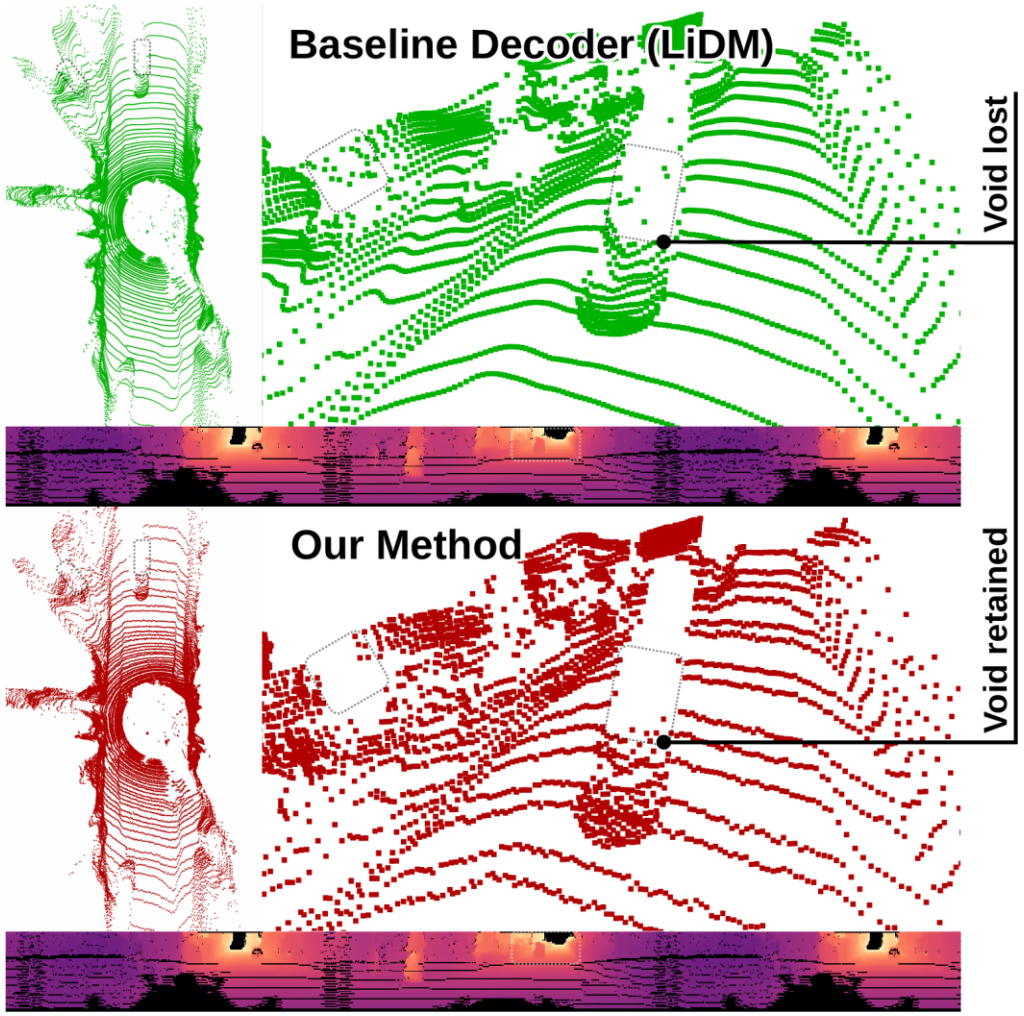
+107%

edge F-score on average over 5 settings, versus +9% for the whole cloud

−23%

smooth-region Chamfer distance on average: no trade-off on flat surfaces

Zero-shot inside a frozen LiDM world model



The generator and its sampled latents stay fixed; only the decoder changes. The latent dynamics were tuned to the old decoder, and CRISP still improves.

−15.5%	−4.0%	−17.5%
FSVD	FRID	edge Chamfer distance

Honest caveat

Edge F-score falls by 20% (0.0486 → 0.0389) while edge Chamfer distance improves: CRISP under-predicts points rather than inventing geometry.

TAKEAWAYS

Flying pixels are a decoder problem

1. Inherited convolutional decoders fail at depth edges, not across the whole cloud

2. Replacing only the decoder with CRISP: −71% / −74% FSVD / FPVD on video VAEs, −71% FRID on LiDM

3. It works zero-shot in a frozen world model: −15.5% FSVD, no retraining

andrea25512.github.io/CRISP

Paper, videos; code and checkpoints will be released on the project page

Come see us: Paris Poster Session 6 · Fri 11 Dec 2026 · 14:30 CET