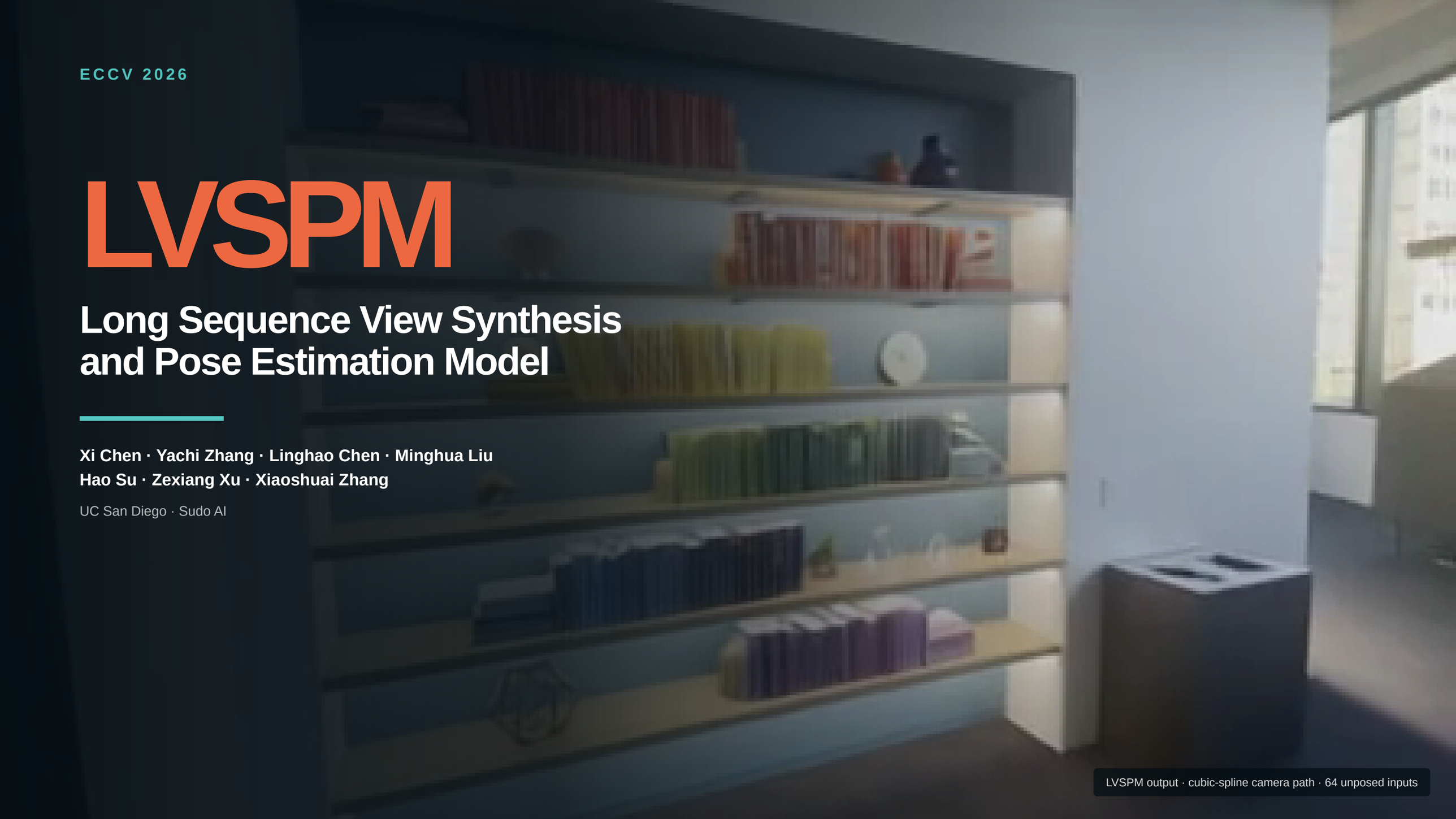
ECCV 2026

LVSPM

Long Sequence View Synthesis and Pose Estimation Model

Xi Chen · Yachi Zhang · Linghao Chen · Minghua Liu
Hao Su · Zexiang Xu · Xiaoshuai Zhang

UC San Diego · Sudo AI



LVSPM output · cubic-spline camera path · 64 unposed inputs

RESULT FIRST

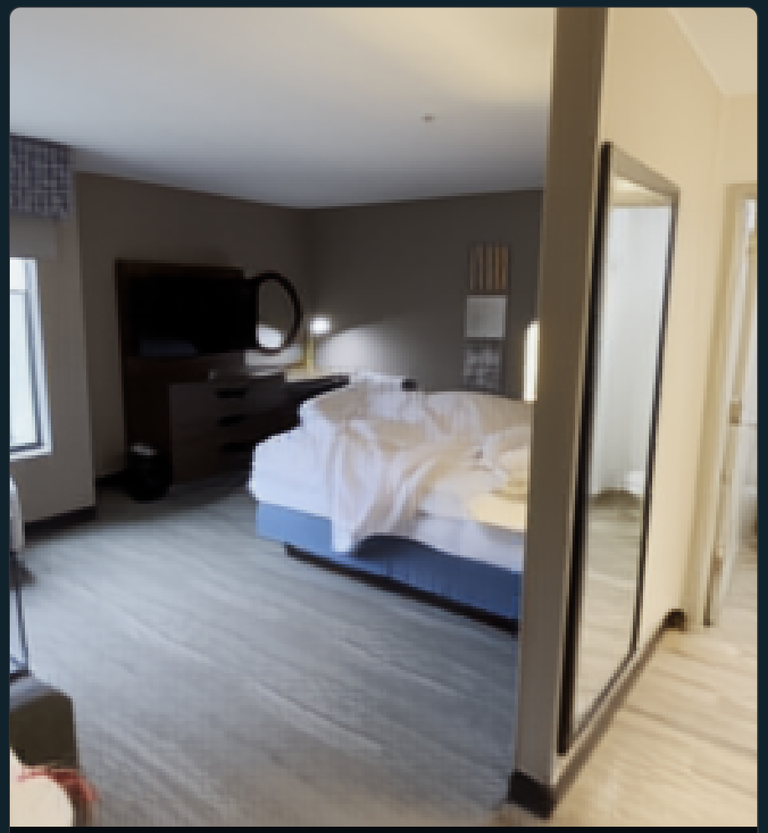
64 unposed images → poses + continuous novel views

One feed-forward scene model.

No input poses · no scene optimization · complete spline path

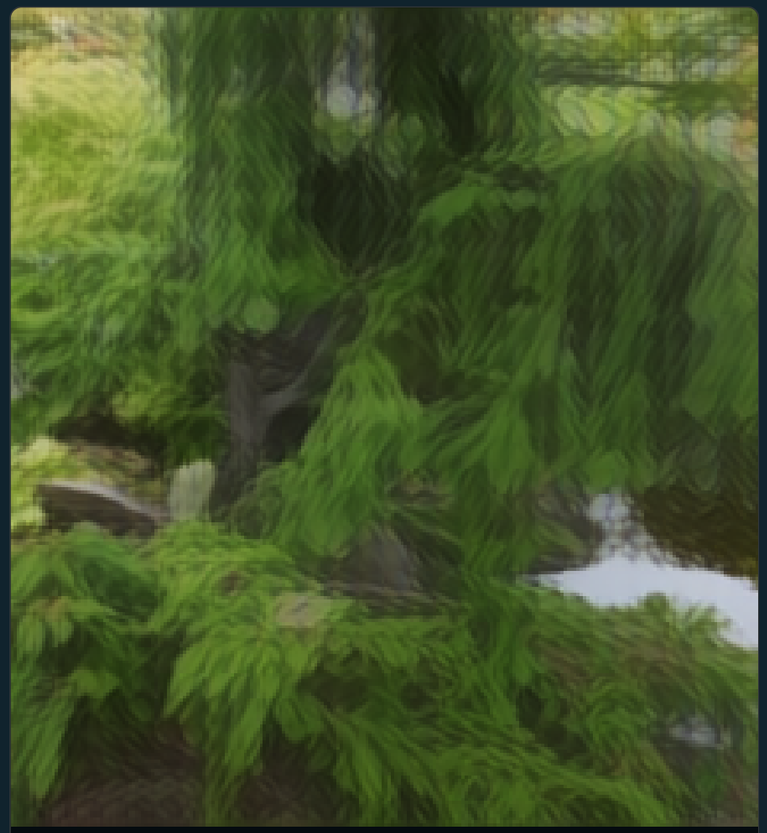
LVSPM OUTPUTS · DL3DV TEST SET

Start with what the model produces.



Indoor sequence

Large interior



Long camera motion

All clips are rendered by LVSPM along cubic-spline interpolated cameras; examples were selected after rendering the 135-scene canonical split.

MOTIVATION

Long image collections should not need a separate calibration pipeline.

16 → 256 images

unknown camera poses

→

LVSPM

shared sequence model

→

camera poses + RGB

at queried target rays

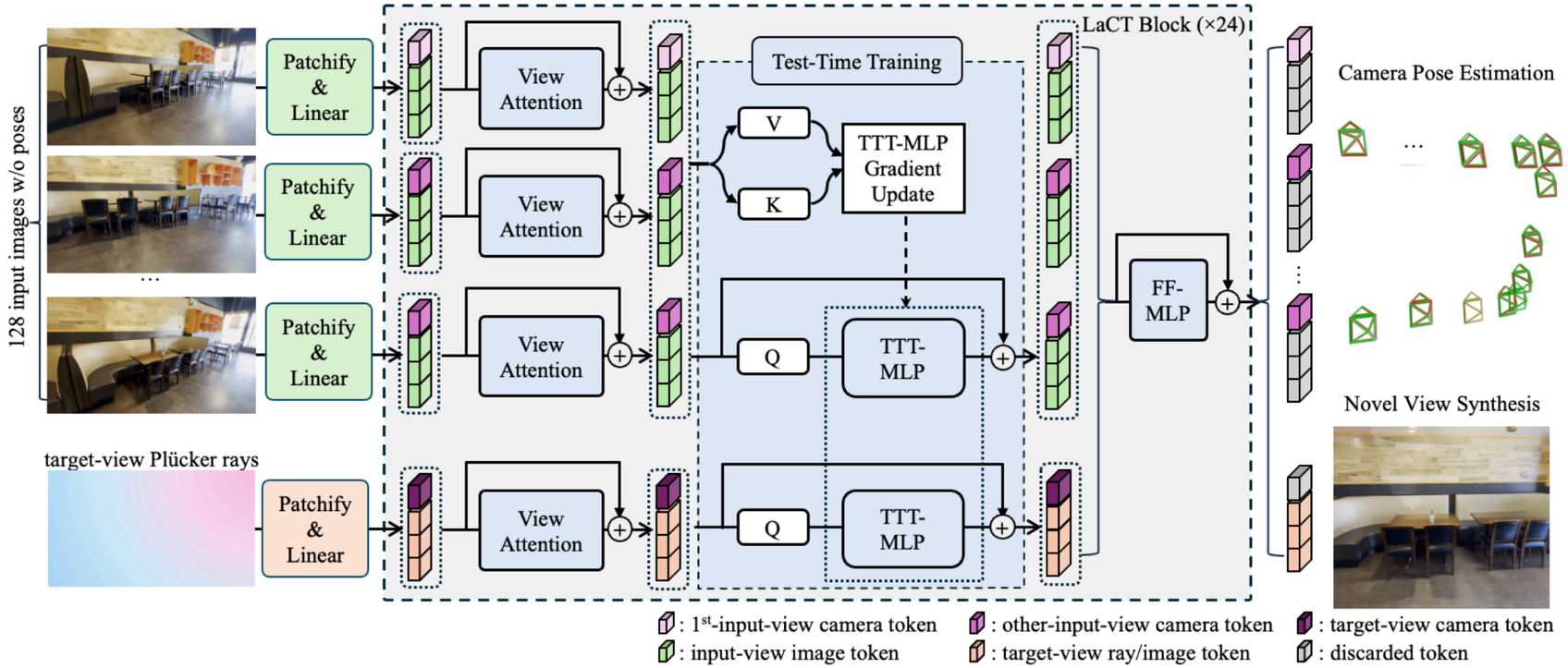
No pose input

No dense 3D labels

No per-scene optimization

CORE METHOD

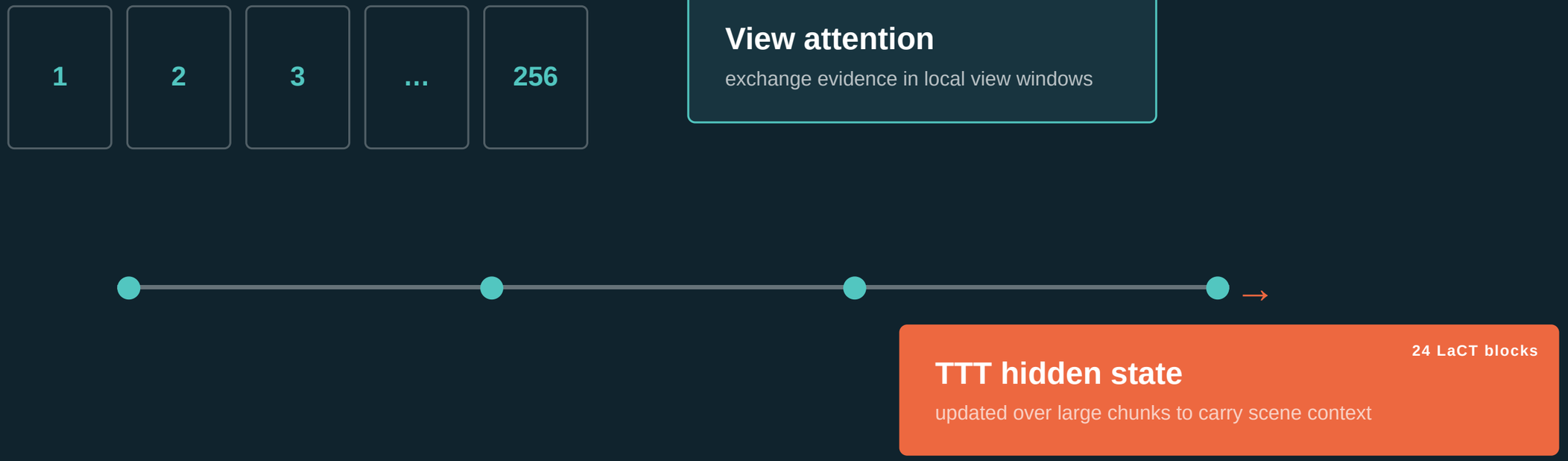
Appearance and cameras share one long-context backbone.



01 Patch + camera tokens	02 LaCT sequence backbone	03 Plücker-ray queries	04 Pose + RGB heads

LONG-CONTEXT DESIGN

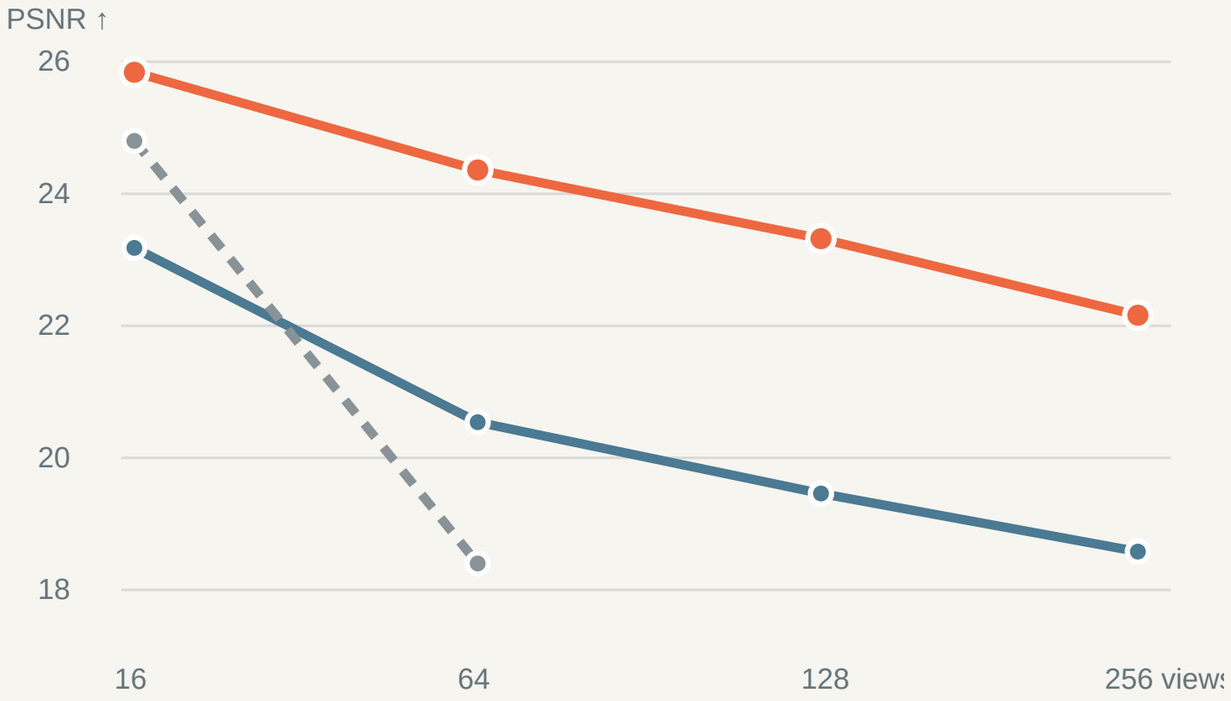
Local attention resolves views. A learned state carries the scene.



The model is trained with sequences up to 128 views and evaluated directly at 256 views.

NOVEL-VIEW SYNTHESIS · 135 DL3DV SCENES

LVSPM stays ahead as the observed scene grows.



+3.59 dB over the pose-free baseline at 256 views

PSNR ↑	16	64	128	256
LVSPM	25.84	24.36	23.32	22.15
AnySplat	23.19	20.54	19.45	18.56
DepthSplat†	24.81	18.40	OOM	OOM

† receives ground-truth camera poses. AnySplat and LVSPM are pose-free.

Canonical-only reduction: exactly 135 tracked scenes per reported setting. Values are recomputed from saved per-scene outputs.

NOVEL-VIEW SYNTHESIS · COMPLETE METRIC REPORT

The advantage is not PSNR-only.

PSNR ↑

	16	64	128	256
LVSPM	**25.84**	**24.36**	**23.32**	**22.15**
AnySplat	23.19	20.54	19.45	18.56
DepthSplat†	24.81	18.40	OOM	OOM

SSIM ↑

	16	64	128	256
LVSPM	**.815**	**.757**	**.714**	**.661**
AnySplat	.759	.647	.597	.548
DepthSplat†	.849	.618	OOM	OOM

LPIPS ↓

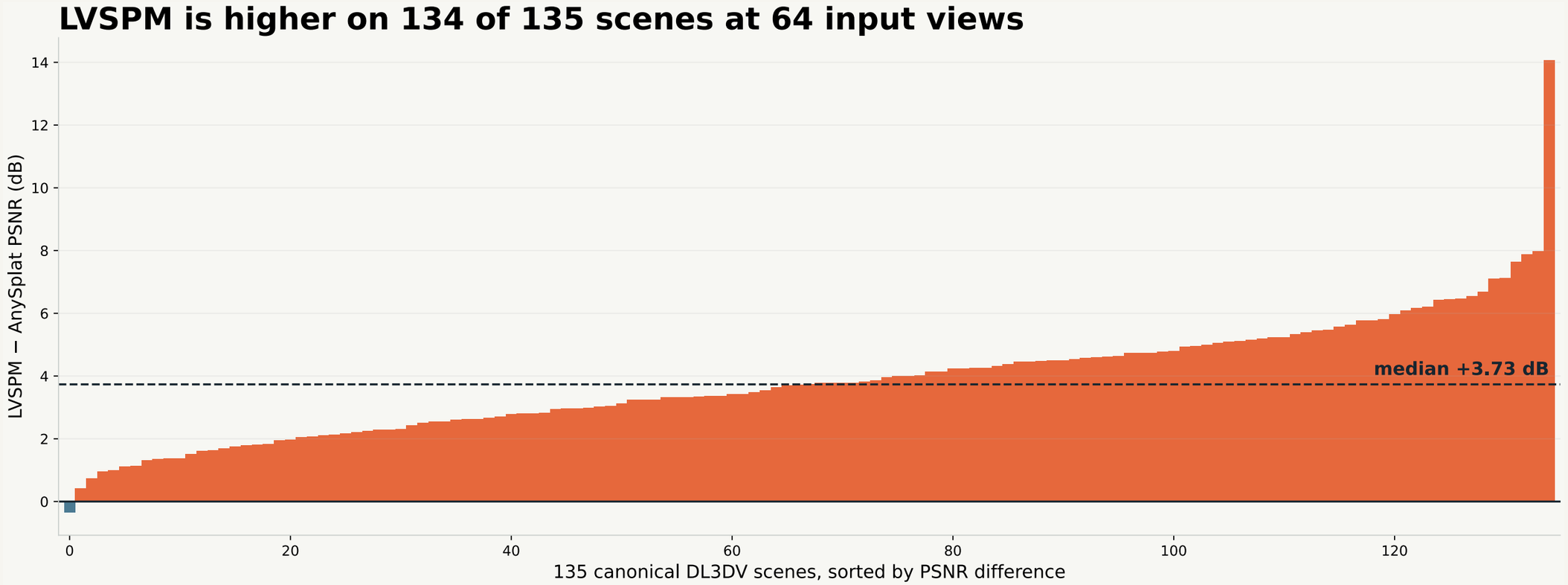
	16	64	128	256
LVSPM	**.150**	**.180**	**.206**	**.244**
AnySplat	.157	.247	.293	.338
DepthSplat†	.133	.362	OOM	OOM

At 256 views **+3.59 dB PSNR** **+.113 SSIM** **−.094 LPIPS** LVSPM vs. AnySplat

Scene means over the same 135-scene canonical split. † DepthSplat receives ground-truth poses; missing cells are measured OOM, not omitted runs.

PER-SCENE EVIDENCE · 64 INPUT VIEWS

The average is backed by nearly every scene.



134 / 135
higher PSNR

129 / 135
higher SSIM

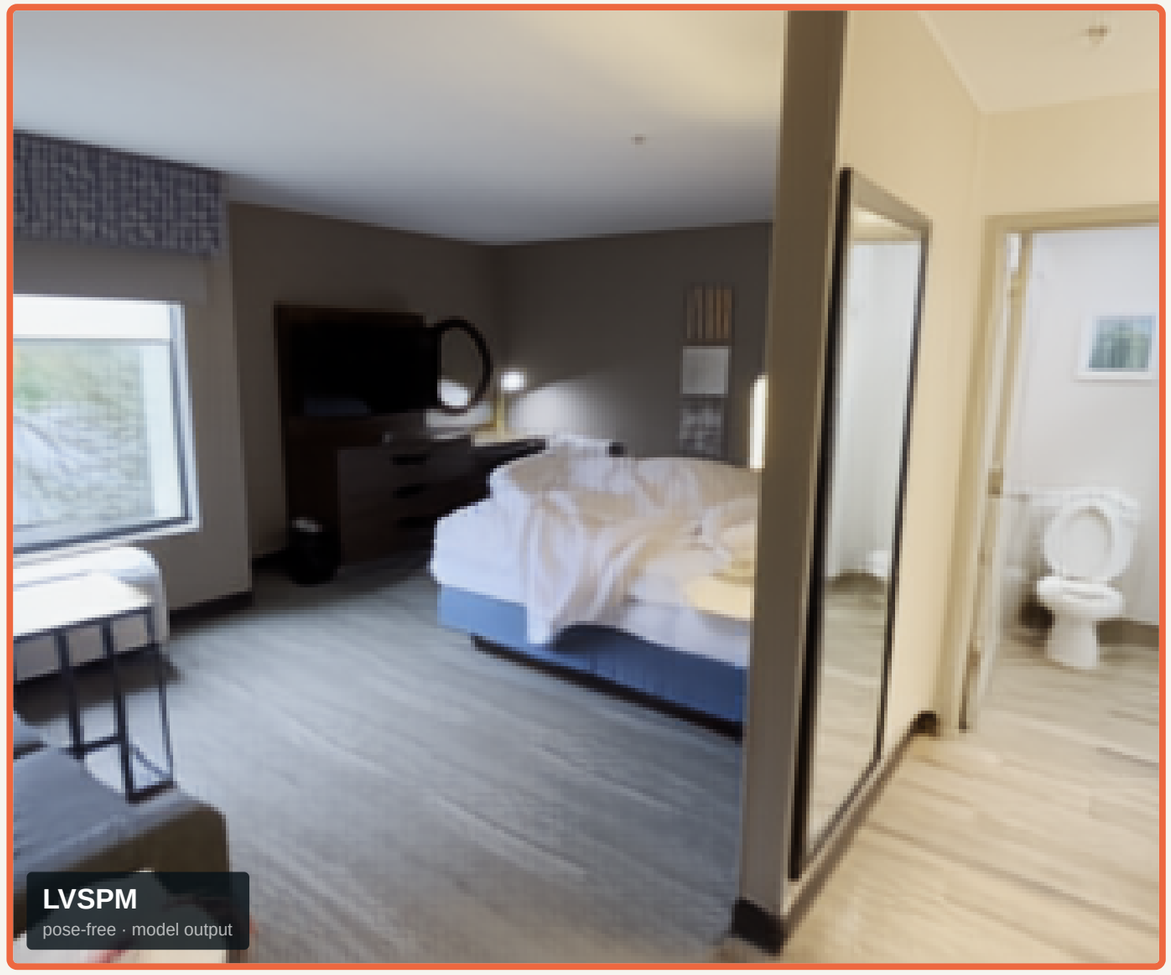
124 / 135
lower LPIPS

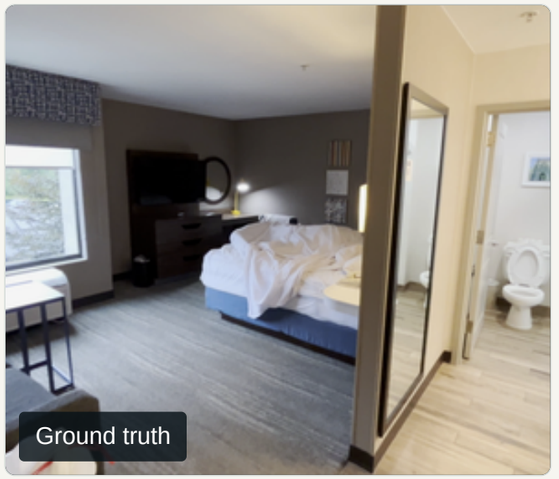
+3.73 dB
median PSNR difference

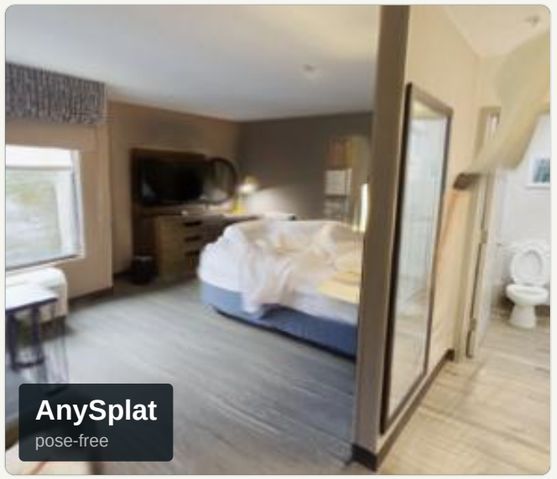
Direct scene-ID join between the complete LVSPM spline run and AnySplat's preserved 64-view outputs; no unmatched or extra scenes enter the counts.

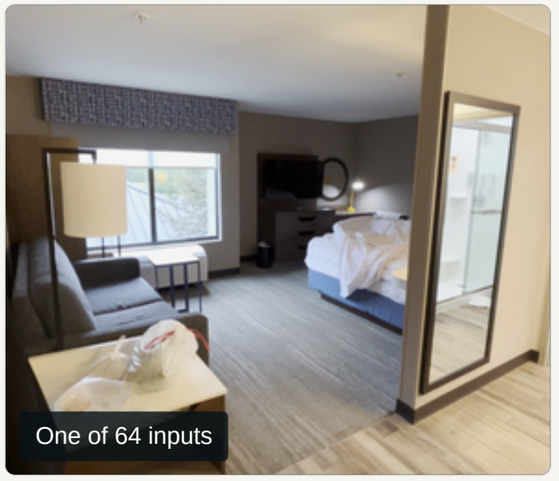
REAL TARGET-ALIGNED COMPARISON · STRONG CASE

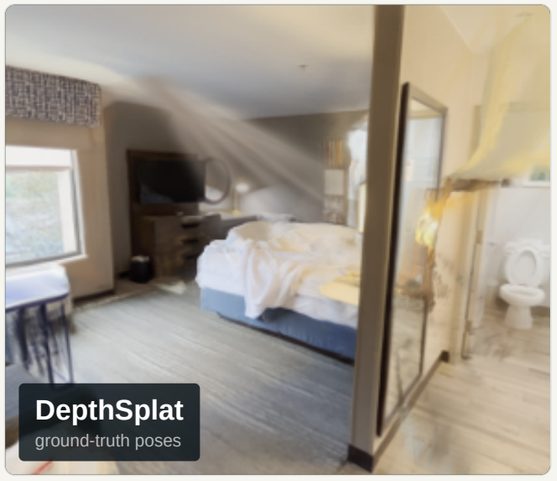
LVSPM first. Baselines on the same held-out target.







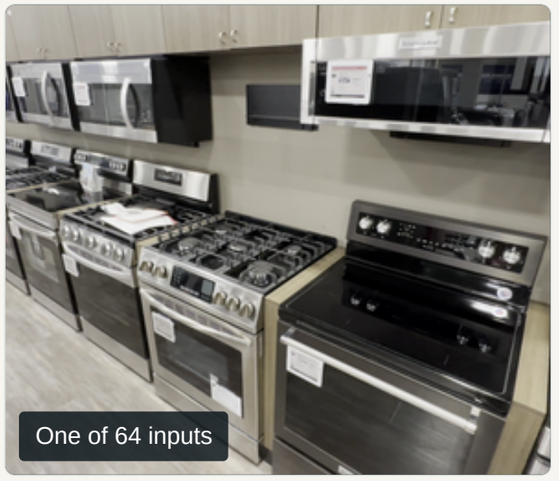




Scene 389a… · LVSPM scene PSNR 27.51 · held-out target verified by pixel-matching saved ground truth.

REAL TARGET-ALIGNED COMPARISON · MEDIAN CASE

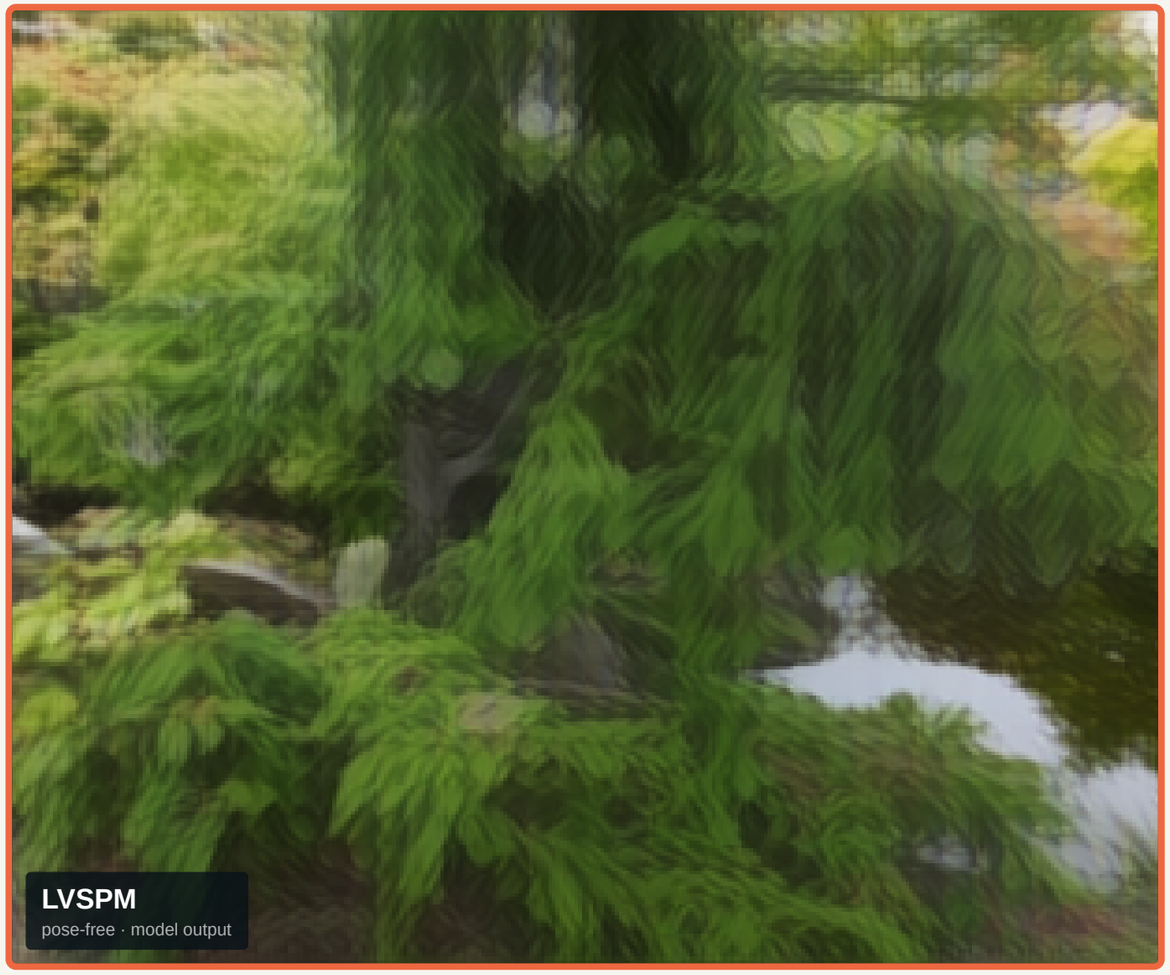
The comparison is not only a best-case gallery.

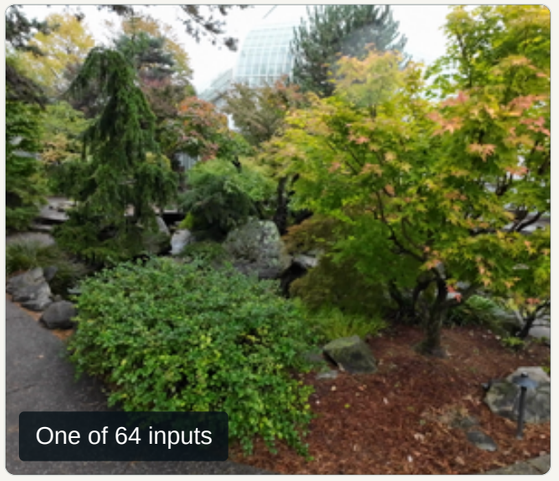


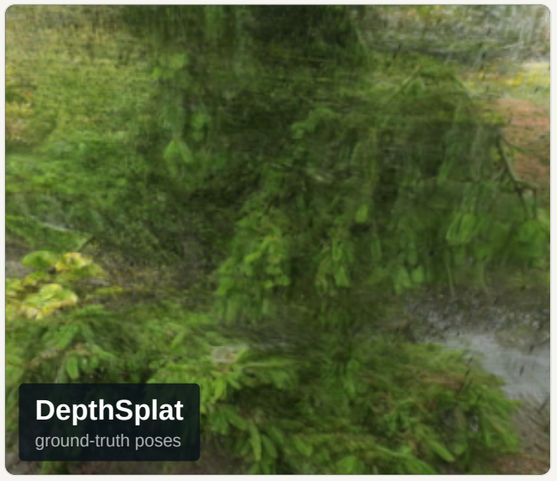
Scene 444d… · exact median of the 135 scenes by stored LVSPM PSNR · scene PSNR 24.19.

REAL TARGET-ALIGNED COMPARISON · HARDEST LVSPM SCENE

Failure modes remain visible.





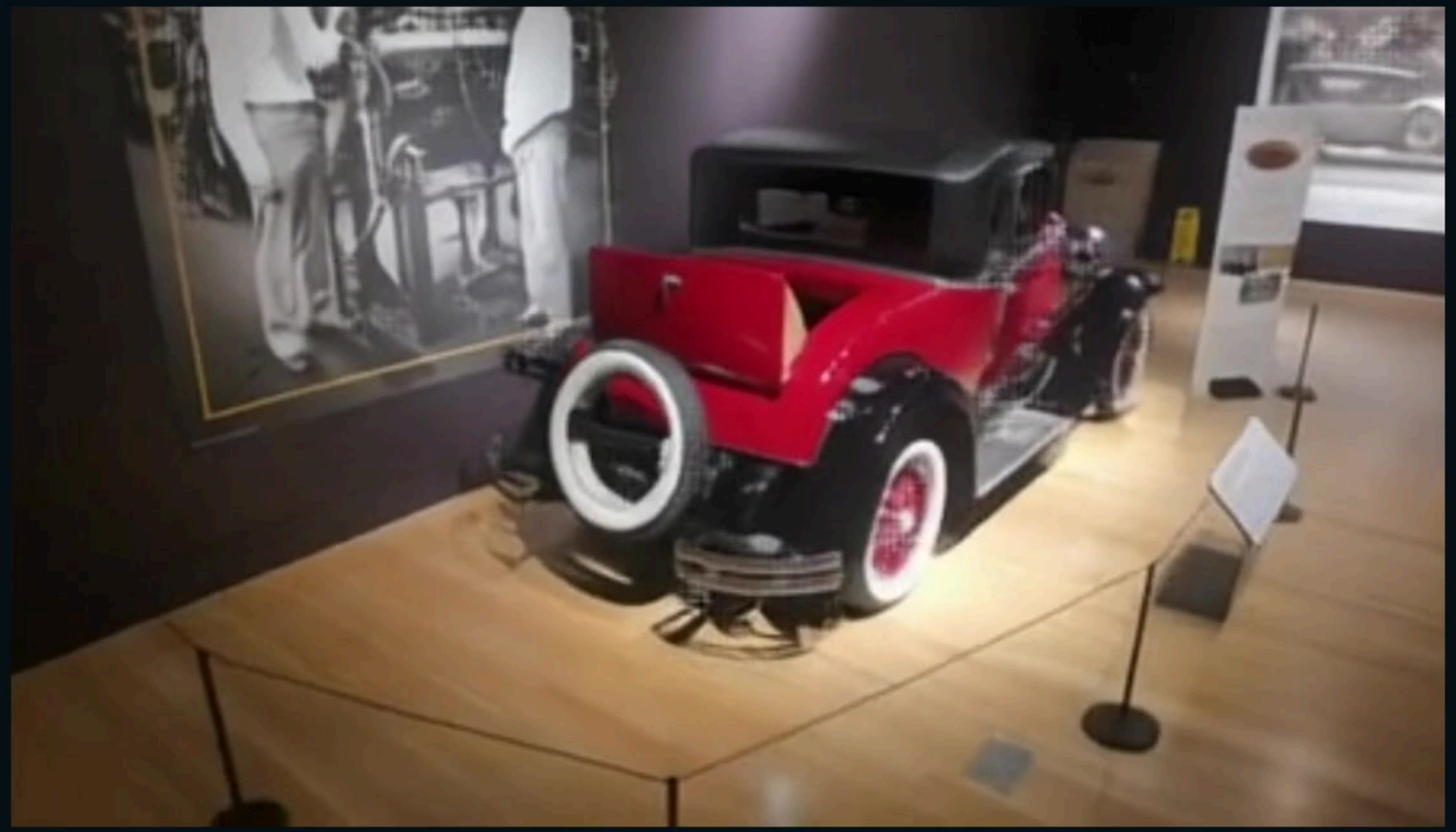


Scene 073f… · lowest LVSPM scene PSNR, 19.32. Fine structure and appearance are visibly degraded.

SYNCHRONIZED NVS VIDEO · HISTORIC CAR · SCENE DAC9…

One path, three reconstructions.

LVSPM | pose-free context



AnySplat | pose-free context

DepthSplat | GT context poses

64 shared input frames

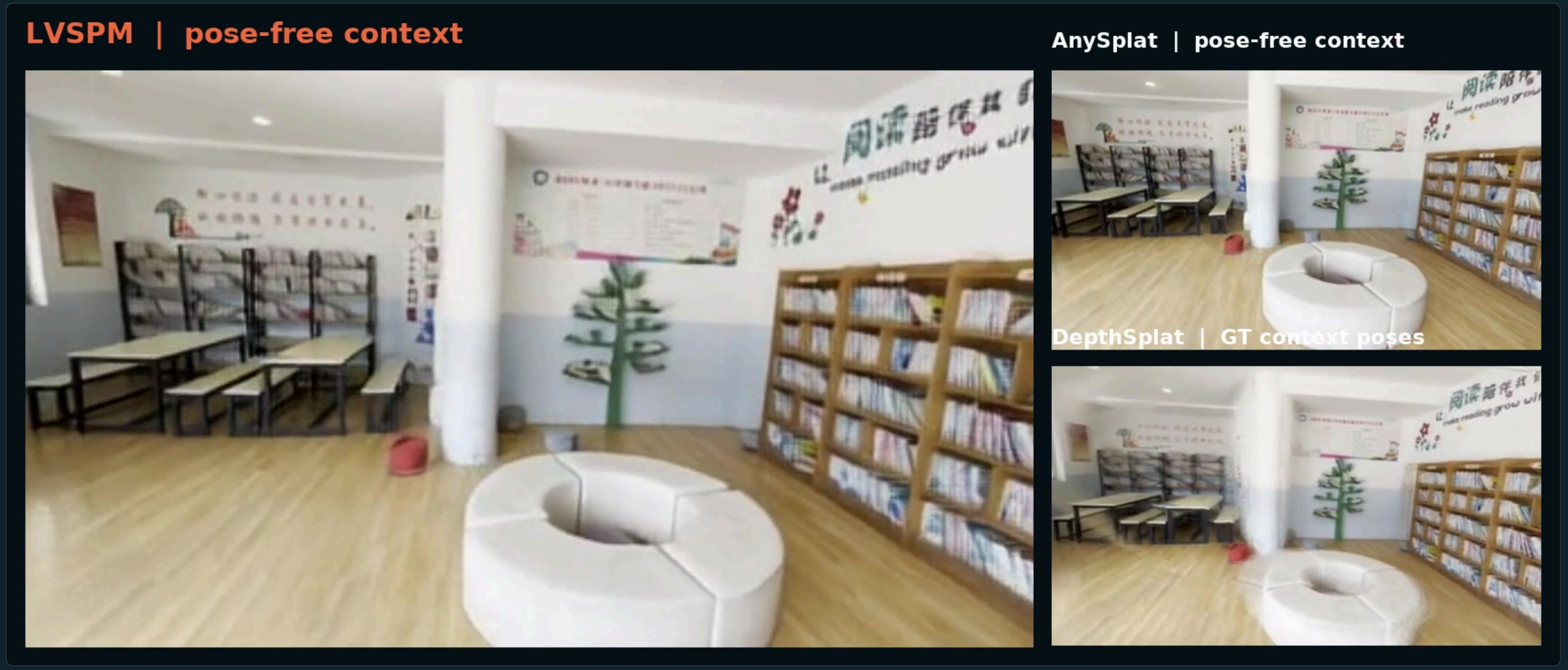
18 sorted held-out cameras

same physical spline path

All three render the exact spline through the same sorted held-out cameras. AnySplat's path is mapped into its reconstructed Gaussian frame by one Sim(3) fitted from context cameras; DepthSplat receives ground-truth context poses.

SYNCHRONIZED NVS VIDEO · LIBRARY · SCENE 2CBF…

The same test in a clean, structured interior.



64 shared input frames

12 sorted held-out cameras

full forward-and-back path

The camera motion is synchronized by normalized spline position, not by cherry-picked frames. The complete round trip is shown, with method and pose-input labels embedded in the video.

POSE ESTIMATION · 53 DL3DV SCENES

Every baseline, every canonical view count.

Relative-pose AUC@3 ↑	16	32	64	128	256
LVSPM	**88.4**	**92.3**	**92.0**	**90.0**	**90.1**
Depth Anything 3	86.2	86.2	74.6	84.5	88.8
VGGT	74.3	89.4	78.4	87.7	85.0
FLARE	43.1	59.6	80.0	73.3	74.6
TTT3R	58.1	62.8	54.8	37.8	30.4
Cut3R	35.9	55.4	47.9	27.2	17.3
Fast3R	17.5	35.3	24.5	17.3	9.7

53 / 53 scenes for every cell

+1.3 points over the next method at 256 views

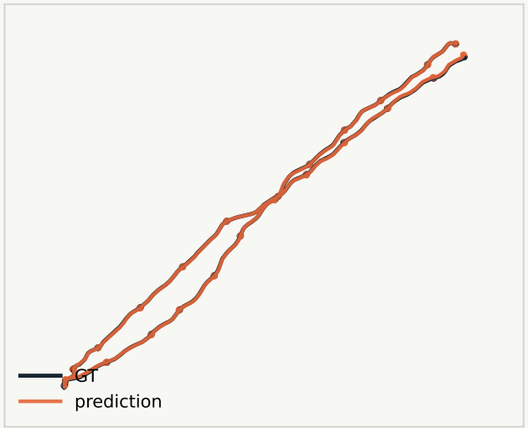
90+ from 32 through 256 views

Re-aggregated from preserved full output trees; the deck does not substitute manuscript values where stored outputs differ.

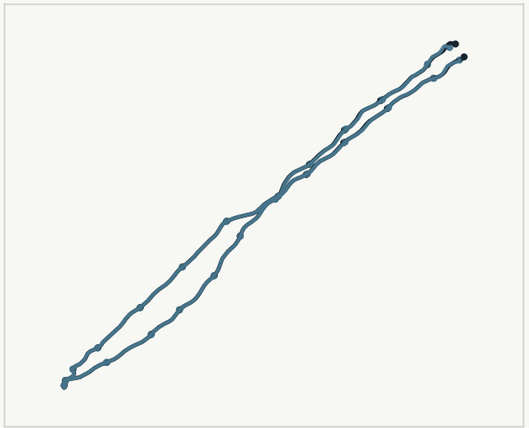
SAVED PREDICTIONS · REPRESENTATIVE 256-VIEW SCENE

Every baseline trajectory, revealed in sequence.

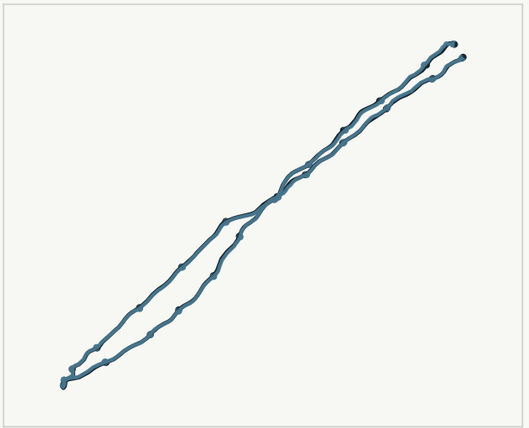
LVSPM
AUC@3 96.9



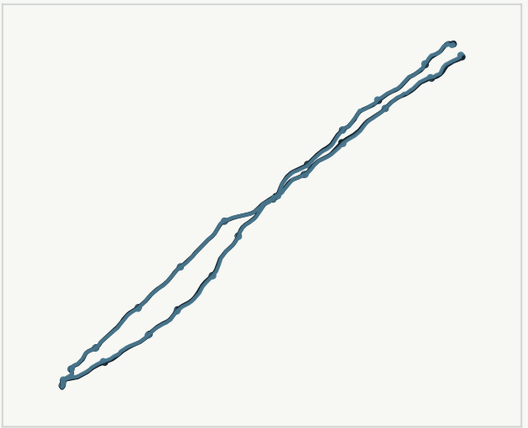
Depth Anything 3
AUC@3 85.2



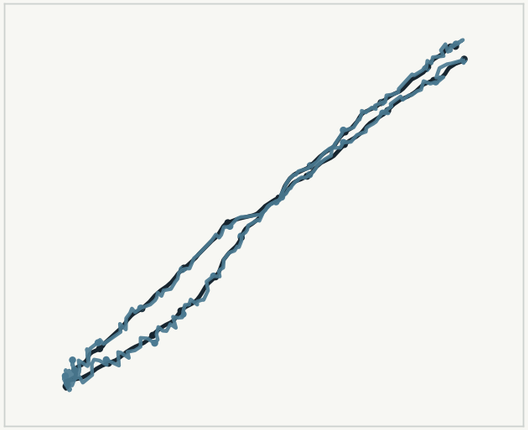
VGGT
AUC@3 86.3



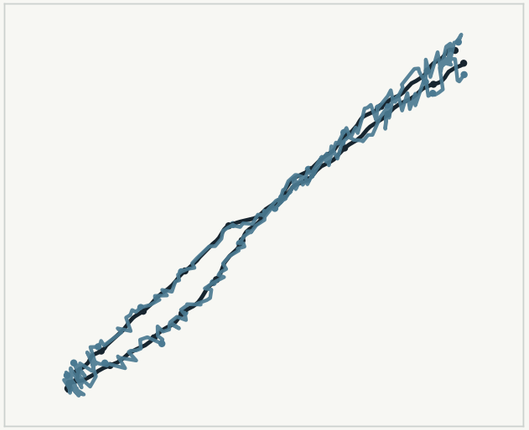
FLARE
AUC@3 78.9



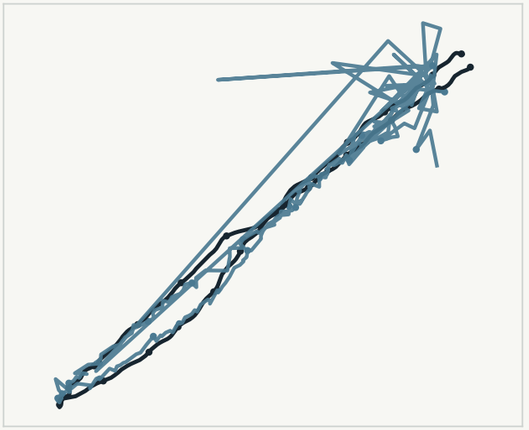
TTT3R
AUC@3 43.0



Cut3R
AUC@3 21.4



Fast3R
AUC@3 10.7



Saved predictions
256 input views

Camera centers only
Similarity-aligned for display

Scene 1881… was selected by LVSPM's margin to the strongest baseline. Camera centers are similarity-aligned only for visualization; AUC values are the saved scene metrics.

256-VIEW POSE ENCODING

LVSPM is the fastest measured model in this set.

13.0×

faster than the slowest measured baseline

58.8 FPS

novel-view rendering after encoding

Encoder-time fields from the saved 256-view output summaries; LVSPM uses the canonical paper/release measurement.

ABLATION · WHAT ACTUALLY MATTERS

Joint RGB supervision and long-context capacity both matter.

camera prediction ↔ **novel-view prediction**

Rendering supplies dense geometric evidence without depth or point-map labels.

Setting	AUC@3 ↑	PSNR ↑	SSIM ↑	LPIPS ↓
DL3DV · 32 views				
without synthetic pretraining	79.9	20.08	.589	.267
LVSPM	**80.2**	**21.12**	**.638**	**.222**
ASE · 32 views				
without RGB loss	**50.9**	—	—	—
half TTT chunk	67.9	24.98	.702	.263
12 TTT layers	59.3	24.18	.673	.288
LVSPM full	**71.9**	**25.01**	**.701**	**.257**

Remaining limits fine structure · reflections · weak texture · multi-stage training

TAKEAWAYS

Long-sequence, pose-free view synthesis in one model.

01

Unified output
camera poses and target RGB

02

Long context
directly evaluated to 256 views

03

Measured impact
stronger NVS, pose, and speed

All comparisons in this deck come from canonical saved outputs or documented release measurements.