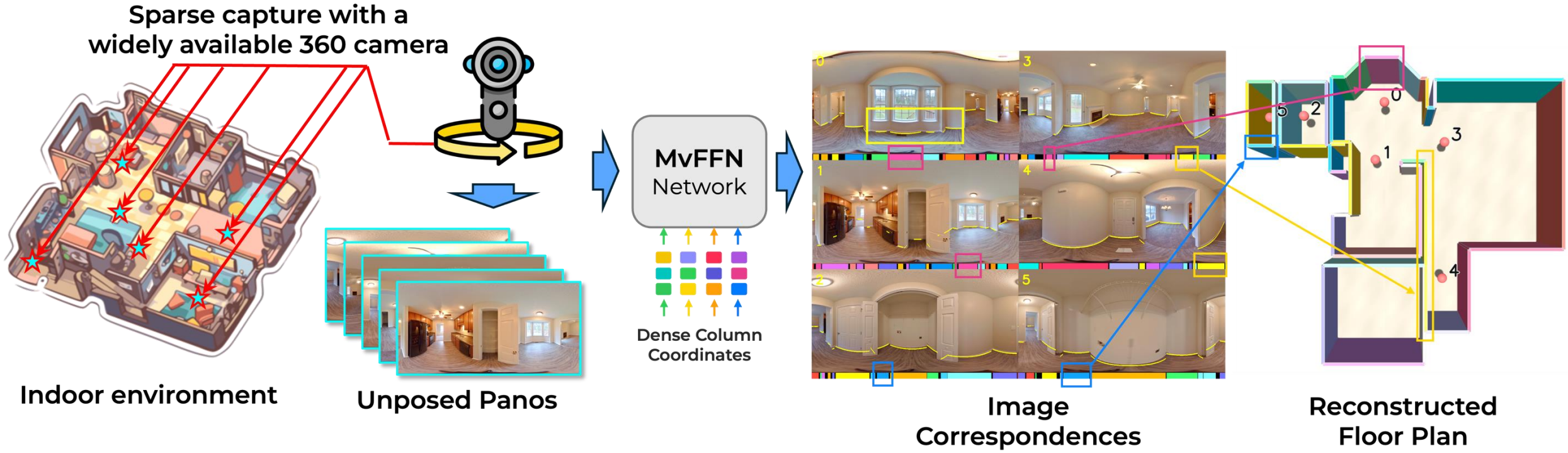
MvFFN

Multi-view Floor-plan Feed-forward Network for Unposed Wide-Baseline Panorama Layout Reconstruction

Yuguang Li, Yichuan Deng, Zixuan Liu, Ivaylo Boyadzhiev, Jacky Wu, Simon Zou, Linda Shapiro, Alex Colburn



Objective: Sparse captures -> precise poses + floor plans

- Predict 3-DoF **camera extrinsic**, **multi-room floor layouts** and **multi-view instance-to-image correspondences**.
  - Currently handles **1 to 8** of panorama **inputs,** can **scale to** ingest **dozens** of inputs.
  - Pairwise visual **overlaps**: **as little as 20%.**
  - Output room layouts may be **closed or** remain **open under occlusion**.
  - **Number of rooms or sequence length are unknown**.
- Objective: Feed-forward, learnable approach.



Unposed Panos

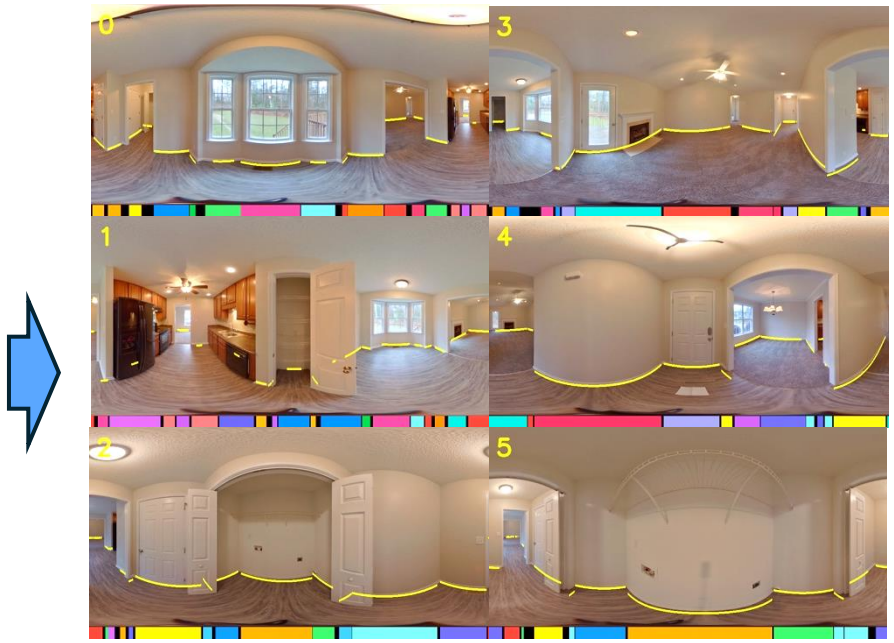
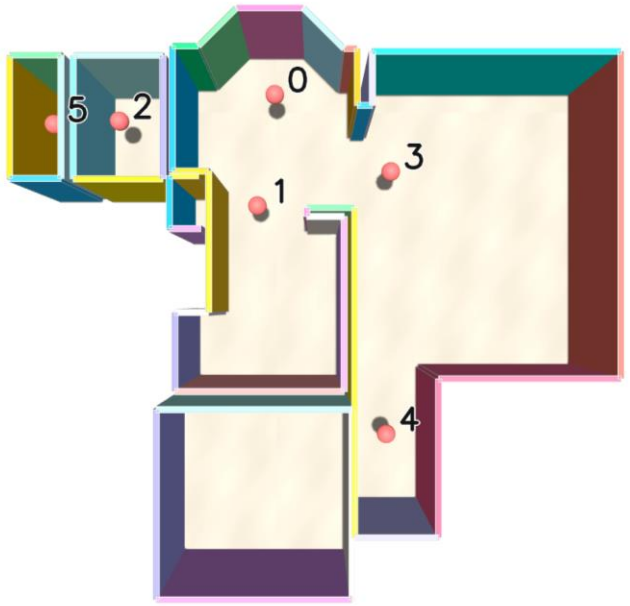


Image Correspondences



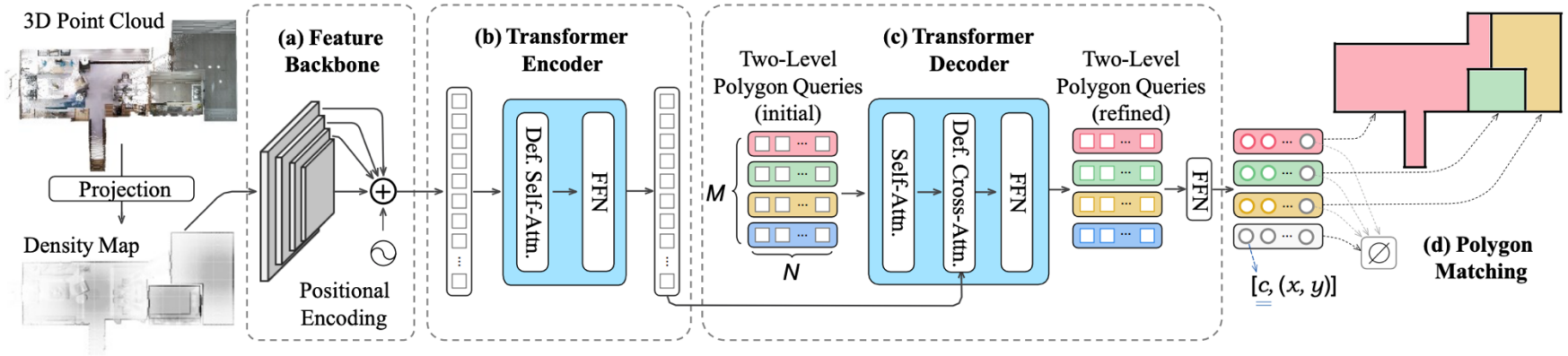
Reconstructed Floor Plan

Starting from BEV is de-facto, but it assumes known camera poses

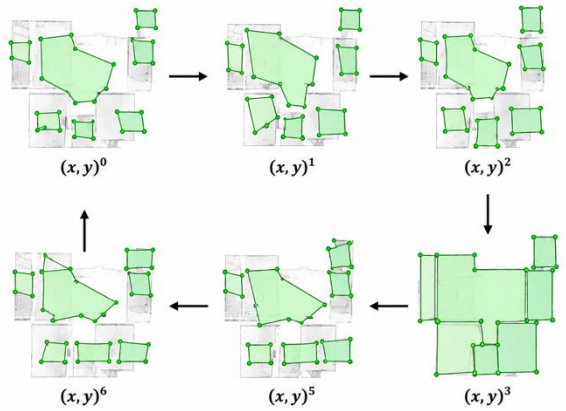
Recover floor-plan key points from rasterized BEV scene maps.

All relies on accurate pose inputs!

RoomFormer (CVPR 2023)



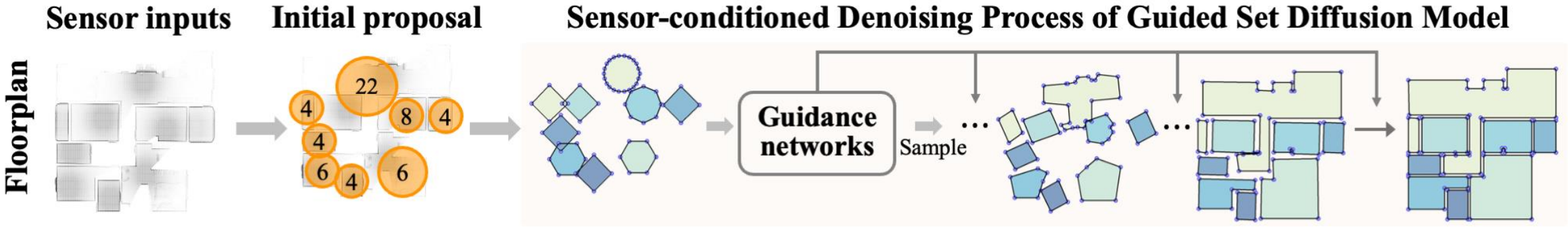
Transformer decoding from BEV density map



Iterative Polygon Refinement

PolyDiffuse (NeurIPS 2024)

Assume room center, number of walls known



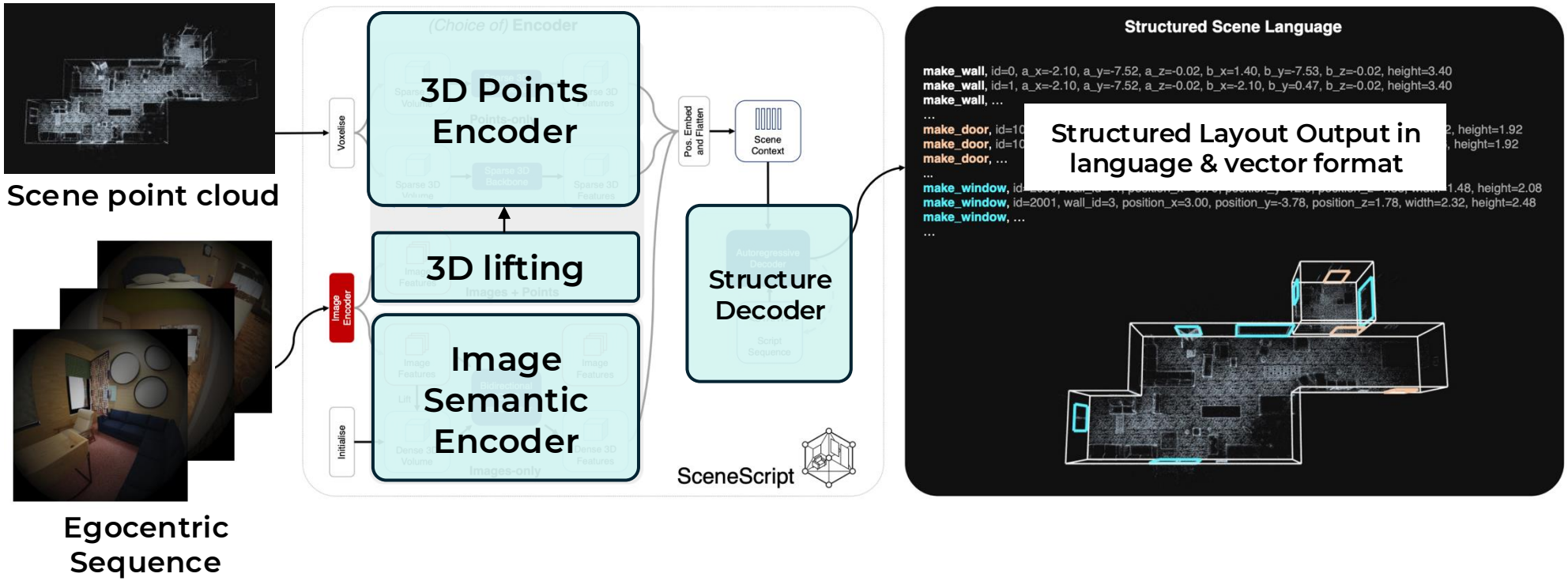
Diffusion-base iterative refinement conditioned by BEV

SceneScript adds image semantics — but still needs posed inputs

SceneScript (CVPR 2024)

- Augment point-cloud geometry with posed multi-view RGB.
- Lift DINOv2 image semantics to improve layout structure understanding.
- Language outputs as a structured scene script.

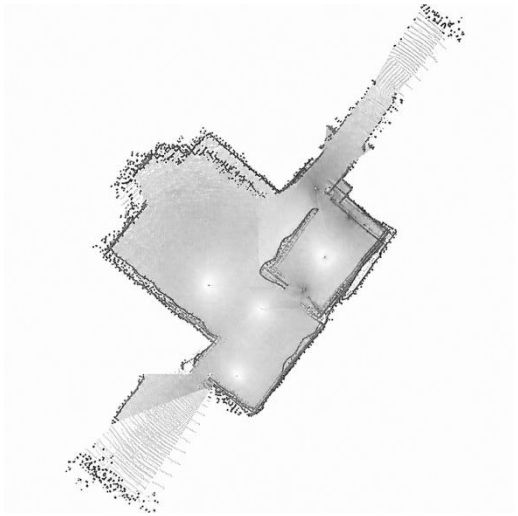
Method	F1 @5cm mean	wall	door	window
SceneCAD '20 [2]	-	0.048	-	-
RoomFormer '23 [52]	0.139	0.159	0.148	0.110
Ours (Point cloud)	0.848	0.930	0.922	0.692
Ours (Lifted features)	**0.903**	**0.943**	**0.959**	**0.806**
Ours (Image-only)	0.661	0.687	0.798	0.497



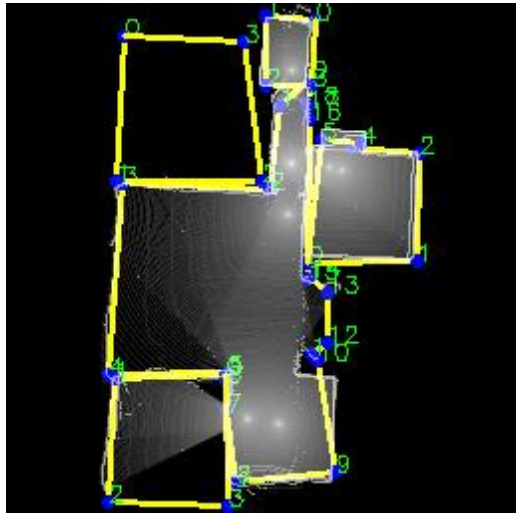
Verdict: Existing pipelines solve layout after pose; this thesis solves them together.

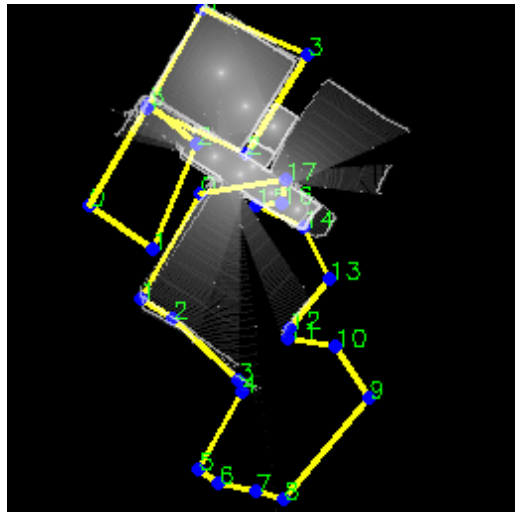
Room instance prediction doesn't work for us

- **Partial layout** observation introduces severe gradient noises.



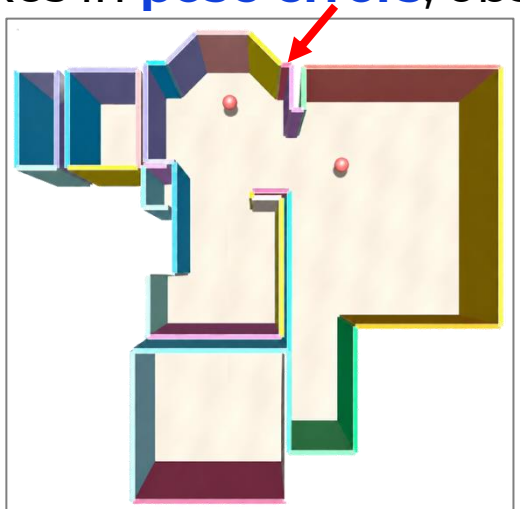
Projected point cloud with noisy depth and poses





RoomFormer prediction with partial observations

- **Fine structural details** are easy to miss in BEV, even seen clearly in panos.
  - BEV map bakes in **pose errors**, obscures structural details.

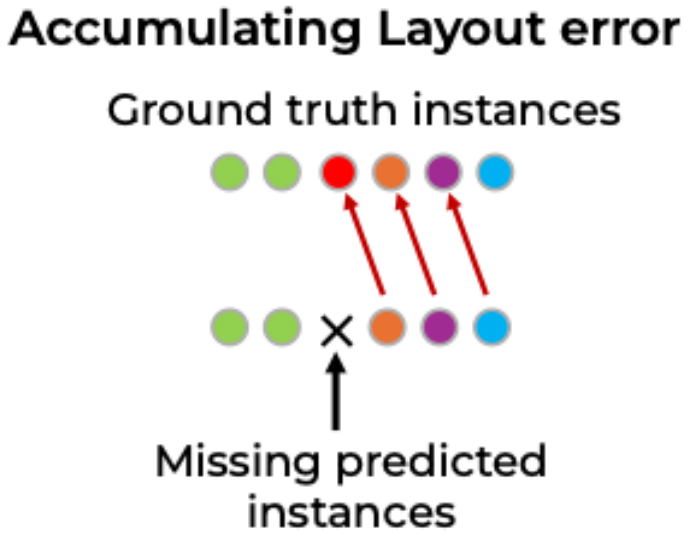




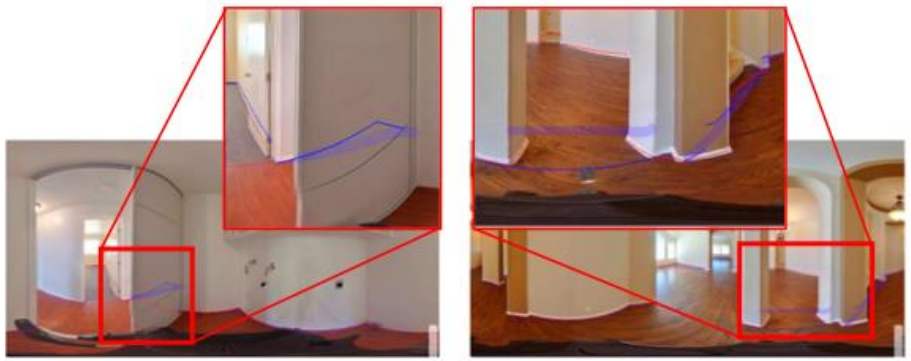
Direct layout prediction doesn't learn complex rooms

- **Sequences matching** with Hungarian produce significant gradient noises.

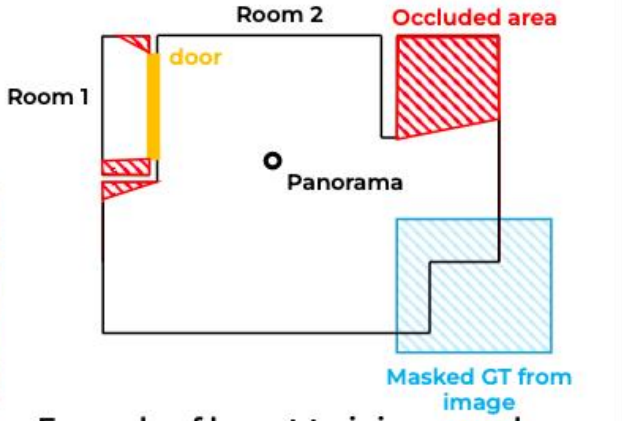




- **Scarce training samples** (~2k training floors). Many **incomplete rooms** from 8-view pano set.
- We **masked out** (~5%) image columns during data cleanup.



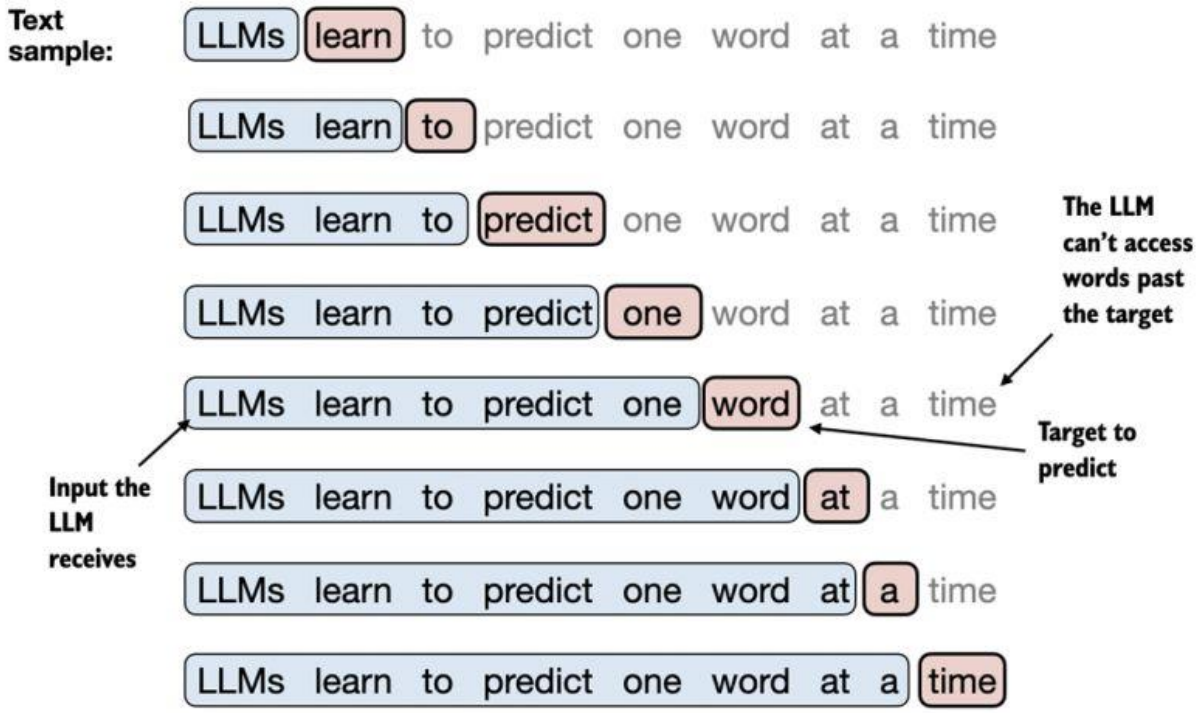
Example Ground Truth data masking for quality control

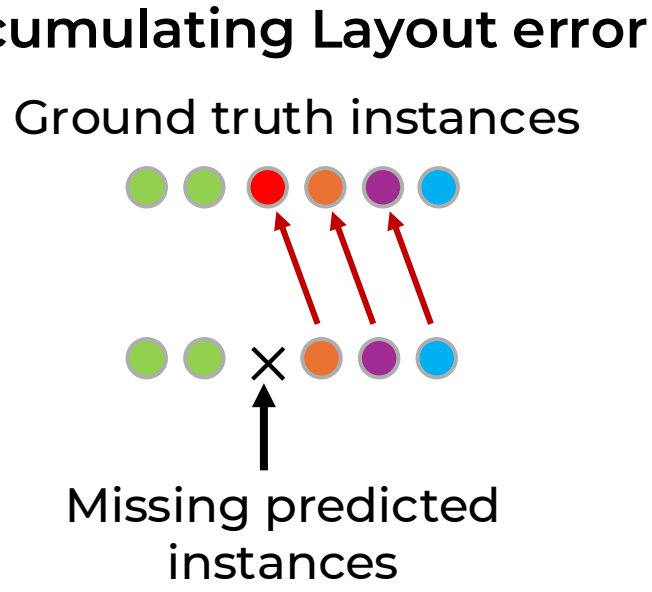


Example of layout training sample with incomplete walls / rooms

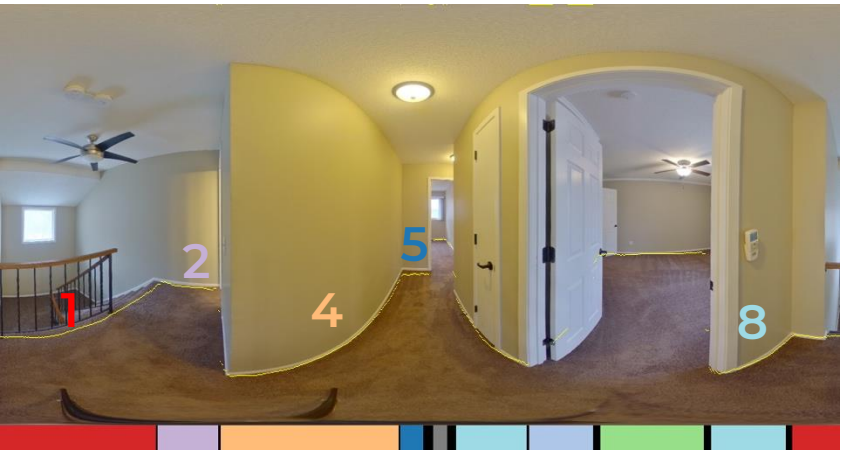
Global layout prediction is the wrong learning signal.

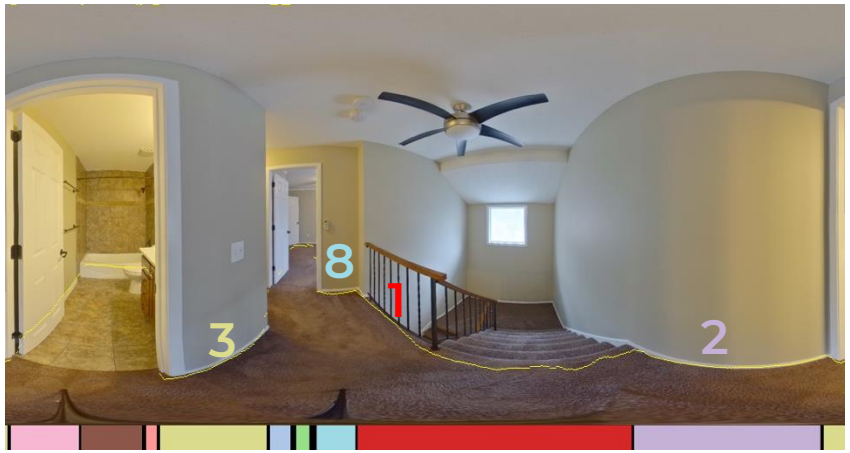
- Sparse global sequence matching is brittle.
- Missing walls make Hungarian matching ambiguous.
- Need a dense local objective, not sparse global sequence prediction.

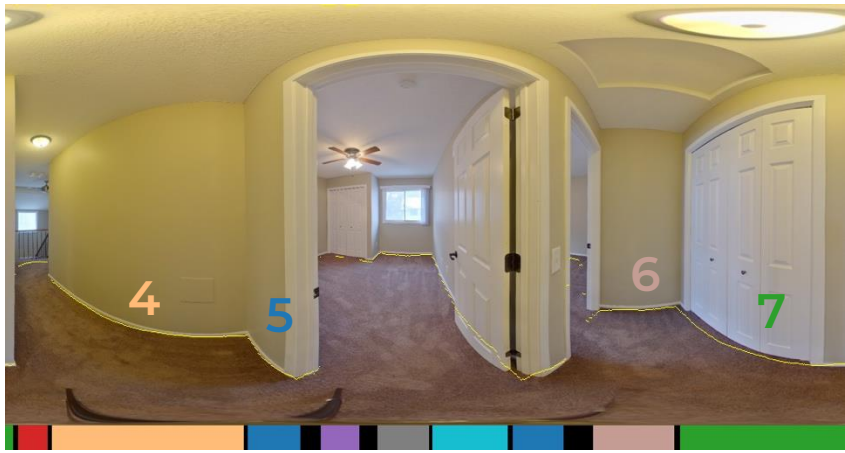




New formulation: Multi-view Wall Tracing



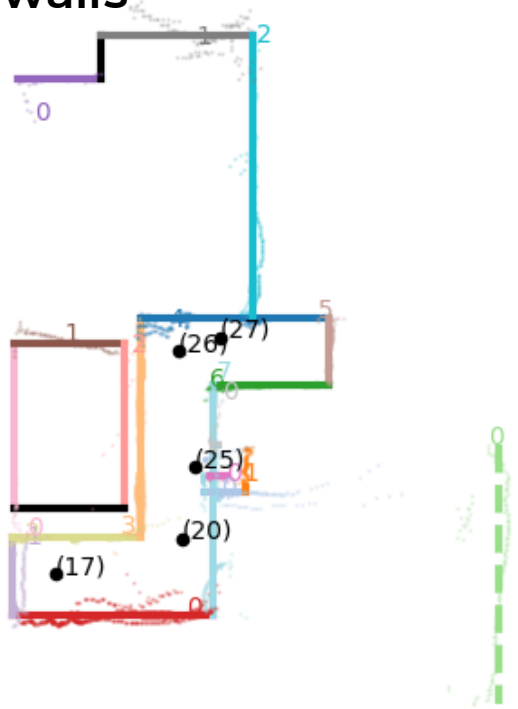




Input panoramas with instance segmentation of ordered walls

- Predicting dense objectives from global room sequences:
  - **Cross-view global wall instance** detection / segmentation
  - **Solving pairwise wall connectivity** by tracing wall instance

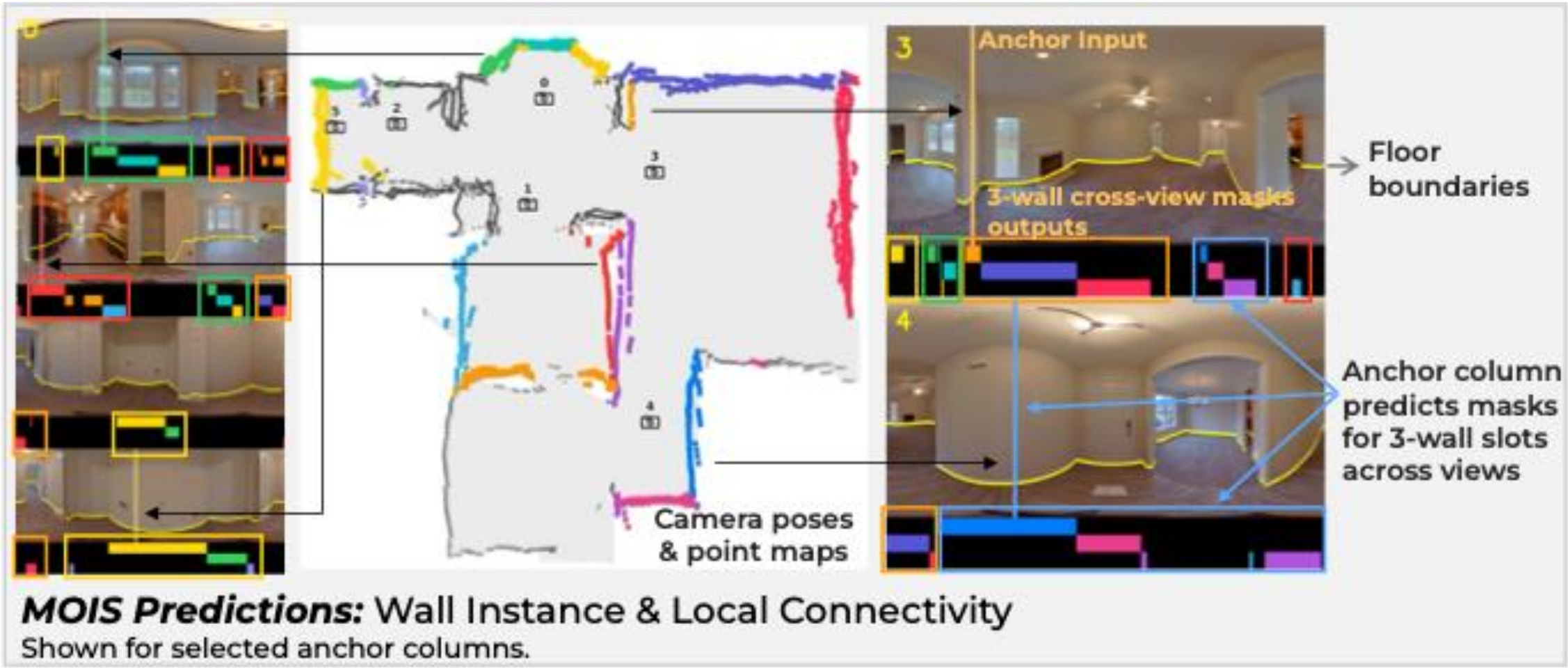
- Doesn't solely rely on BEV, can **tolerate much larger pose ambiguity**.



Room-level wall graph

MOIS Reformulation: make the hard problem local, dense and learnable

- MOIS dense objectives: wall instances + wall connectivity.
- Decoder input: query tokens encoded with column positions.
- Decoder output: Cross-view segmentation on the physically adjacent k walls starting from this query column.



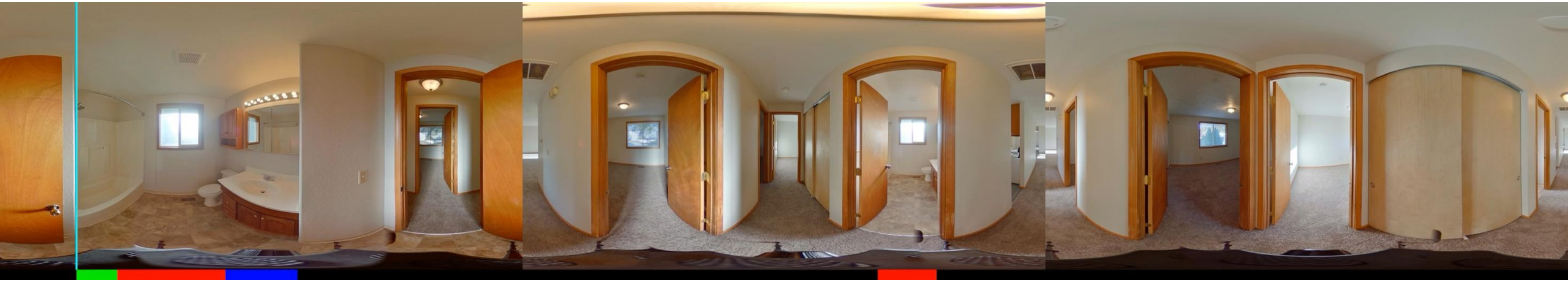
MOIS dense objective visualized for a sample.

MvFFN predict dense MOIS objective in a single call, shaped as N x W x N x W x 3

Dense column position query

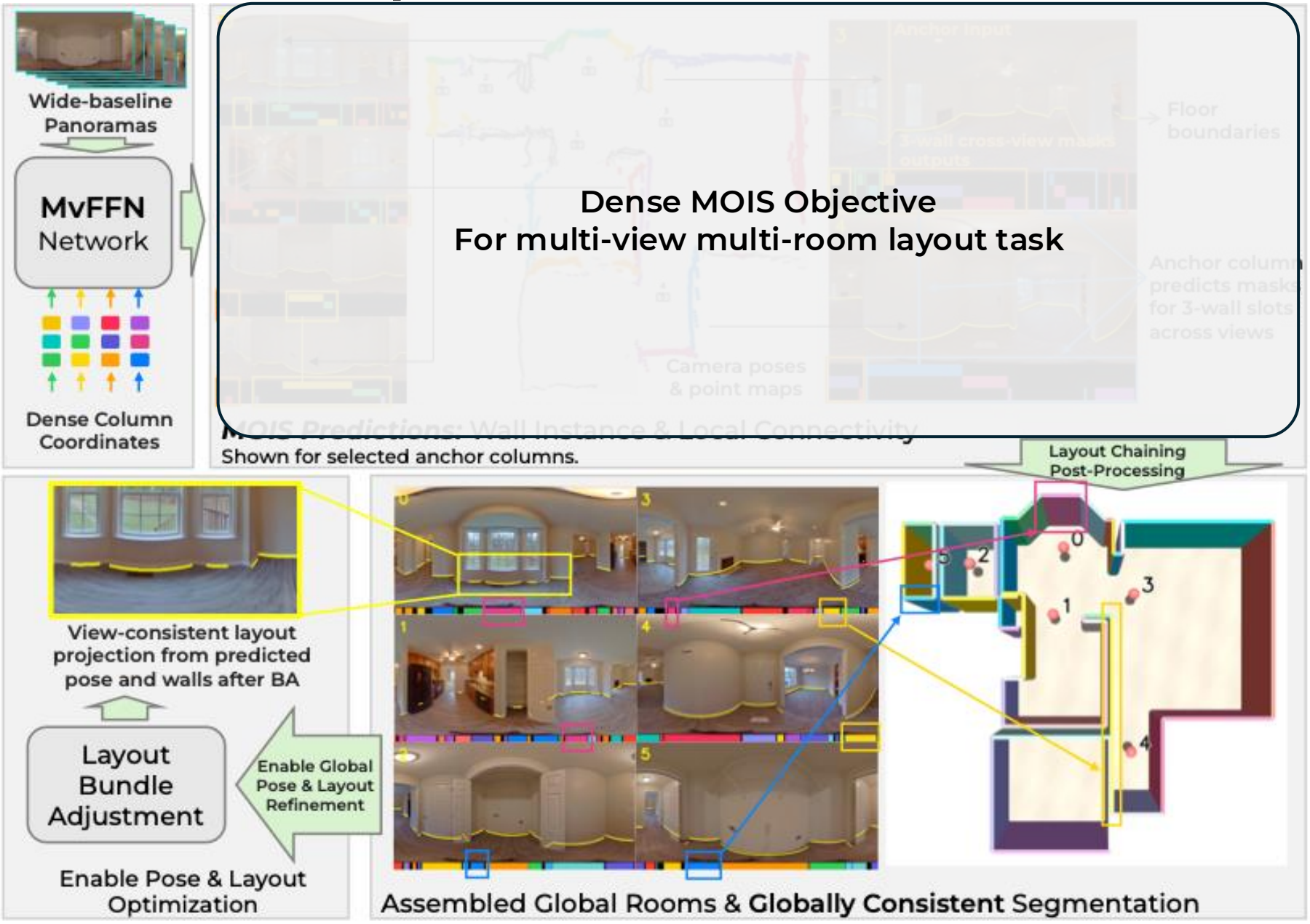
Cross-view segmentation for 3 walls per-query

A training sample, where the whole video demonstrates the prediction of one inference call.



Masked areas in ground truth – Not used in training

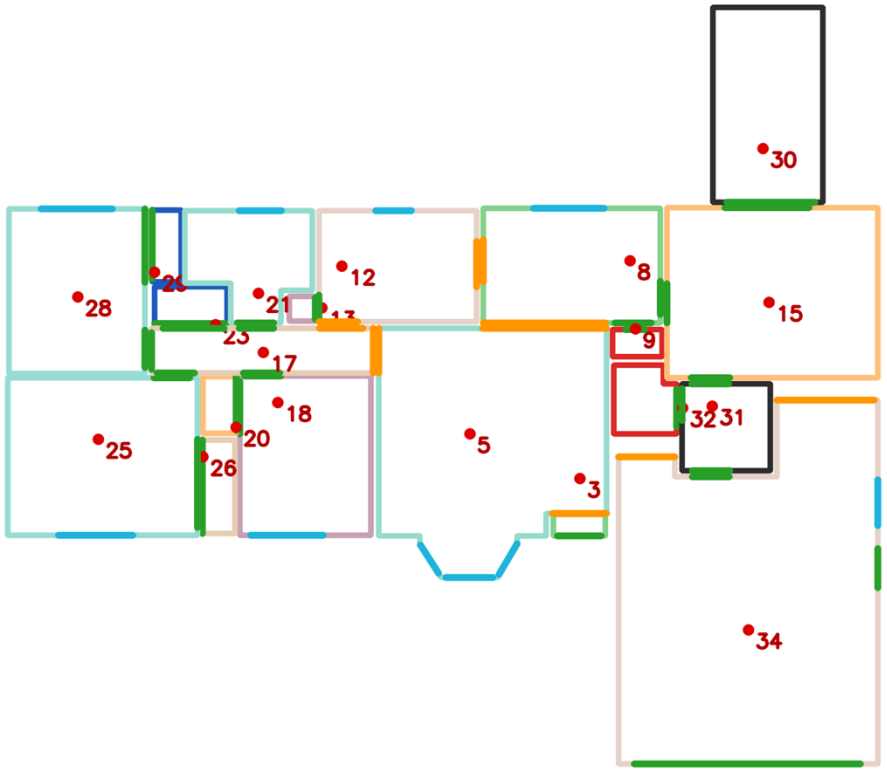
MOIS-based Pipeline Overview



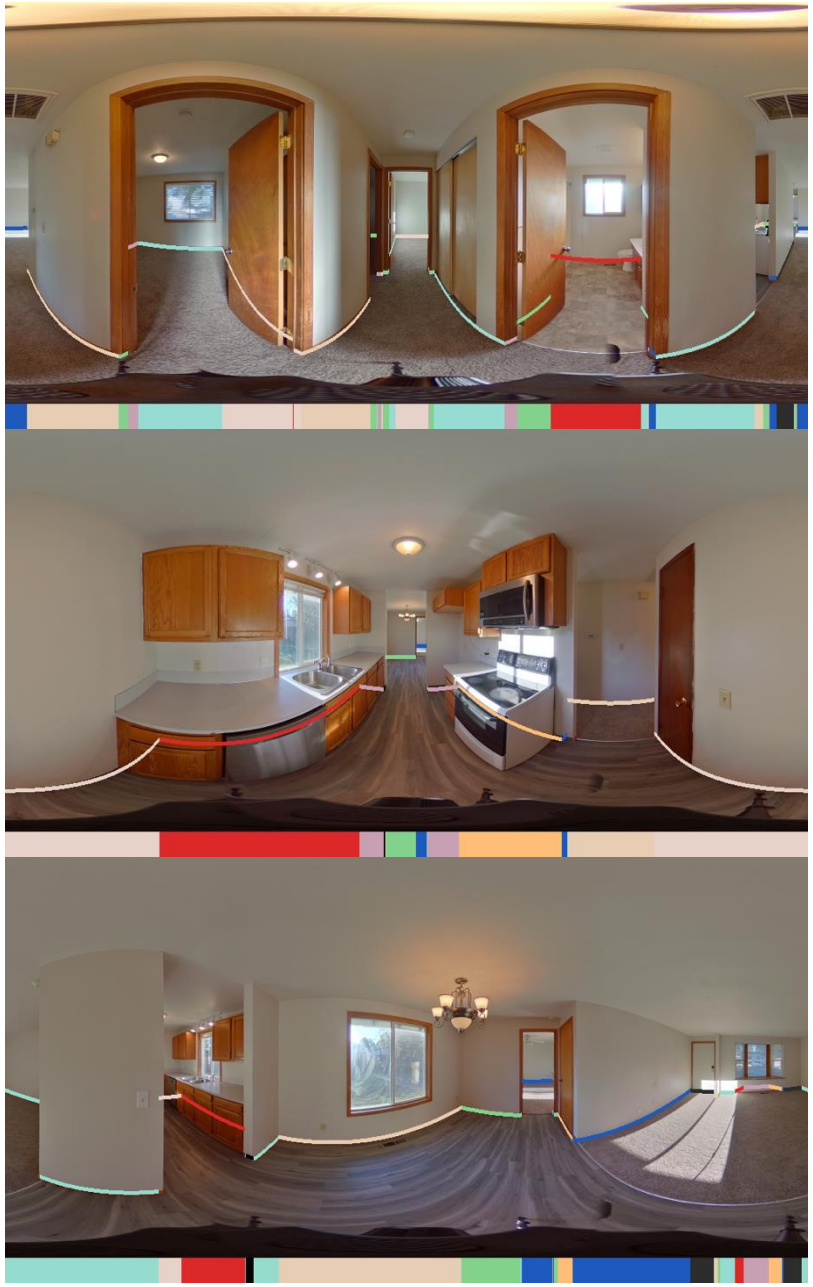
- MOIS enable the effective model convergence with noisy limited training data
- Potentials with better decoder header, using MOIS as a foundation task

Introducing ZInD-XV dataset

- Extension of ZIND dataset: 2706 floors and 66837 panoramas.
  - Merge smaller rooms joined by openings into open spaces.
- Cross-view globally unique wall instances and per-view segmentation.
- 40hr+ annotation for masking.



Original ZIND floor plans

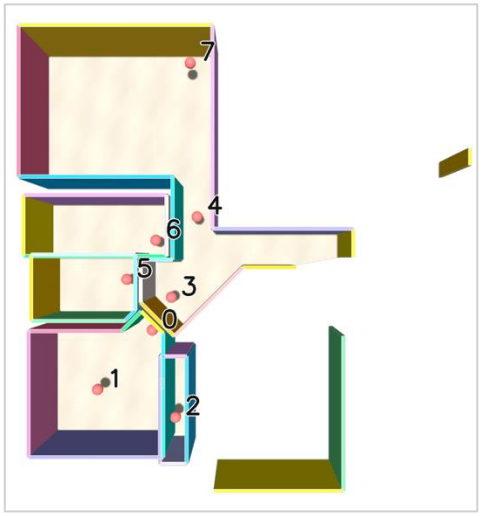


Panoramas with floor boundaries & cross-view instance segmentation

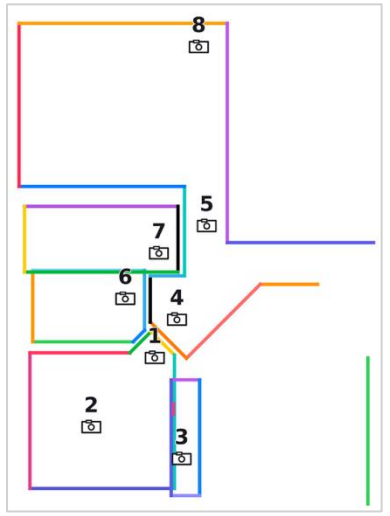
MvFFN: multi-room, multi-view floor plan reconstruction from wide baseline unposed panoramas



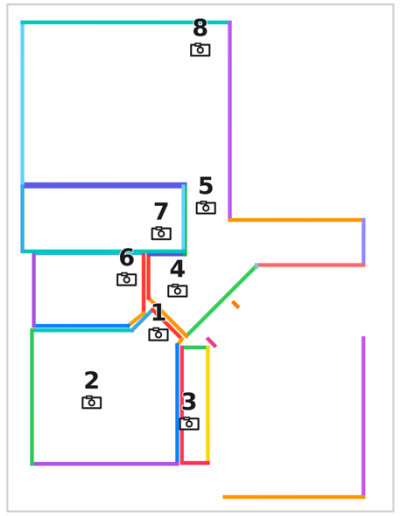
Input panoramas with predicted wall-instance segmentation



Pred layout from *MvFFN* (rendered)

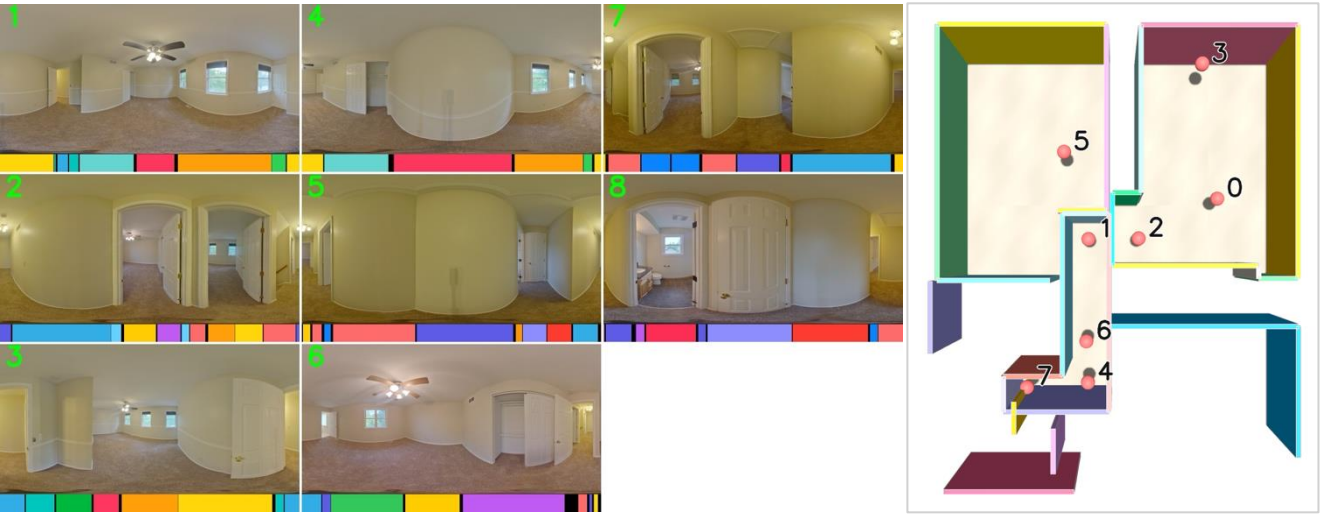


Pred layout from *MvFFN* (lines)



GT layout

MvFFN Accuracy Preview



Robust Pose Prediction Accuracy

Methods	8v Pos (cm)↓			8v Rot (°)↓		
	Mn	Std	Med	Mn	Std	Med
CovisPose+GST	17.3	37.3	**11.8**	1.9	1.6	**0.8**
VGGT	53.4	63.1	27.8	13.3	17.7	4.9
MvFFN	**17.0**	**15.6**	13.8	**1.3**	**0.8**	1.2

Layout Accuracy

Modular Pipeline			8-view wall				8-view junction			
pose	geom.	layouts	Prec.↑	Rec.↑	F1 ↑	Dist ↓	Prec.↑	Rec.↑	F1 ↑	Dist ↓
MvFFN			**89.4**	**87.5**	**88.4**	**19.0**	**81.0**	**69.8**	**75.0**	**19.1**
CP		*RF*	35.0	42.4	38.3	79.0	15.7	18.6	17.0	136.8
CP		*PR*	31.8	25.0	28.0	86.2	13.8	10.5	11.9	133.4
CP	*LGT*	*LGT-PP*	76.3	72.7	74.5	24.4	61.7	56.7	59.1	24.4

MvFFN has two jobs: geometry first, layout next

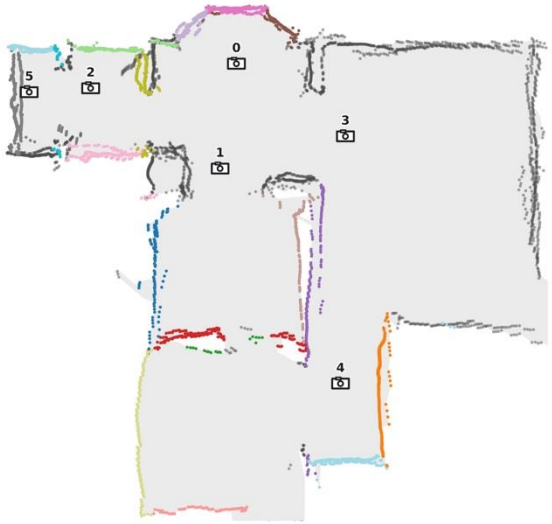
The model needs to understand:

- Foundation: Multi-view dense points & poses.
  - Joint semantic and geometry understanding.
- Task specific: Decode into globally unique wall instances in order.
  - Can tolerate more pose & geometry uncertainty.



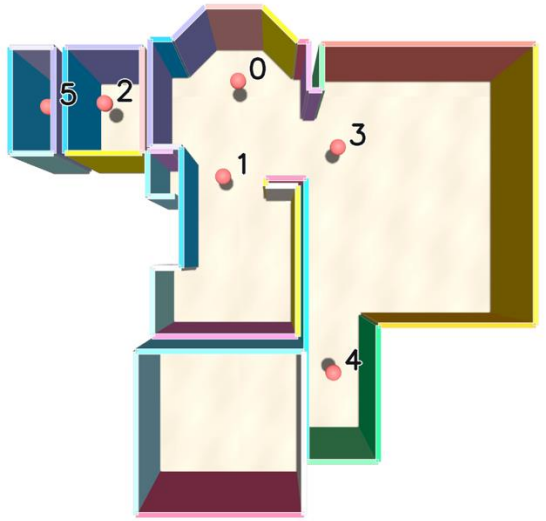
Unposed Panos

Multi-view geometry transformer encoder



Dense global geometry & poses

Multimodal layout decoder



Structured layouts as ordered wall instances



Part 1 – Geometry Foundation

Multi-view shared geometry

Geometry transformers give us a starting point

- VGGT (CVPR 2025) predicts camera poses and dense 3D point maps.
- Inputs: N adjacent or medium baseline perspective images
- Outputs: camera poses and dense point maps
- MvFFN adapts this idea to sparse 360° panoramas.



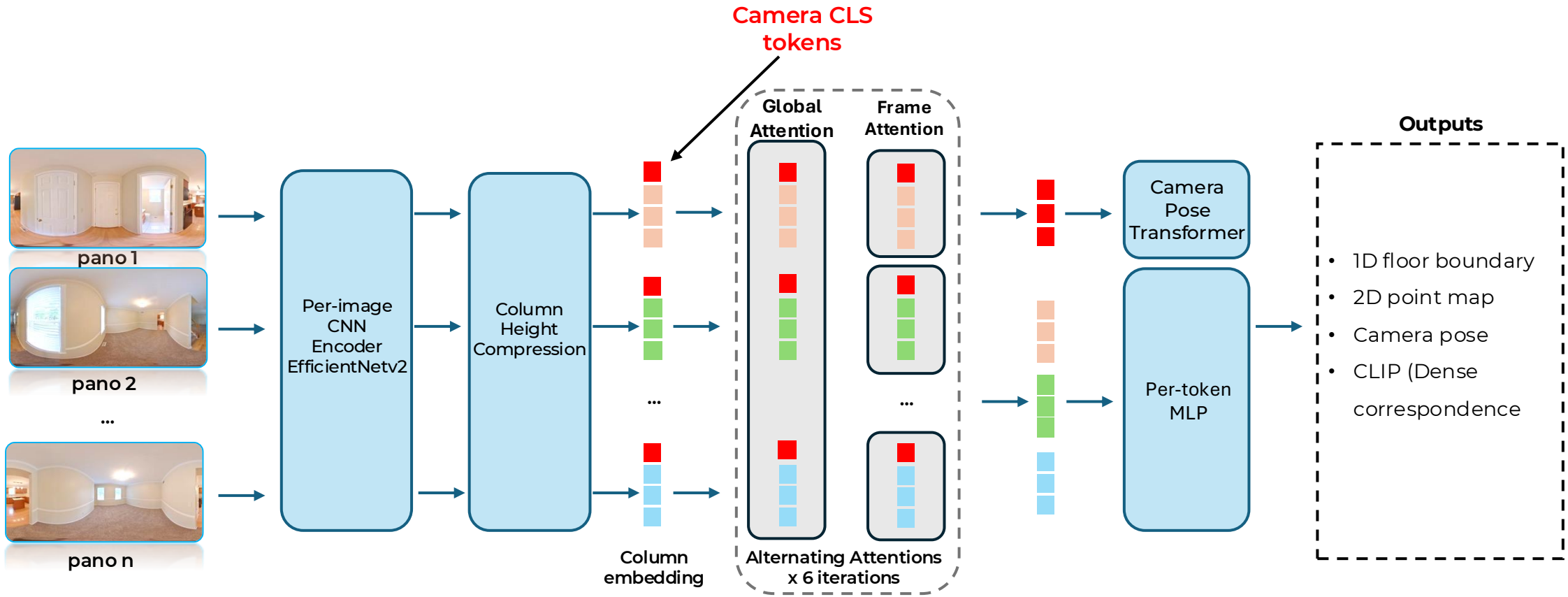
Input video frames



Output global 3D point maps & camera poses

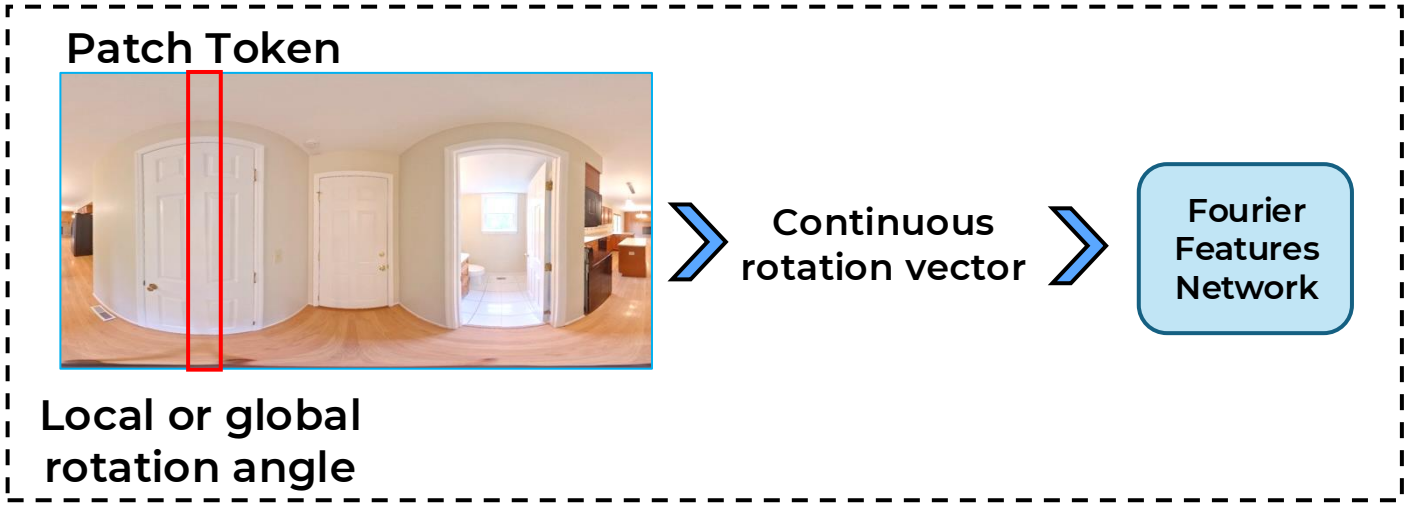
Alternating attention for multi-view geometry

- **Alternating attention**: allows more efficient attention within each frame and across panoramas.
- **Camera CLS tokens**: predict camera pose from shared context.
- **Training objective**: Pointmap L2 loss + confidence loss drives cross-view geometric alignment.

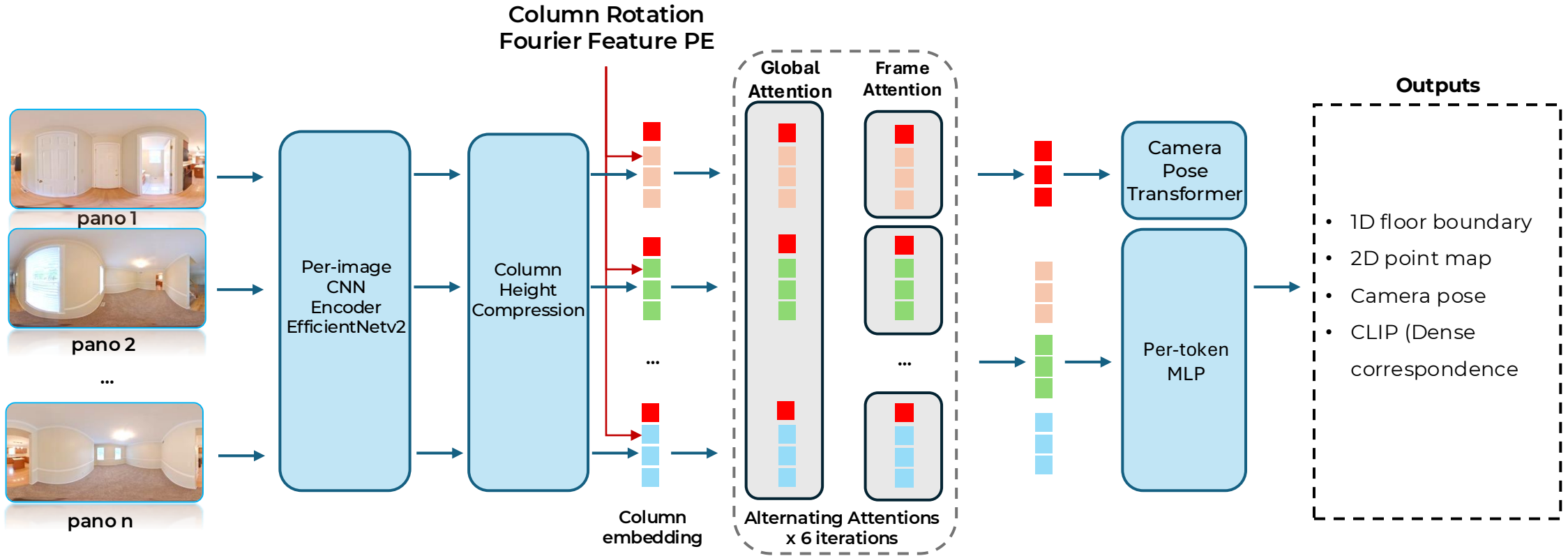


Flash Attention

Panoramas need rotation-aware column tokens

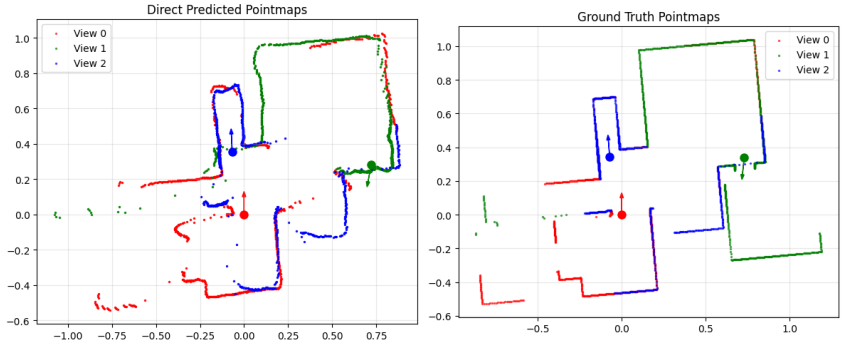


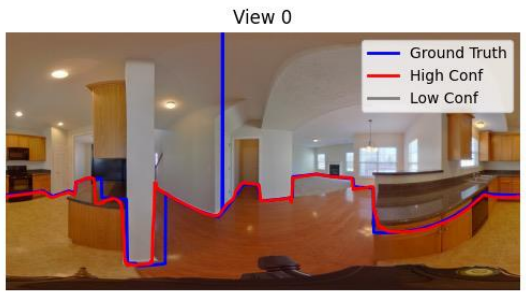
Patch PEs as Fourier features of local/global horizontal rotation of patches.



Qualitative examples

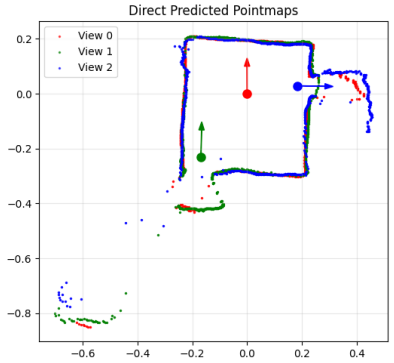
- Curriculum training over the view counts: Start with 2 - 3 views, finish with 5 - 8 views.
- 3-view geometry is stable.

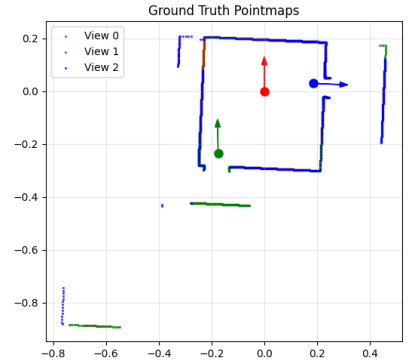


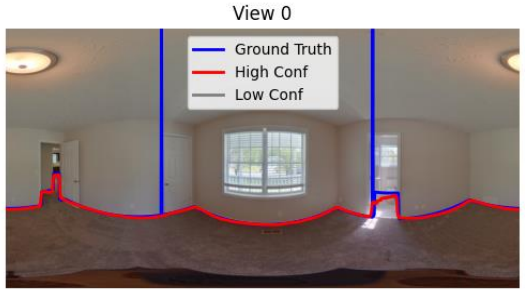






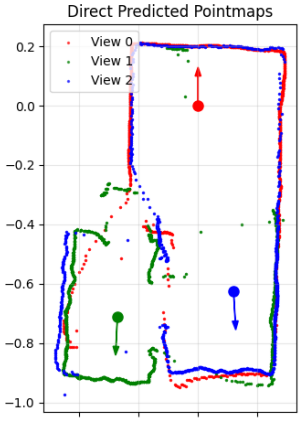


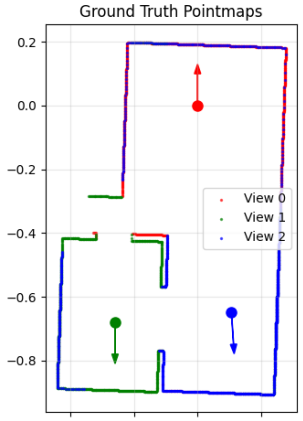












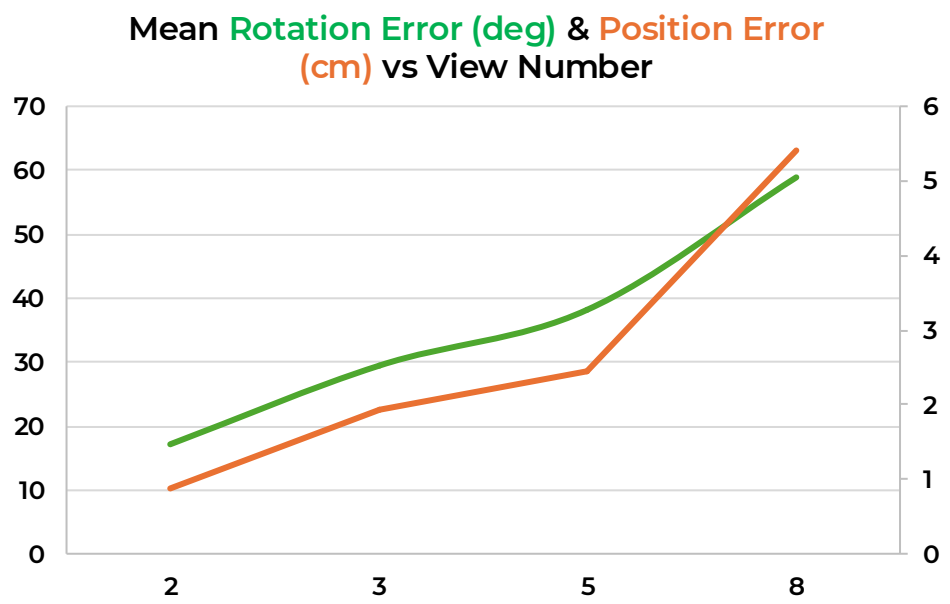


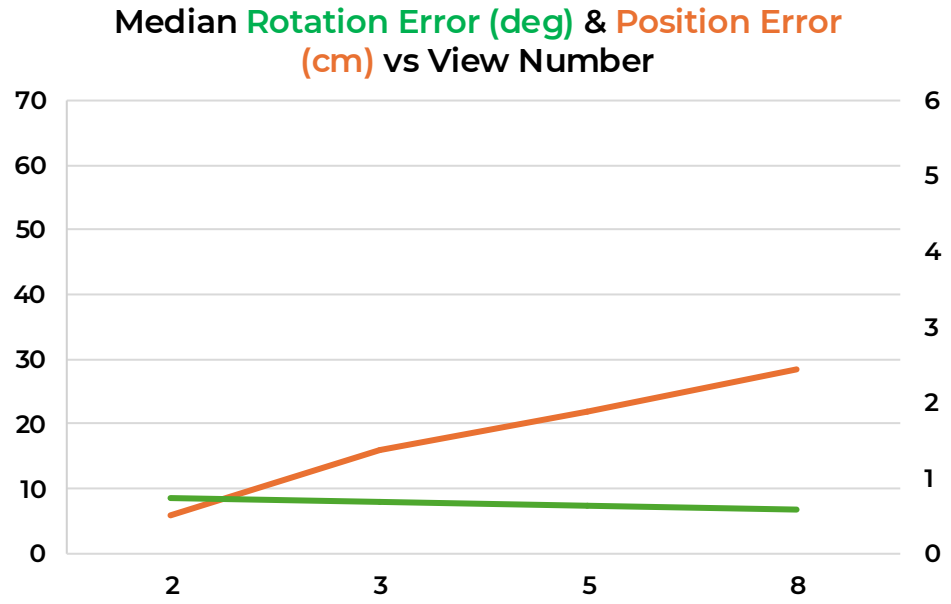




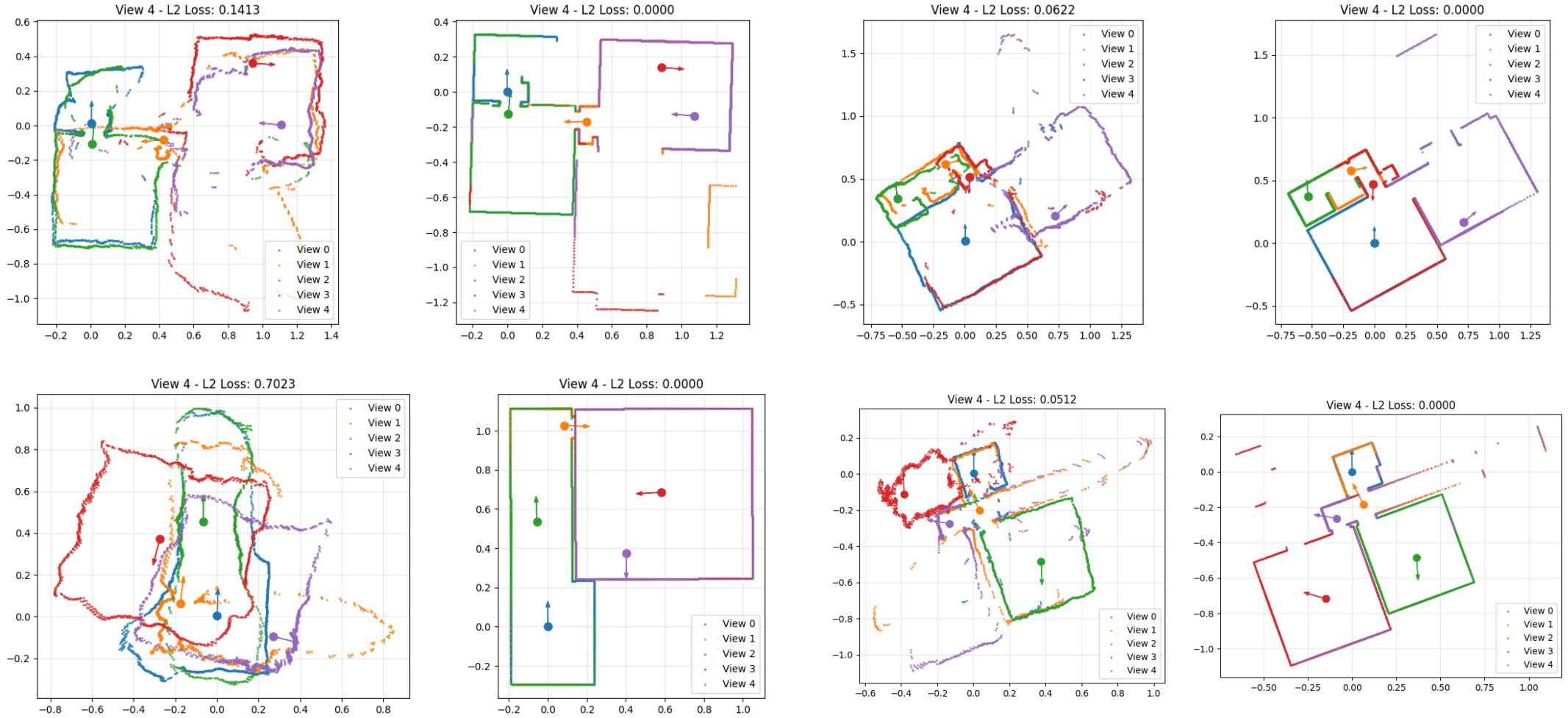
Pose accuracy drops quickly as number of views increases

- Outliers grow quickly from 2 to 8 views.
- **Severe data scarcity** for highly data-hungry multi-view tasks:
  - ZInD: 55k training images, 2k floor scenes
  - VGGT: 17 datasets, 50+ million images, 100k+ unique 3D scenes
- MvFFN uses geometry as support, while the main objective is layout reconstruction.





5 views expose the data-scarcity problem

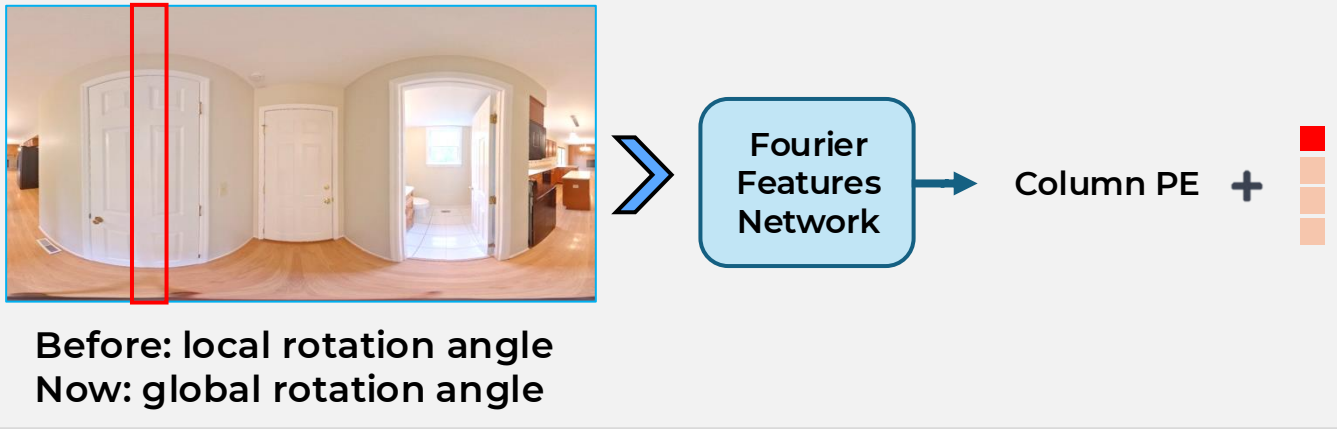


Predicted point maps vs. ground-truth geometry

Inductive Biases for Severe Data-Scarce Multi-view Training

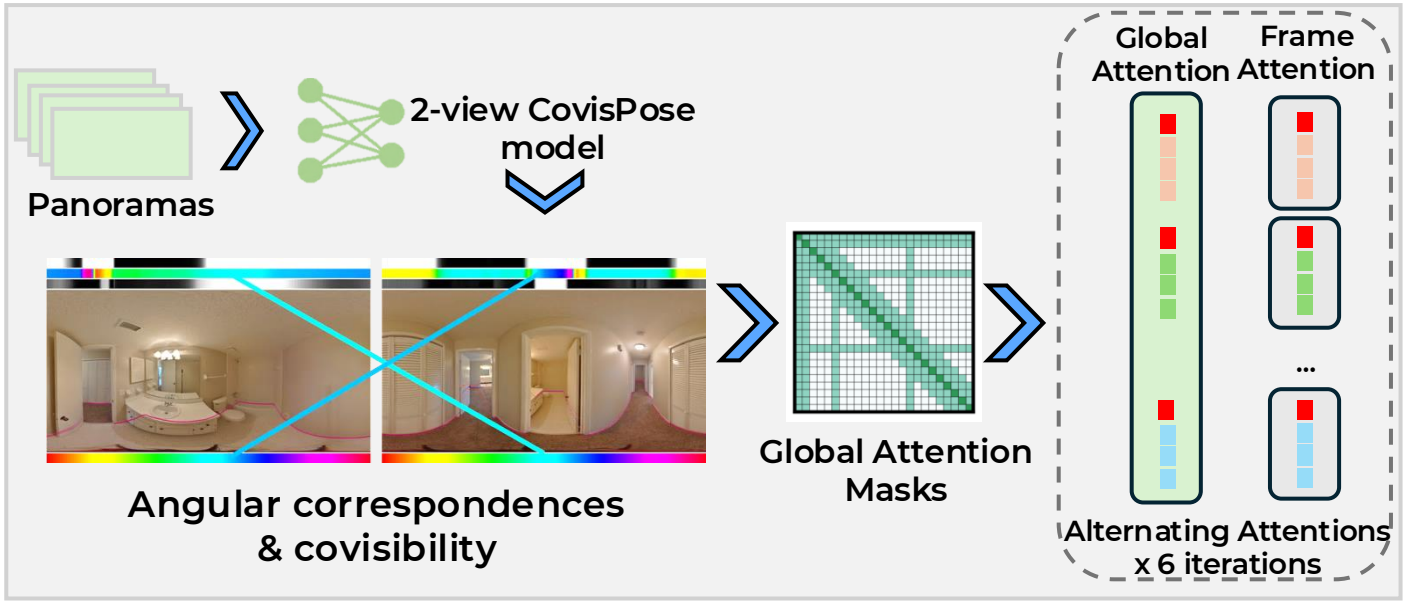
Improvement 1: Reduce pose task DoF, use 2-view rotation as a global cue.

- Estimate global rotation from 2-view MvFFN / CovisPose + angle refinement
- Inject global rotation as column PE



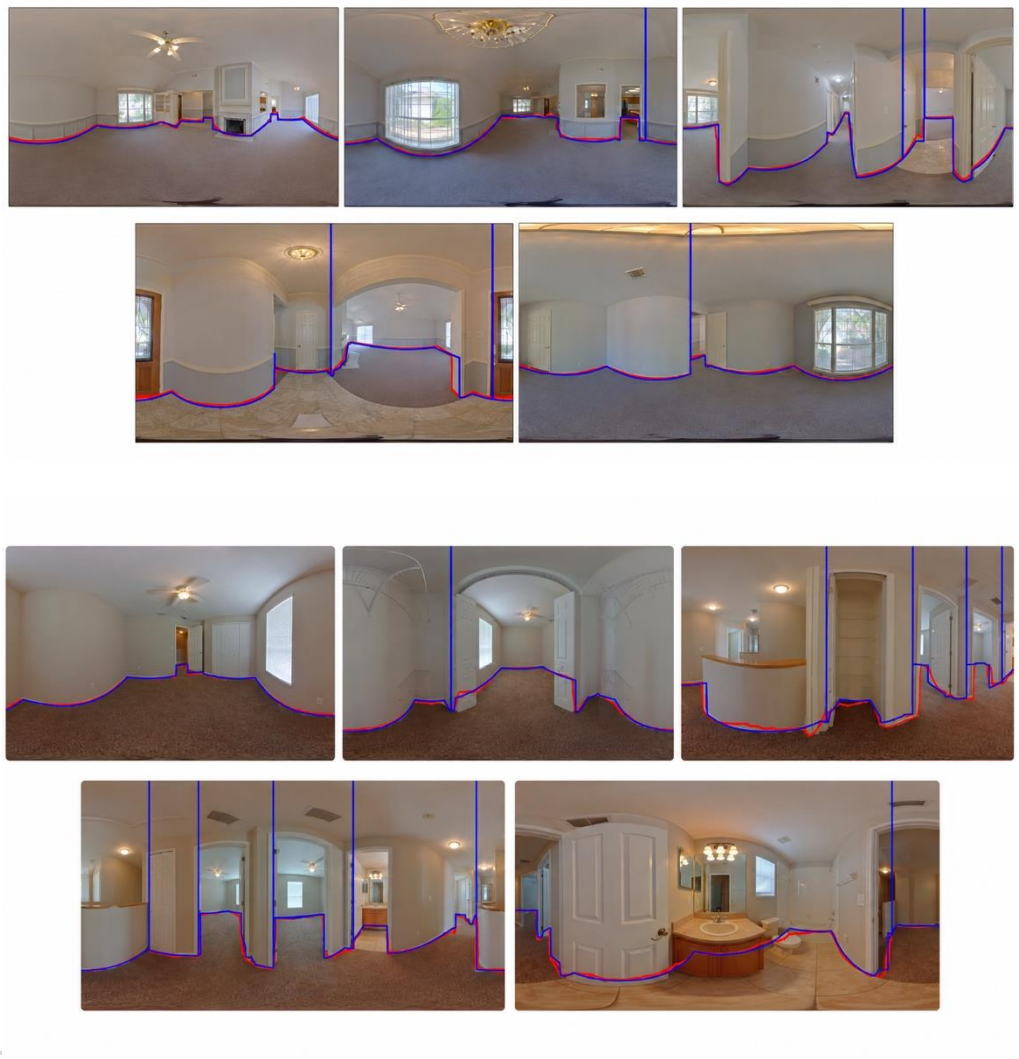
Improvement 2: Reduce from full attention, use dense correspondence priors as attention mask.

- Expand each angular correspondence match to an 80-px wide attention mask.
- Mask unrelated columns during global attention.

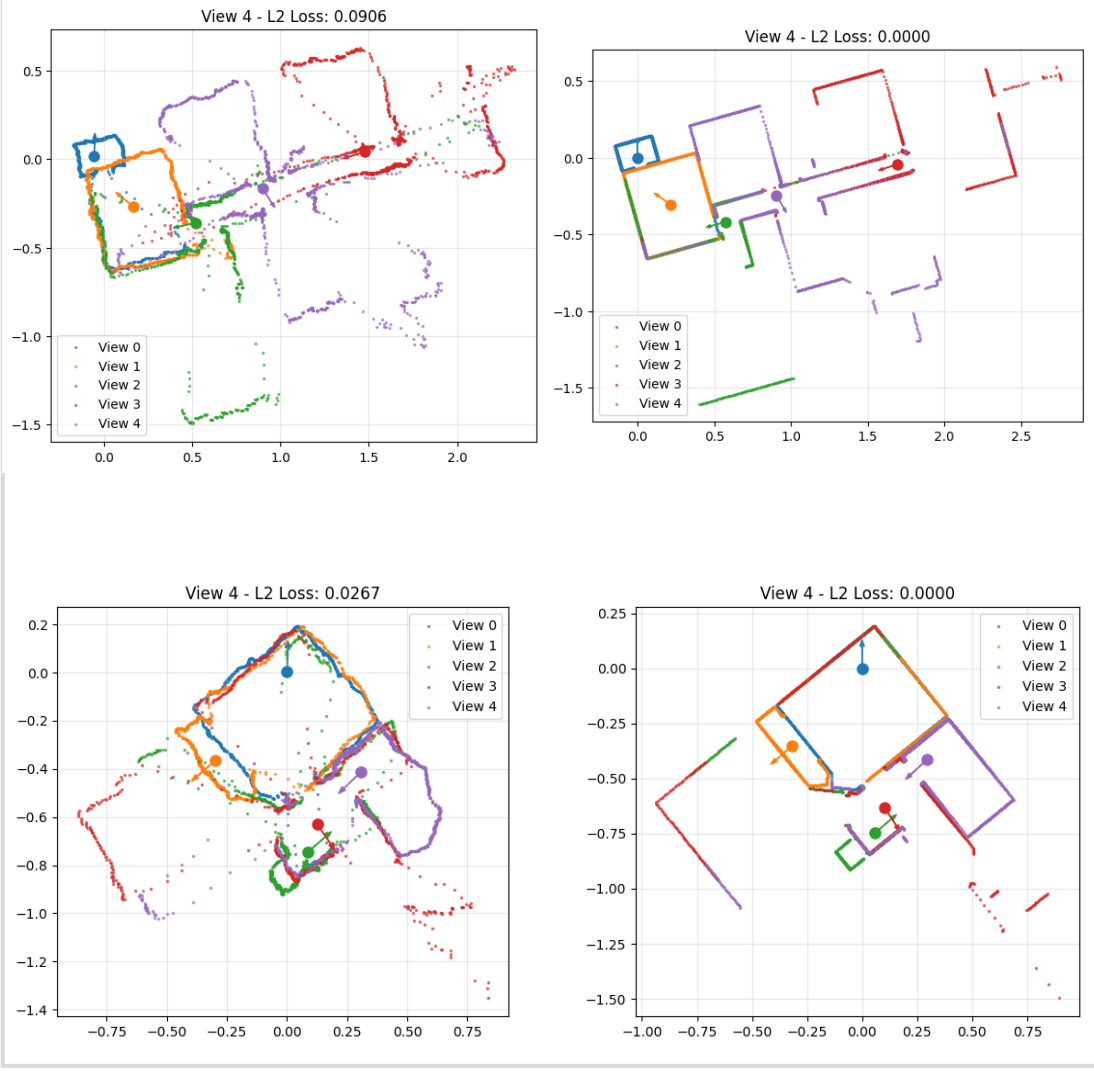


Inductive Bias Enables 5-8 View Convergence

Qualitative examples: global point maps and camera poses become stable under limited training data.



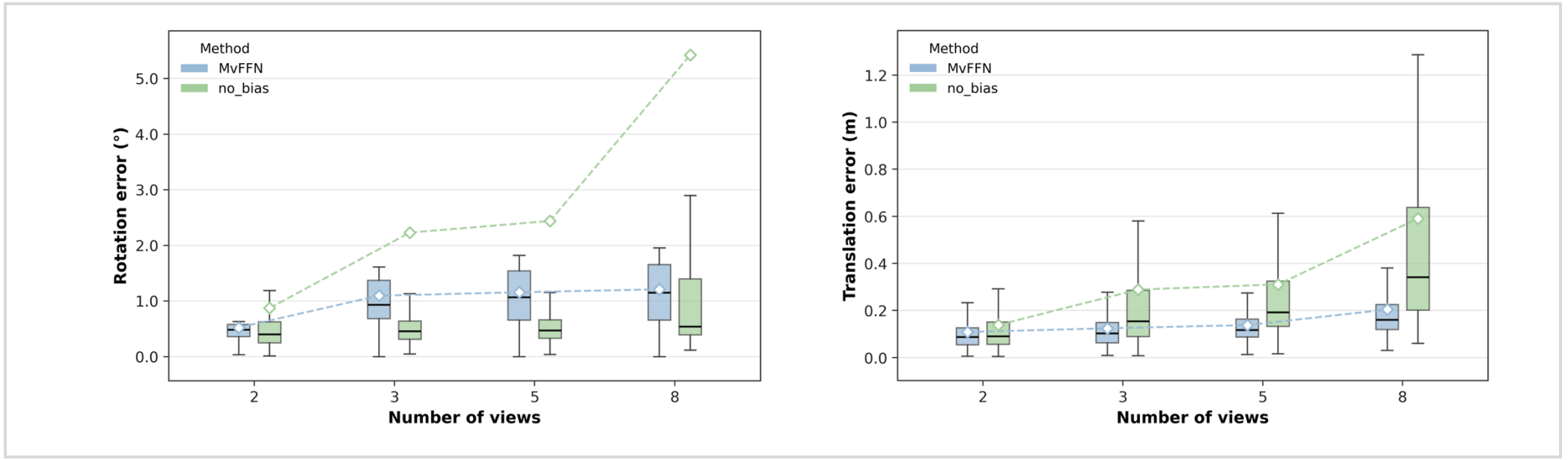
Floor Boundaries Predictions



Global Pointmaps Predictions

Inductive Bias Improves Training with Data Scarcity

Priors reduce error growth and variance as the number of input views increases



Ablation of different inductive Biases

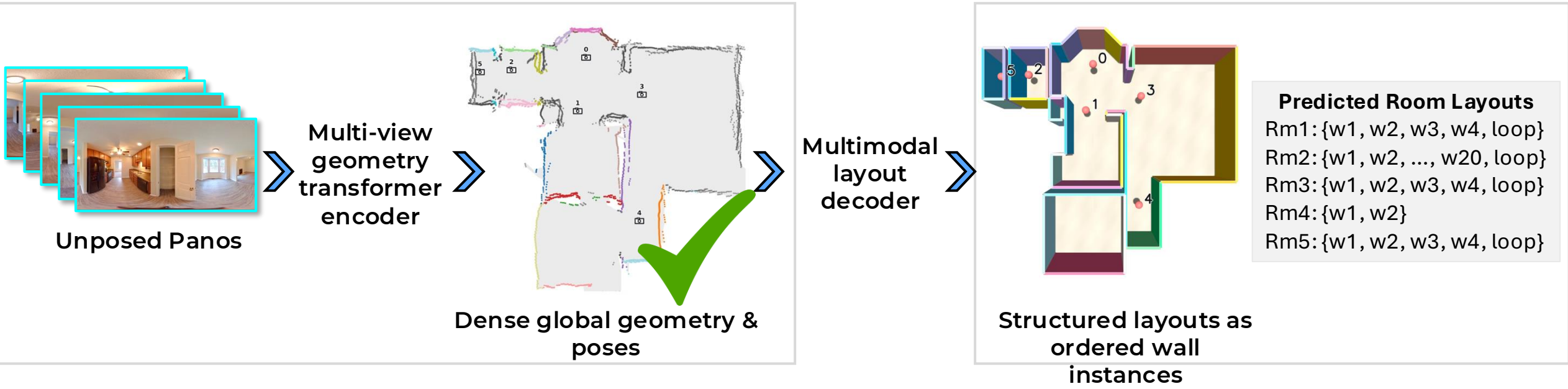
Table 1: Camera pose evaluation on ZInD-XV dataset.

Methods	5v Pos (cm)↓			5v Rot (°)↓			8v Pos (cm)↓			8v Rot (°)↓		
	Mn	Std	Med	Mn	Std	Med	Mn	Std	Med	Mn	Std	Med
CovisPose+GST	12.6	23.3	**9.3**	1.3	0.9	**0.7**	17.3	37.3	**11.8**	1.9	1.6	**0.8**
VGGT	30.7	39.4	13.3	11.1	16.0	4.0	53.4	63.1	27.8	13.3	17.7	4.9
MvFFN	**12.5**	**10.8**	10.2	**1.2**	**0.6**	1.1	**17.0**	**15.6**	13.8	**1.3**	**0.8**	1.2

Pose accuracy with inductive bias.

Takeaway: geometry works, but data is scarce

- We have built a strong multi-view geometry backbone.
  - Scales to 8-view training data with limited samples.
  - Inductive bias is required under current public-data limits.
  - We can train without inductive bias when we have enough data.
  - Unblock our main task.

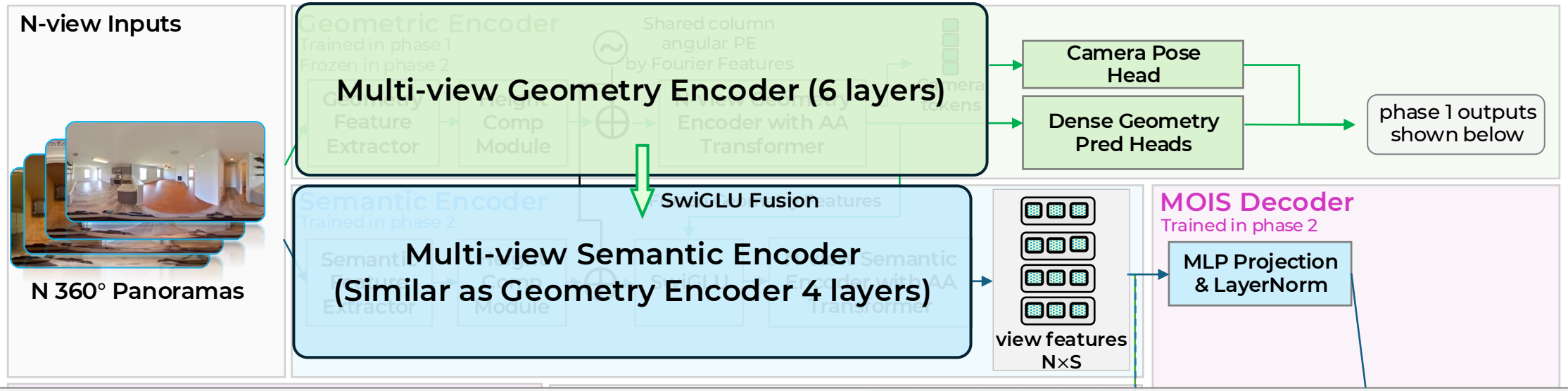




Part 2 – Layout Objective

Multi-view wall instance tracing

MvFFN – Our prediction network around MOIS objective



Insights - Geometry and Semantic Attention are different:

- Geometry Attention: Push uniqueness for each point
- Semantic Attention: Aggregating points of similar semantics

Trained in 2 stages:

1. Geometry only
2. Freeze geometry & train semantic with the layout decoder

We only visualize the *MOIS* outputs for a single column query

MvFFN – Ablation Study

Table 4: Ablation study of *MvFFN* on `ZInD-XV`, comparing encoder design, geometry fusion, and staged training. – denotes pose metrics identical to *MvFFN*.

Method	Geom Enc	Sem Enc	Geom Fusion	Train Stages	5v pos (cm)↓			5v rot (°)↓			5v wall				5v junction			
					Mn	Std	Med	Mn	Std	Med	Prec↑	Rec↑	$F1$ ↑	Dist↓	Prec↑	Rec↑	$F1$ ↑	Dist↓
Shared-Enc	✓	✗	✗	2	35.8	50.1	26.4	2.1	1.3	2.0	86.4	69.8	77.2	31.2	72.9	50.4	59.6	39.9
No-Fusion	✓	✓	✗	2	–	–	–	–	–	–	**92.6**	22.7	36.5	28.0	**84.3**	14.4	24.6	21.4
1-stage	✓	✓	✓	1	46.8	50.5	30.1	2.5	1.5	1.9	87.7	71.9	79.0	33.9	75.8	53.1	62.5	44.1
MvFFN	✓	✓	✓	2	**12.5**	**10.8**	**10.2**	**1.2**	**0.6**	**1.1**	88.5	**87.7**	**88.1**	**13.9**	79.2	**68.2**	**73.3**	**13.8**

Annotation Made Easy with AI for HM3D-XV dataset

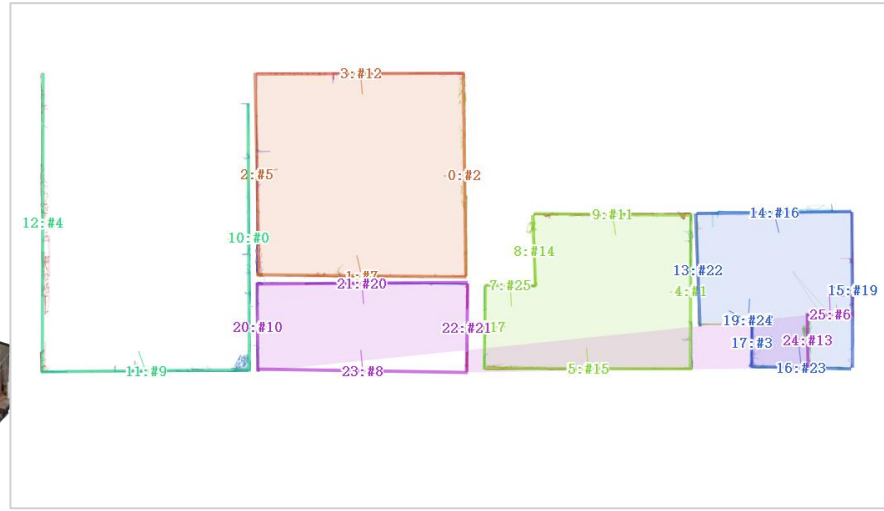
- Extension of HM3D dataset: mesh to wall instance with a human-in-the-loop process.
- Coding agent enables complex annotations with web GUI within hours.

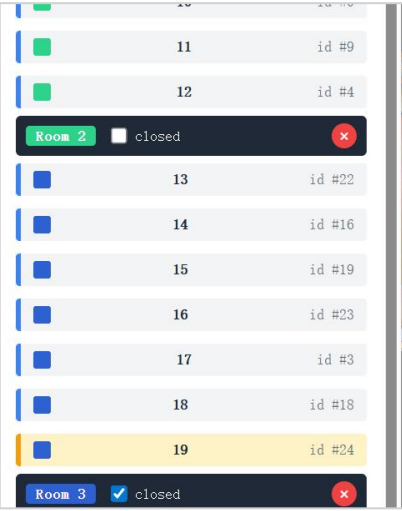


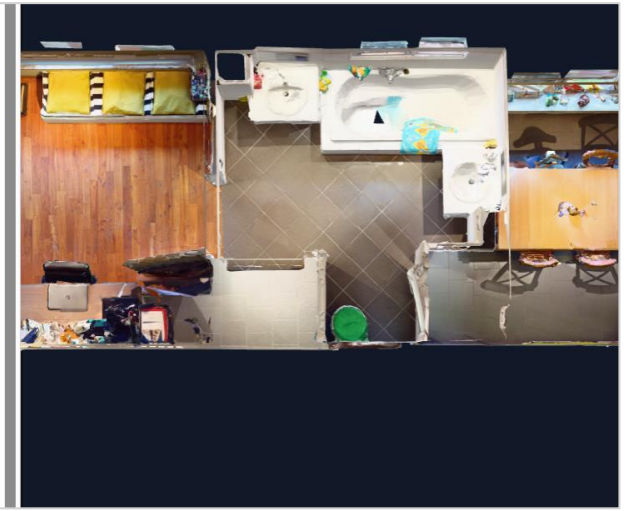
HM3D-XV Panoramas



3D mesh from HM3D dataset





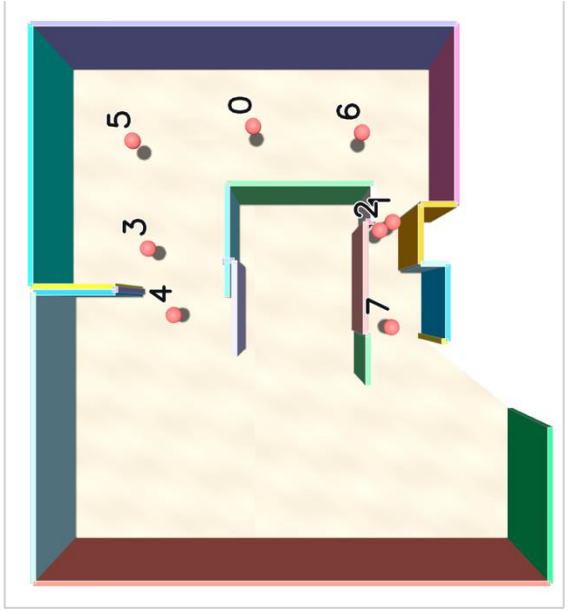


Quick Annotation Web UI Written with Claude in 2 hrs

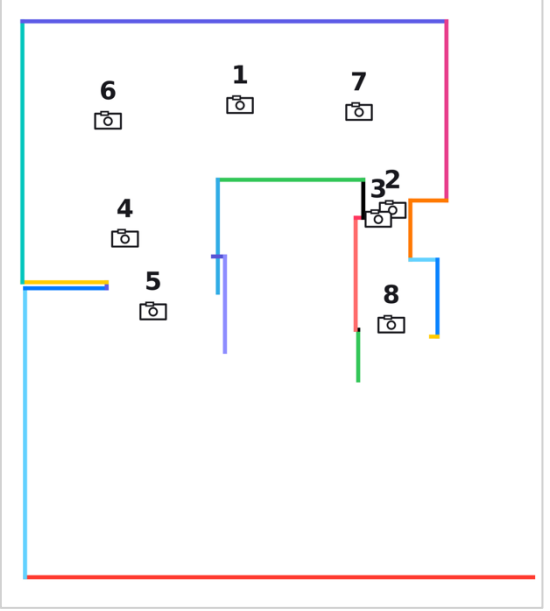
Qualitative Results



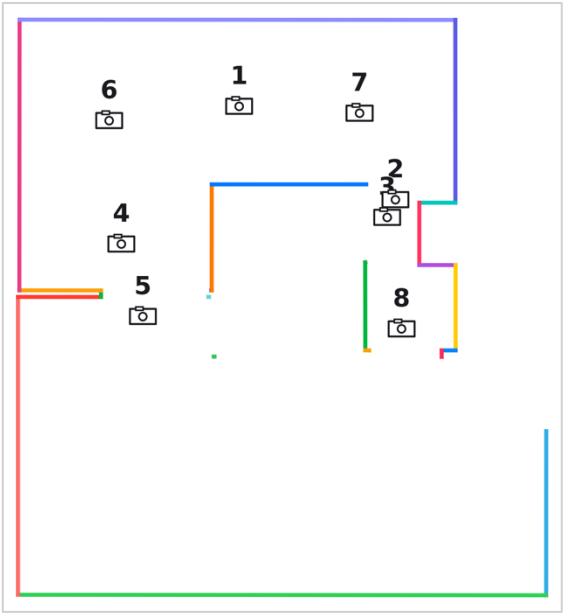
Input panoramas with predicted wall-instance segmentation



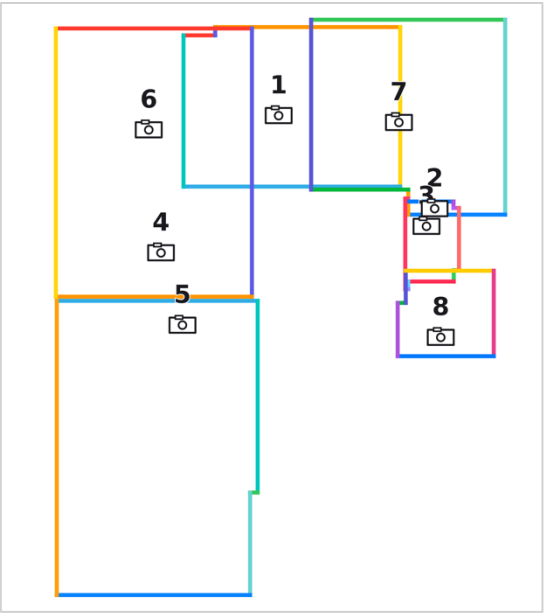
Pred layout from *MvFFN* (rendered)



Pred layout from *MvFFN* (lines)



GT layout

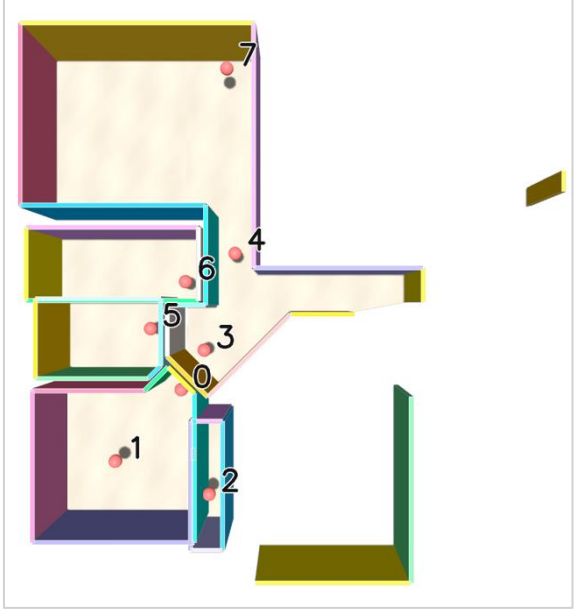


Pred layout from *CovisePose+LGTNet*

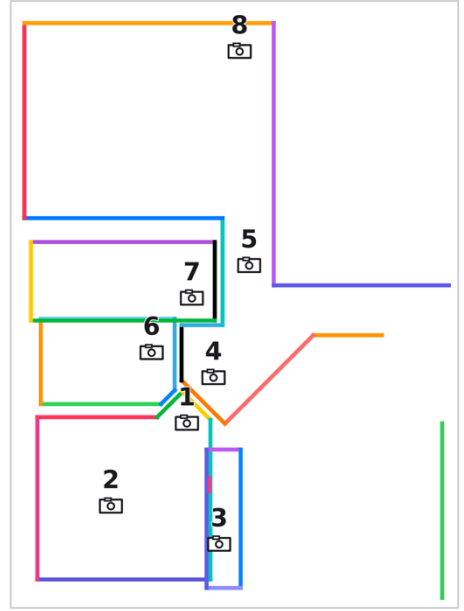
Qualitative Results



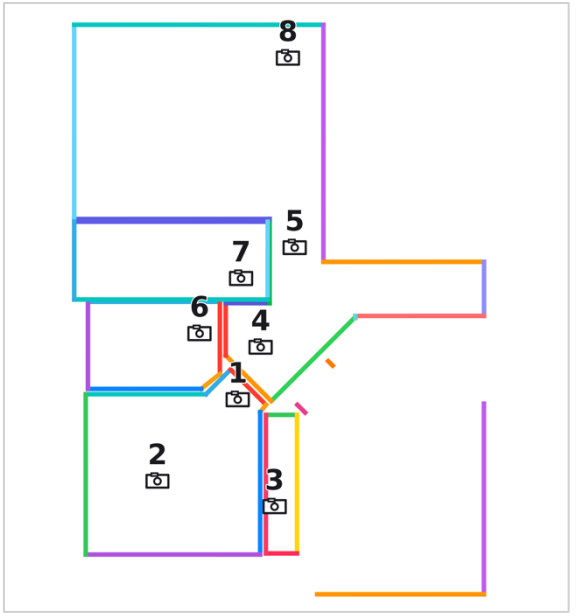
Input panoramas with predicted wall-instance segmentation



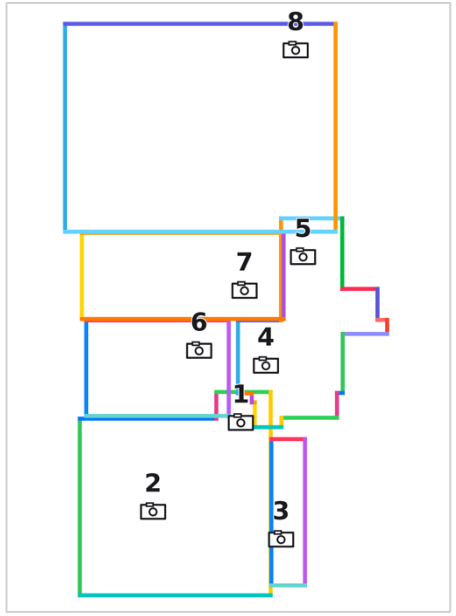
Pred layout from *MvFFN* (rendered)



Pred layout from *MvFFN* (lines)



GT layout

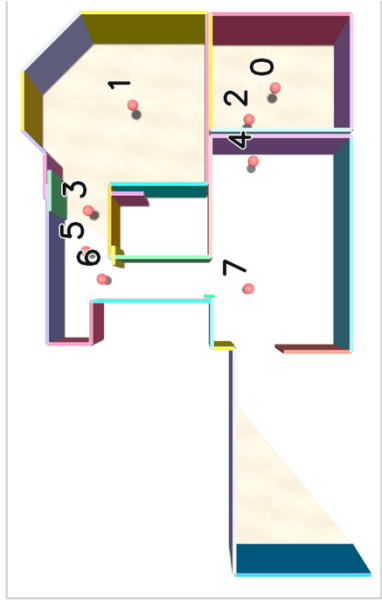


Pred layout from *CovisePose+LGTNet*

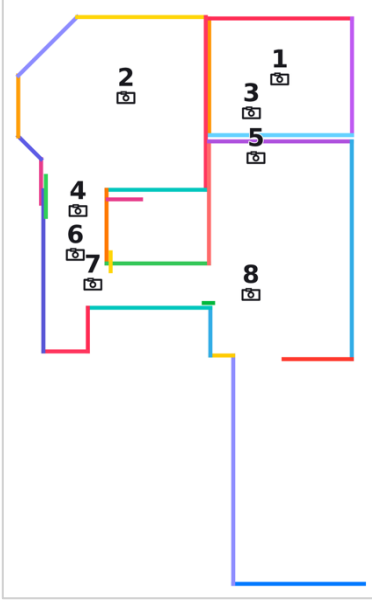
Qualitative Results



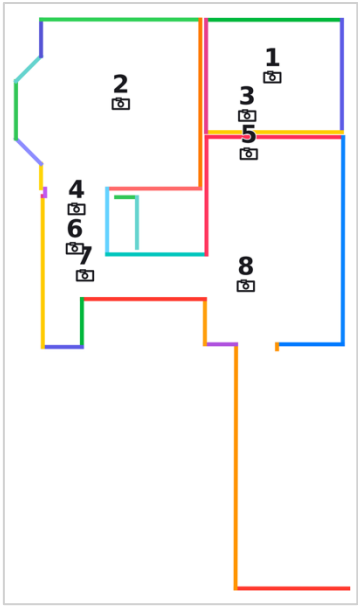
Input panoramas with predicted wall-instance segmentation



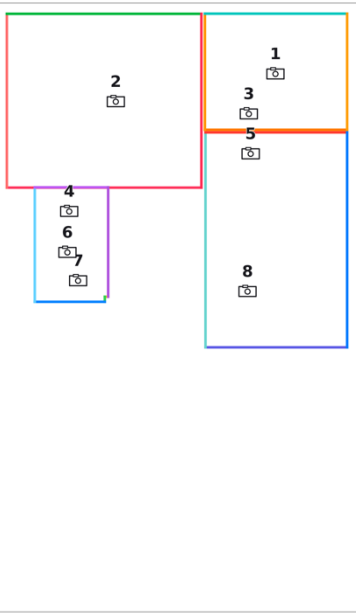
Pred layout from *MvFFN* (rendered)



Pred layout from *MvFFN* (lines)



GT layout



Pred layout from *CovisePose+LGTNet*

Layout Accuracy

Table 2: Floor-layout evaluation on ZInD-XV. Distance is in cm. Geom. denotes the intermediate geometric input and Layout denotes the vectorization backend. GT-pose denotes a control group.

Modular Pipeline			5-view wall				5-view junction				8-view wall				8-view junction			
pose	geom.	layouts	Prec.↑	Rec.↑	F1 ↑	Dist ↓	Prec.↑	Rec.↑	F1 ↑	Dist ↓	Prec.↑	Rec.↑	F1 ↑	Dist ↓	Prec.↑	Rec.↑	F1 ↑	Dist ↓
MvFFN			**88.5**	**87.7**	**88.1**	**13.9**	**79.2**	**68.2**	**73.3**	**13.8**	**89.4**	**87.5**	**88.4**	**19.0**	**81.0**	**69.8**	**75.0**	**19.1**
CP		*RF*	31.9	44.2	37.1	84.9	15.4	20.9	17.7	137.8	35.0	42.4	38.3	79.0	15.7	18.6	17.0	136.8
CP		*PR*	33.6	29.3	31.3	86.6	17.4	14.9	16.1	129.7	31.8	25.0	28.0	86.2	13.8	10.5	11.9	133.4
CP	*LGT*	*LGT-PP*	75.6	71.4	73.4	21.3	60.0	54.8	57.3	21.3	76.3	72.7	74.5	24.4	61.7	56.7	59.1	24.4
GT	*LGT*	*LGT-PP*	78.8	70.5	74.4	19.2	62.4	54.3	58.1	19.7	80.5	70.8	75.3	22.7	65.1	55.5	59.9	22.7

Layout Accuracy Gain Comes from MOIS Formulation & MvFFN

MvFFN + MOIS predicted better multi-view instance segmentation & wall connectivity

Multi-view 3-slot wall segmentation IOU

Method	5-view inputs			8-view inputs		
	wall 1	wall 2	wall 3	wall 1	wall 2	wall 3
MvFFN	**81.9**	**71.1**	**57.8**	**75.4**	**63.8**	**47.7**
LGT + CP	63.7	37.8	20.3	60.4	34.5	18.5

Baseline Pipeline

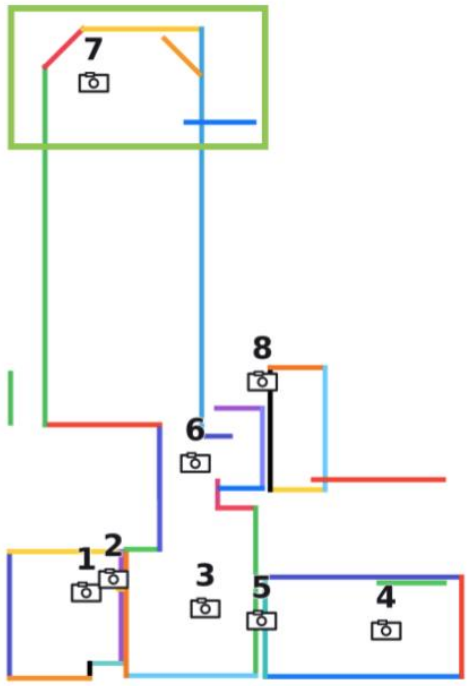


Single view wall segmentation & connectivity From LGT-Net

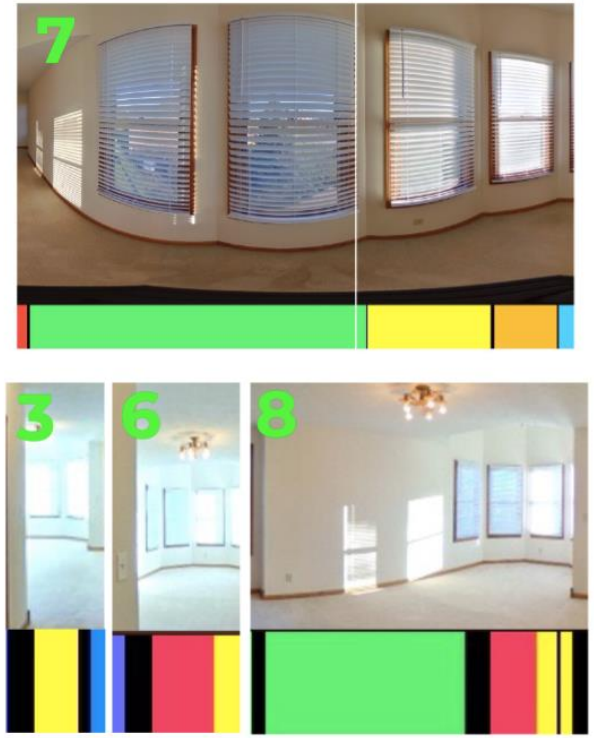


Dense correspondences by 2-view model CovisPose

Failure Cases

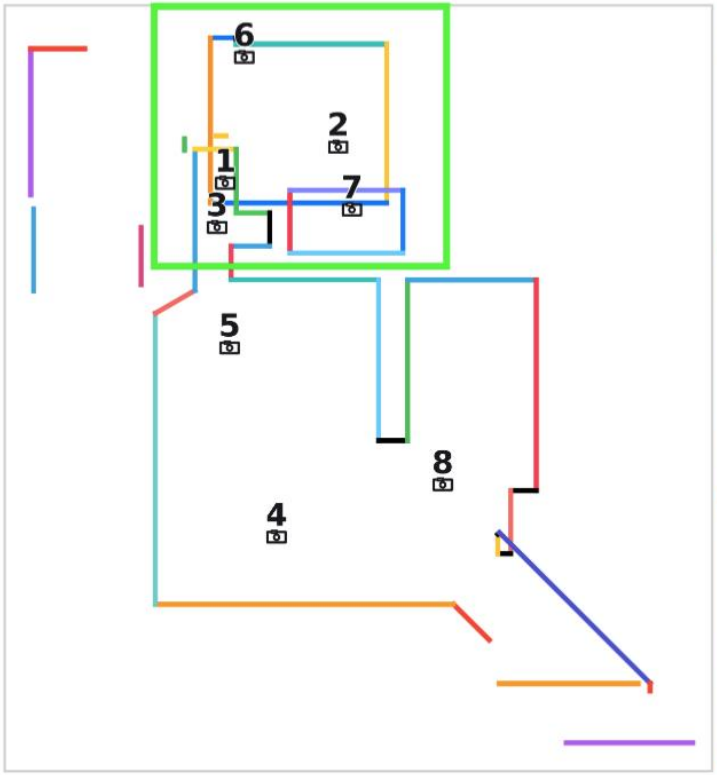






Failure case caused by incorrect MOIS wall-instance prediction, where MvFFN fails to distinguish different wall instances across views.

Failure Cases





Failure case caused by geometric errors in the initial MvFFN predictions, where inaccurate poses and point maps lead to distorted layout geometry.