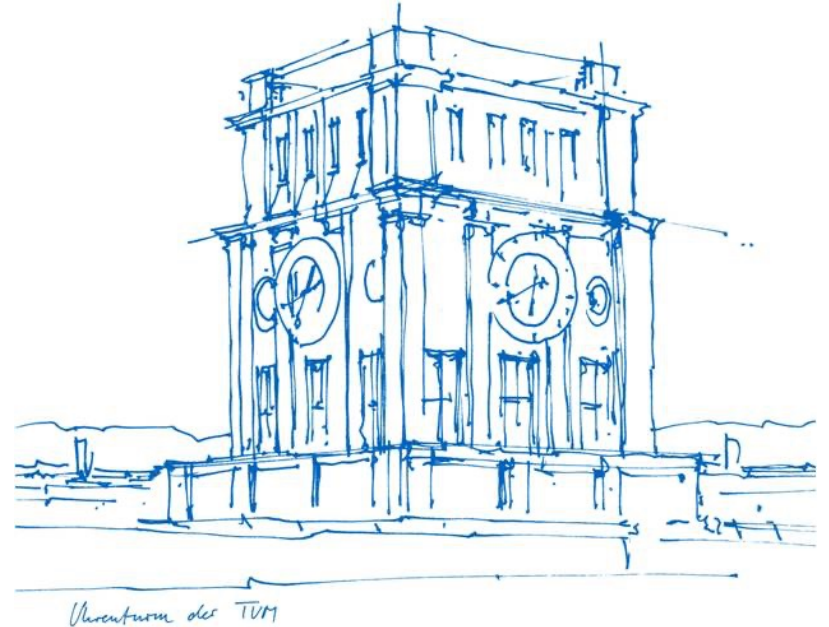


Framework For 3D Reconstruction Understanding

Vy Hong

Osama Salem

Munich, 30. January 2026



Phase 1: Comparative Performance Analysis

Objective: Benchmark state-of-the-art visual navigation paradigms.

- **Hybrid:** MAST3R
- **Feed Forward:** MapAnything, VGGT
- **Self-Supervised:** SPFSplat, RayZer

Phase 2: Mechanistic Investigation

Objective: Isolate MapAnything to analyze internal feature dynamics and robustness.

- **Ablation Studies** (Input contributions)
- **Sensitivity & Robustness**
- **Interpretability Analysis**

Related Work: MAST3R

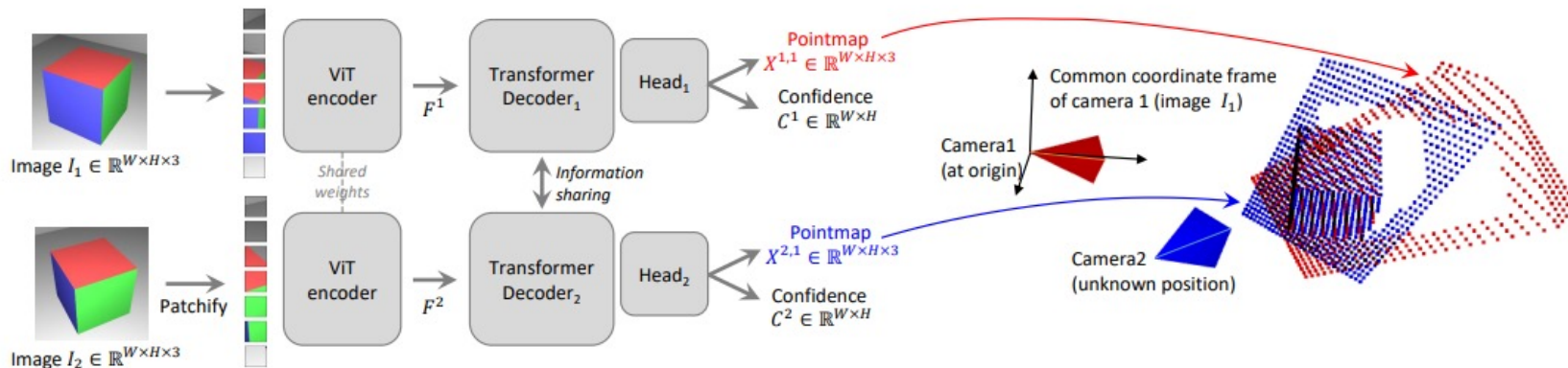
ViT Based Approach: Processes image pairs to predict dense **3D pointmaps** and confidence scores.

Geometric Pipeline:

- **Intrinsics:** Robust estimation via **Weiszfeld** algorithm.
- **Pose:** Solved via **PnP-RANSAC**.

Global Optimization: Registers multiple views by constructing a pairwise graph and applying Global Bundle Adjustment.

Limitation: Pairwise processing and iterative Bundle Adjustment result in high computational cost, limiting real-time performance.

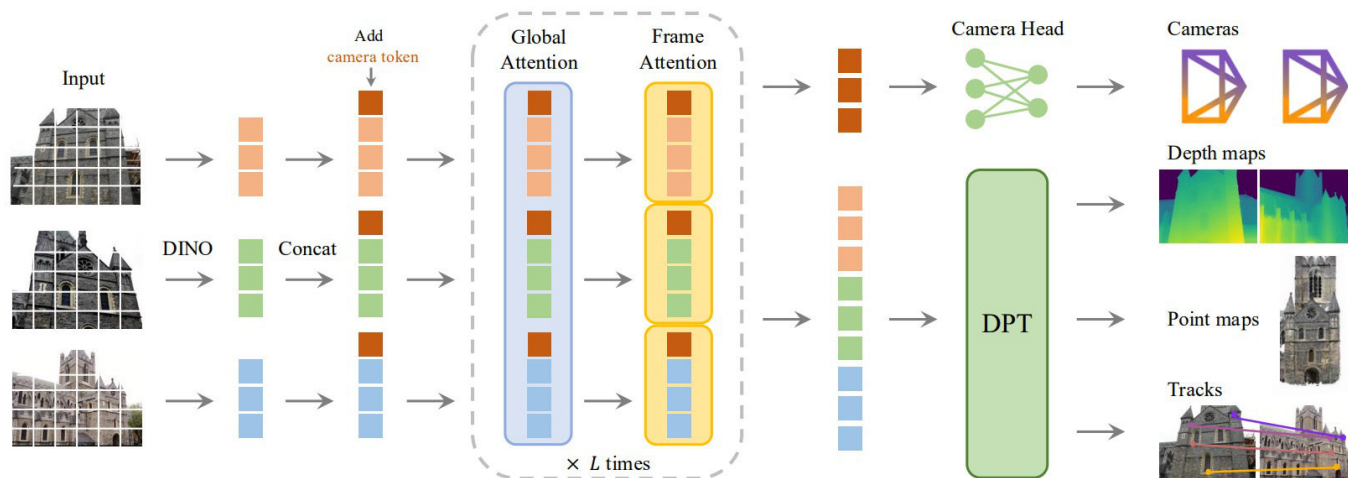


Related Work: VGGT

Architecture: Encodes frames into tokens (DINOv2) and applies **interleaved Global/Frame Attention** in a single forward pass, replacing slow iterative optimization.

Joint Output: Simultaneously regresses camera poses, depth, and point clouds for geometric coherence.

Input Rigidity: Strictly **RGB-only**. The architecture cannot leverage available geometric priors (e.g., sensor depth, known intrinsics) to improve robustness.



VGGT: Visual Geometry Grounded Transformer

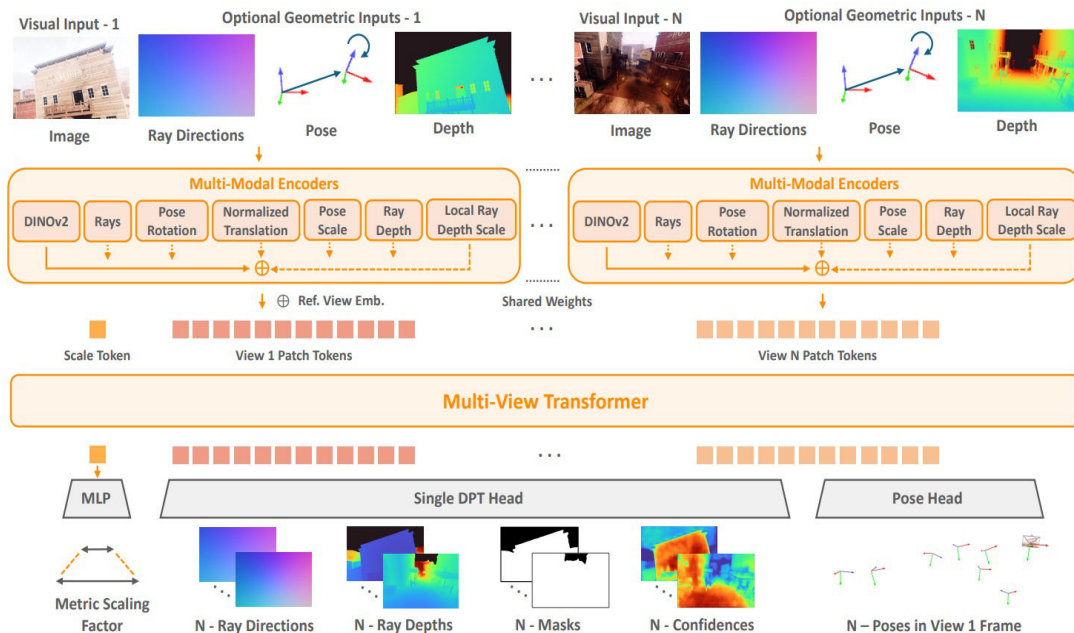
<https://vgg-t.github.io/>

Related Work: MapAnything

Multi-Modal Input: Fuses RGB (**DINOv2**) with optional geometric priors (e.g., sensor depth, ray directions) for adaptability.

Joint Prediction: A Multi-View Transformer simultaneously outputs metric depth, confidence, and camera poses

Supervision Bottleneck: Training relies on **COLMAP** for ground-truth poses. This process is slow, expensive, and error-prone, preventing scalable self-supervised learning on unposed video.



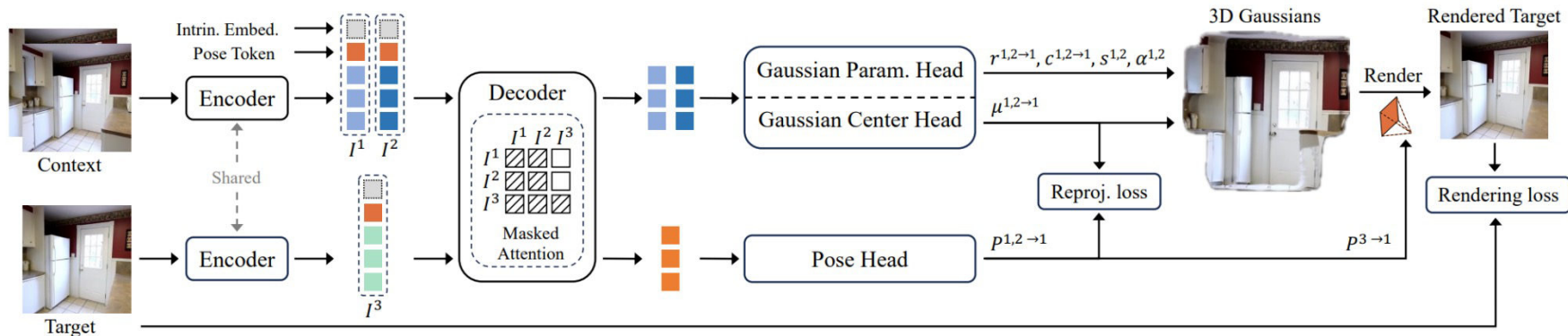
Related Work: SPFSplat

Asymmetric Transformer: Target views cross-attend to Context views for localization, while Context inputs focus on generating 3D Gaussian Splats.

Dual Output: Simultaneously regresses camera poses (for all views) and scene geometry (from Context) relative to a canonical frame.

Photometric Supervision: Eliminates the need for pre-computed COLMAP poses.

End-to-End Learning: Minimizes **re-rendering loss** (pixel difference between rendered Splats and target image) to jointly optimize geometry and pose.



Efficient Self-Supervised Pose-Free 3D Gaussian Splatting from Sparse Views

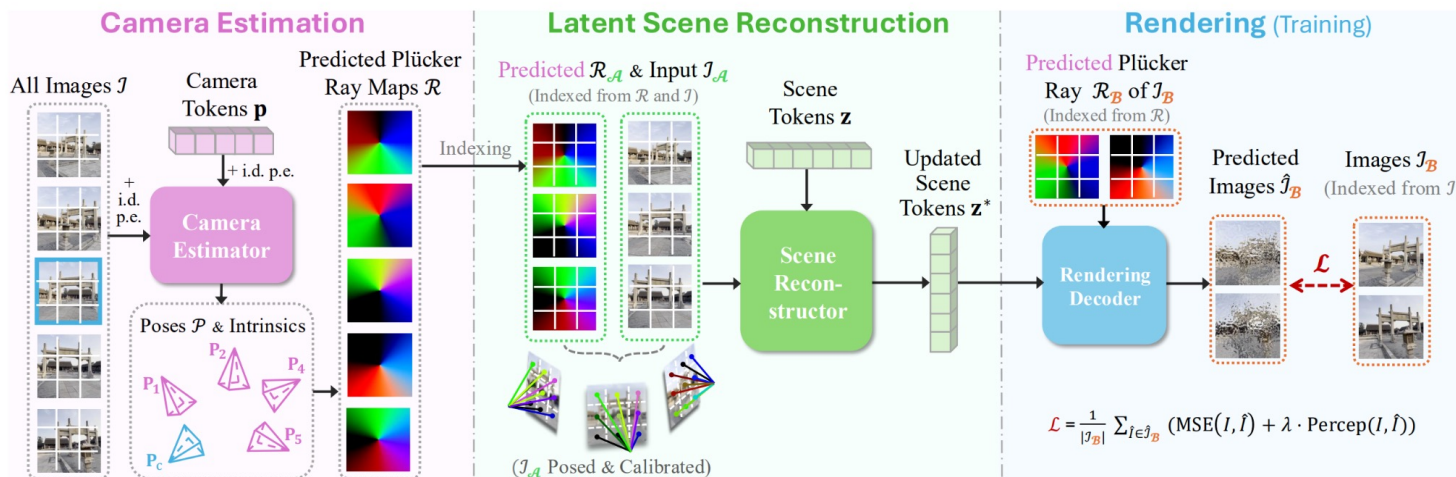
<https://ranrhuang.github.io/spfsplatv2/>

Related Work: RayZer

Latent Reconstruction: Encodes the scene into latent tokens rather than explicit geometry, enabling a compact, structure-free representation.

Ray-Based Estimation: Predicts cameras as Plücker Ray Maps which index the latent tokens to retrieve view-dependent features.

Uninterpretable Poses: Predicted poses are not constrained to SE(3). Because the model outputs implicit ray mappings rather than rigid rotations and translations. <https://github.com/hwjiang1510/RayZer/issues/1>



Phase 1 - Comparative Performance Analysis

Core Objective To systematically benchmark which architectural paradigm yields optimal 3D reconstruction and pose estimation performance on **sparse, unordered image sets**.

Experimental Setup

Dataset: ScanNet++ (High-fidelity, metric, static indoor scenes).

Sampling: Evaluated across {2, 4, 8, 16} views.

Task: Joint recovery of camera poses and 3D geometry from unposed RGB inputs.

Evaluation Metrics

Camera Pose: Trajectory Drift (ATE) & Rotation Accuracy (AUC @ 5°).

3D Geometry: Point Cloud Position Error & Depth Map Consistency.

Global Scale: Absolute Relative Error (AbsRel) to measure metric scale recovery.

Phase 1 - Quantitative Results

Geometric Priors Win: MapAnything dominates all settings (>74% accuracy). Integrating sensor depth guarantees stability, even with just 2 views.

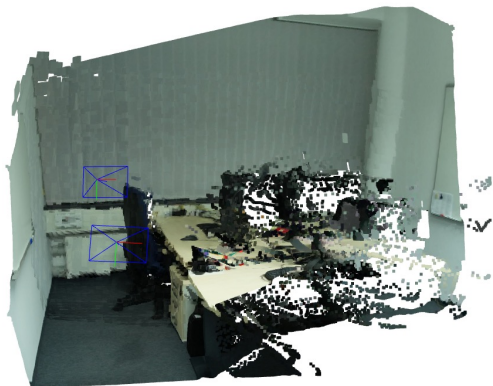
The Photometric Trap: Self-supervised methods achieve near-zero pose accuracy. Minimizing photometric loss alone does not guarantee correct SE(3) geometry.

Training Context Matters: MAST3R lags at N=2. This proves that training on **long multi-view sequences** builds stronger geometric priors than pairwise training, allowing better generalization to sparse pairs.

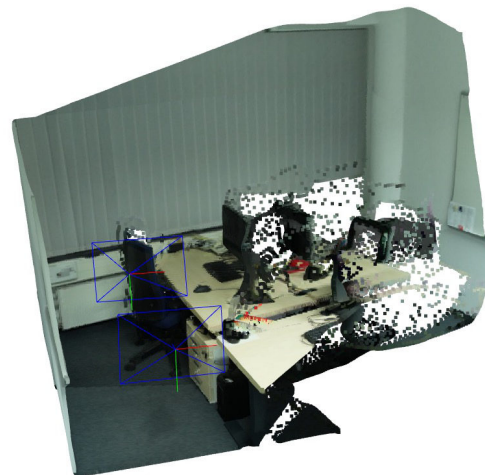
#Views	Method	Pose Accuracy		Scale	Point Accuracy		Depth Accuracy		Ray Error
		ATE RMSE	AUC@5	Abs Rel	Abs Rel	Inlier Ratio	Abs Rel	Inlier Ratio	Degrees
2	RayZer	0.3049	0.00%	-	-	-	-	-	22.7088
	SPFSplat	0.1531	0.00%	0.4530	0.4085	13.01%	0.2083	11.94%	-
	MASt3R	0.0602	12.40%	0.1249	0.2223	24.21%	0.1772	14.65%	7.7124
	VGGT	0.0197	44.80%	0.4890	0.0847	55.18%	0.0860	31.14%	1.9323
	MapAnything	0.0064	78.80%	0.0217	0.0497	71.43%	0.0479	67.36%	0.2322
4	RayZer	1.0807	0.00%	-	-	-	-	-	34.9026
	SPFSplat	0.2631	0.93%	0.5599	0.5649	11.57%	0.2717	9.67%	-
	MASt3R	0.1025	12.93%	0.1228	0.2305	26.78%	0.2134	14.24%	7.4933
	VGGT	0.0337	46.67%	0.5288	0.0801	61.92%	0.0839	34.93%	1.6006
	MapAnything	0.0099	77.00%	0.0155	0.0425	77.80%	0.0428	68.82%	0.2356
8	RayZer	1.2634	0.05%	-	-	-	-	-	25.6562
	SPFSplat	0.3629	0.41%	0.6513	0.6925	9.58%	0.3577	7.63%	-
	MASt3R	0.1520	15.83%	0.1450	0.3153	28.46%	0.2830	18.02%	5.5966
	VGGT	0.0389	42.99%	0.5577	0.0751	64.87%	0.1032	32.87%	1.5391
	MapAnything	0.0115	75.63%	0.0139	0.0422	80.89%	0.0559	71.82%	0.2271
16	RayZer	1.2595	0.00%	-	-	-	-	-	26.4189
	SPFSplat	0.3882	0.33%	0.7039	0.7593	9.07%	0.4237	6.36%	-
	MASt3R	0.1135	24.31%	0.1220	0.2295	38.33%	0.2473	24.09%	4.4116
	VGGT	0.0440	37.61%	0.5886	0.0697	67.26%	0.0997	28.18%	1.5832
	MapAnything	0.0119	74.17%	0.0114	0.0367	85.76%	0.0471	74.05%	0.2251

Phase 1 - Qualitative Results

MASt3R



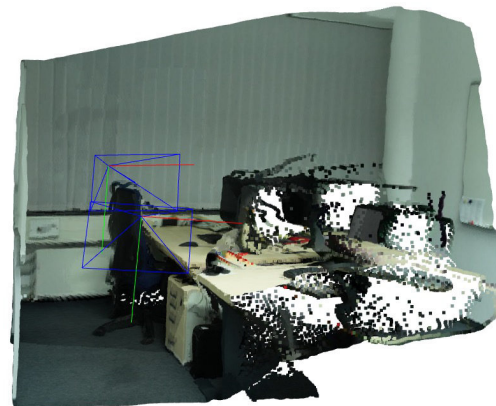
VGGT



SPFSplat



MapAnything



Phase 2 - Mechanistic Analysis of MapAnything

Motivation: Why this Deep Dive?

Universal Design: MapAnything fuses RGB with optional geometric priors (e.g., sensor depth, known poses).

The Goal: Phase 1 proved *performance*. Phase 2 explains *mechanism* by isolating the contribution of each input type.

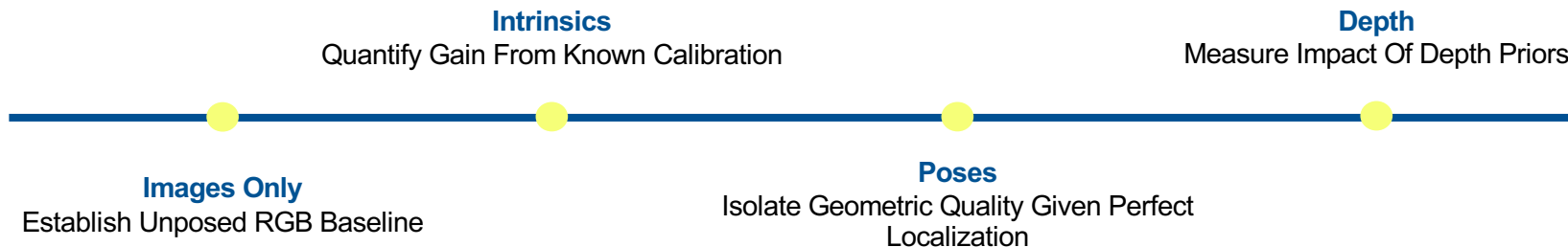
Core Objectives

Quantitative Ablation: How much do specific priors (Depth, Poses) improve the RGB baseline?

Robustness Analysis: Is the model stable when these geometric inputs are noisy or corrupted?

Interpretability: Do attention maps confirm the model physically "attends" to these cues?

Phase 2.1 - Input Ablation Study



Experimental Setup

Source: ScanNet++ (50 Sequences).

Input: Fixed at 16 Views for fair comparison

Evaluation Metrics

Pose: Drift (ATE) & Rotation Accuracy (AUC).

Geometry: Depth Error (AbsRel) & Inlier Ratio.

Scale: Absolute Relative Error.

Ray: Angular Error (deg).

Phase 2.1 – Quantitative Results

The "Trust" Factor: The model actively utilizes every provided modality. It does not ignore priors but effectively fuses them to refine the corresponding output.

Poses Lock Trajectory: Providing poses (I+P) boosts Pose Accuracy from 74% to 96%. The model trusts the localization prior, overriding its internal estimator to fix the path.

Intrinsics Sharpen Rays: Adding calibration (I+K) drops Ray Error significantly (0.22° to 0.16°). Even if the trajectory doesn't change, the internal ray geometry becomes much more consistent.

Depth Drives Structure: Dense depth (I+D) drastically improves geometric metrics (Depth Accuracy 74% to 94%). The model anchors its visual features to the provided surface.

Method	Pose Accuracy		Scale	Point Accuracy		Depth Accuracy		Ray Error
	ATE RMSE	AUC@5	Abs Rel	Abs Rel	Inlier Ratio	Abs Rel	Inlier Ratio	Degrees
I	0.0119	74.17%	0.0114	0.0367	85.76%	0.0471	74.05%	0.2251
I + D	0.0082	81.57%	0.0044	0.0180	95.16%	0.0120	94.63%	0.2406
I + D + K	0.0082	82.22%	0.0048	0.0175	95.16%	0.0109	94.92%	0.1696
I + K	0.0118	75.27%	0.0119	0.0364	85.79%	0.0466	74.60%	0.1650
I + P	0.0033	96.48%	0.0045	0.0303	88.63%	0.0450	77.19%	0.2572
I + D + P	0.0021	97.20%	0.0024	0.0109	97.01%	0.0112	95.05%	0.2527
I + P + K	0.0032	96.66%	0.0046	0.0297	88.73%	0.0444	77.49%	0.1649
I + D + P + K	0.0021	97.28%	0.0025	0.0101	97.03%	0.0096	95.39%	0.1579

Conclusion: MapAnything is a true generalist. It improves exactly where you give it help (Poses = Trajectory, K = Rays, Depth = Geometry).

Phase 2.2 – Sensitivity & Robustness

1. Photometric Perturbation (RGB)

Method: Add Gaussian noise to input images.

Goal: Test if pose estimation survives sensor noise and feature corruption.

2. Geometric Corruption (Depth)

Input Structure: Depth is split into a Normalized Map (structure) and a Scalar Token (metric scale).

Method:

- **Structure:** Apply checkerboard and random masking to the normalized map.
- **Scale:** Perturb the scalar token values.

Goal: Verify if the model can inpaint missing structure and correct metric scale errors.

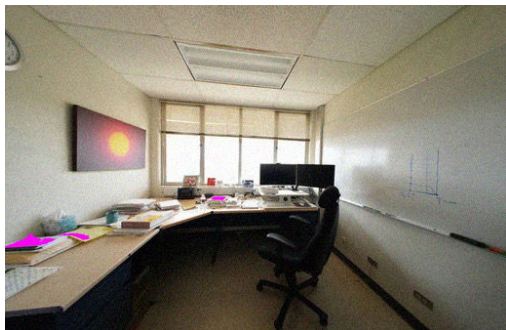
3. Calibration Noising (Rays)

Method: Apply relative noise to intrinsic matrix values.

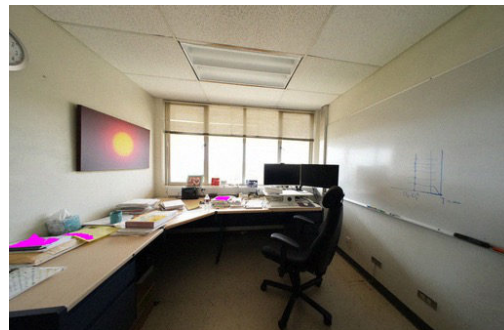
Goal: Evaluate resilience to partial loss of calibration data.

Phase 2.2 – Photometric Perturbation

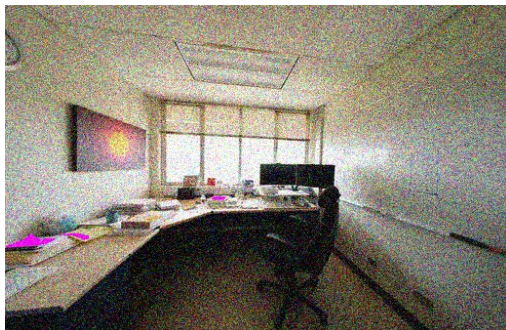
$\sigma = 15$



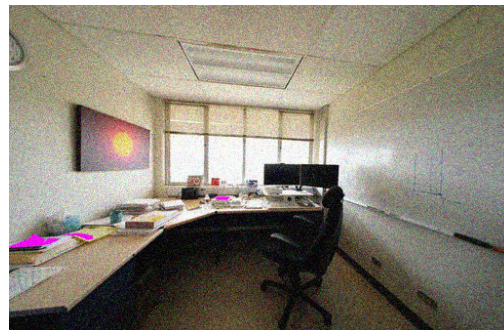
$\sigma = 5$



$\sigma = 65$



$\sigma = 35$

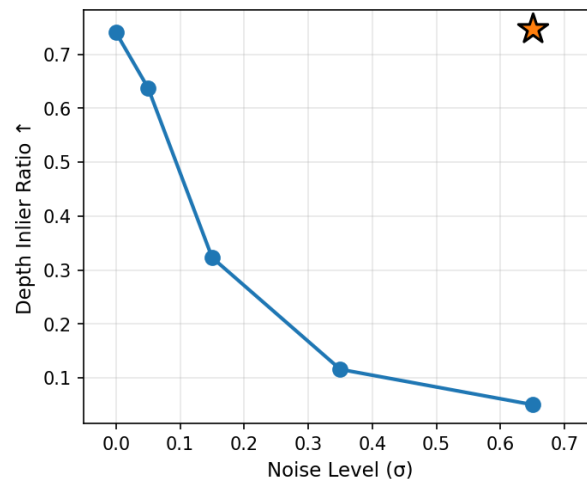
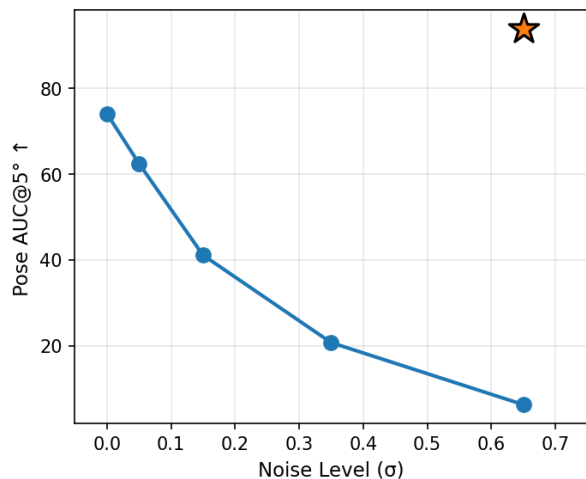
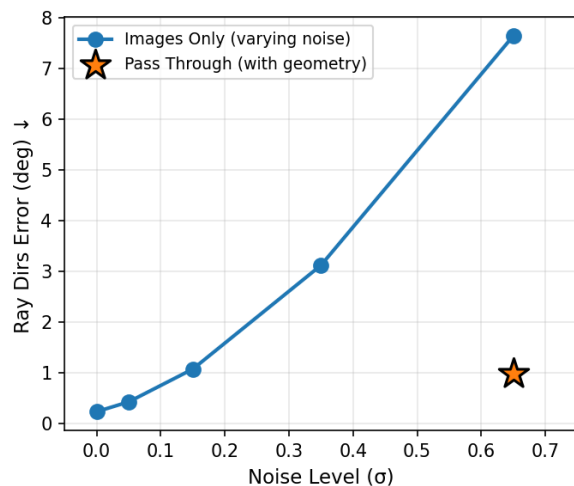


Phase 2.2 – Photometric Perturbation

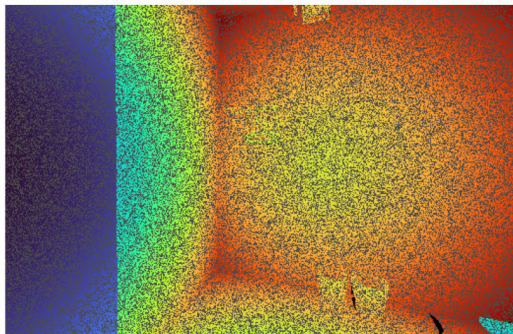
Visual Fragility: Purely visual estimation collapses under high noise.

Geometric Anchor: Providing priors effectively "overrides" the corrupted visual features even when images are unrecognizable.

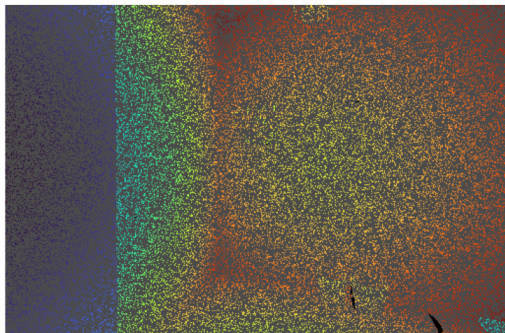
Conclusion: The model demonstrates a dynamic attention mechanism, shifting reliance from corrupted RGB to stable Geometry.



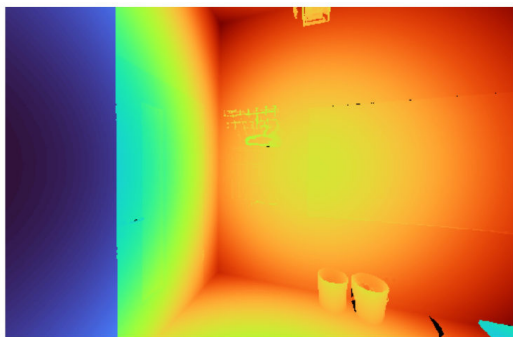
Phase 2.2 – Depth Masking



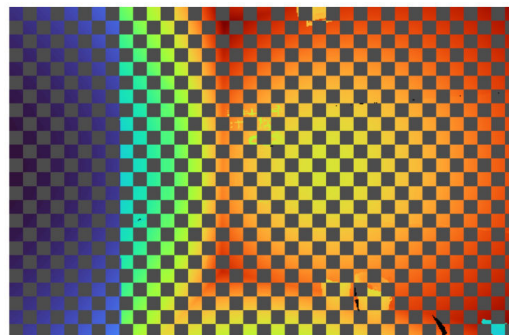
$\sigma = 40$



$\sigma = 80$



$\sigma = 20$



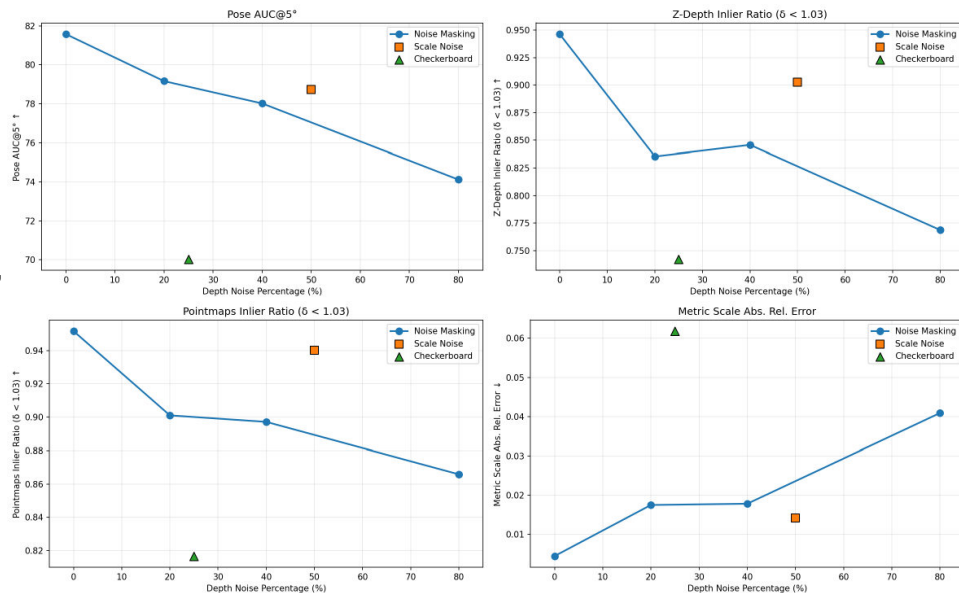
Checkerboard

Phase 2.2 – Depth Masking

Scale Sensitivity: Corrupting the single **Scale Token** (Orange Square) causes an immediate drop in accuracy across all metrics. The model relies heavily on this scalar value for global metric alignment, even if the feature map is perfect.

Smooth Convergence: As random masking increases (Blue Line), performance decays smoothly and **converges to the RGB-only baseline** (Pose AUC ~74%). The model gracefully hands over control from "Geometry" to "Vision" as data vanishes.

Structured "Poison": Checkerboard Masking (Green Triangle) degrades performance (~70% AUC) **below** the RGB-only baseline. Structured, artificial occlusion acts as adversarial noise, misleading the model more than simply having no data at all.



Phase 2.2 – Calibration Noising

Method

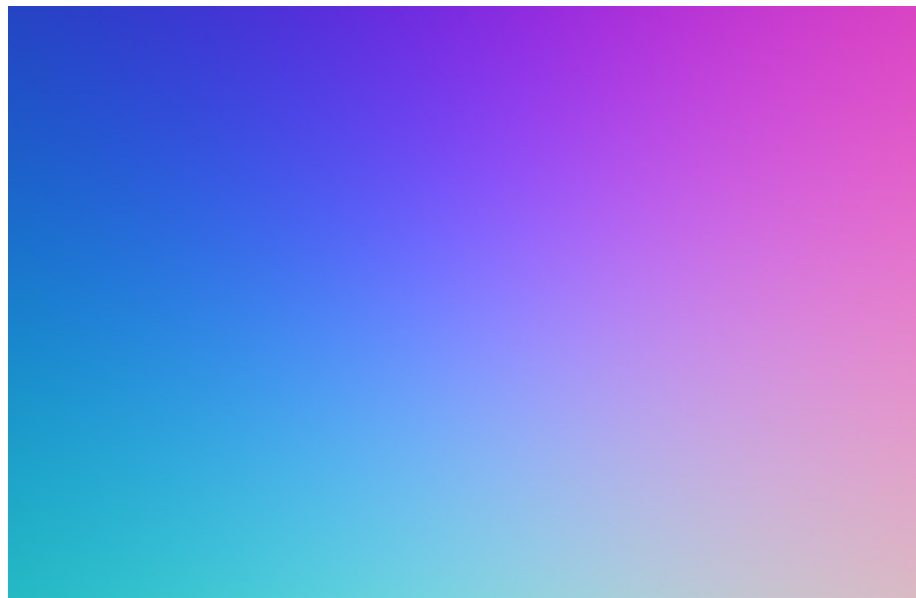
- **Inject Noise:** Apply relative noise ($\sigma \in [0, 50\%]$) to the focal lengths (f_x, f_y) of the intrinsic matrix K .

Objective

- **Simulate "Lying":** Test robustness to incorrect camera specs (e.g., wrong FOV), rather than missing data.

Hypothesis

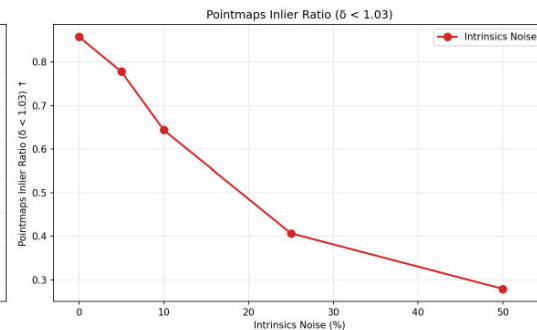
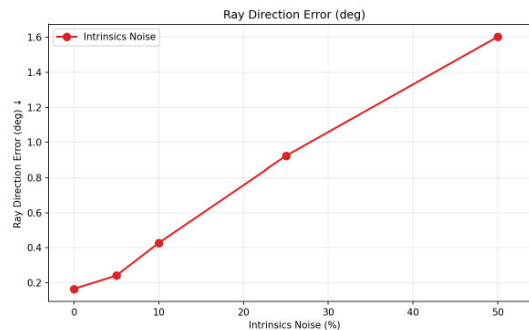
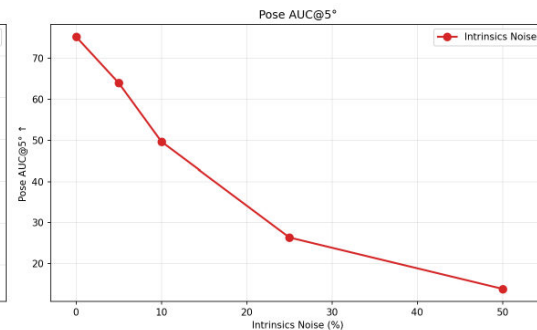
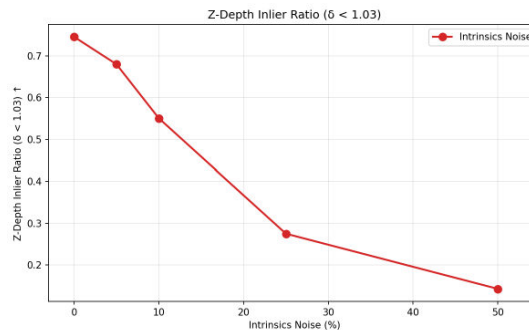
- **Blind Trust vs. Vision:** Does the model follow the misleading geometric instructions, or can it "self-correct" using visual cues?



Phase 2.2 – Calibration Noising

Bad Priors > No Priors: Incorrect calibration is more harmful than no calibration at all. Adding just **10% noise** drops Pose AUC to **~50%**, significantly below the uncalibrated baseline (74%).

No Safety Net: The model blindly trusts the provided intrinsics without validation. Unlike depth masking, there is no recovery mechanism; performance collapses linearly as the calibration error increases.



Phase 2.3 – Interpretability

1. Attention & Correlation Maps

Goal: Investigate what parts of the image the model attends to.

2. Feature Maps

Goal: Investigate the internal representation.

3. Gradient CAM & SHAP

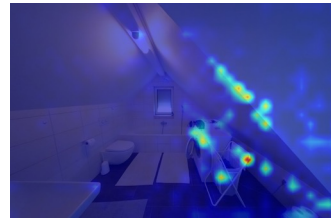
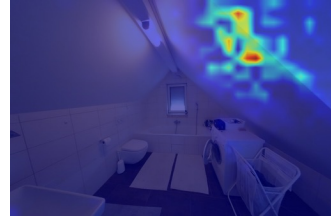
Goal: Visualize which features influence the prediction.

4. Saliency

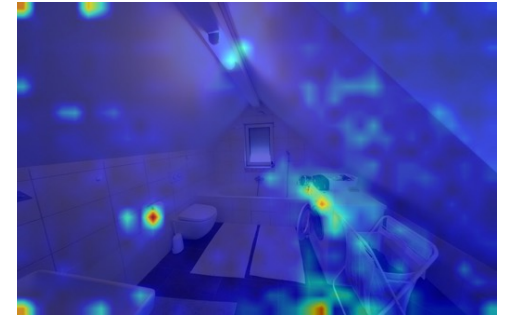
Goal: Isolate modality influence on the prediction.

Phase 2.3.1 – Attention

What patches attend to



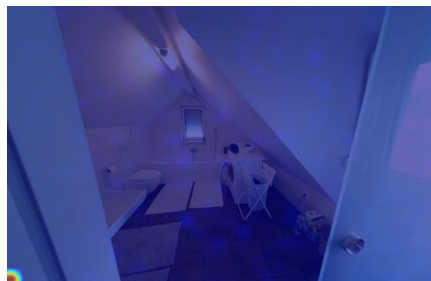
Averaged per View



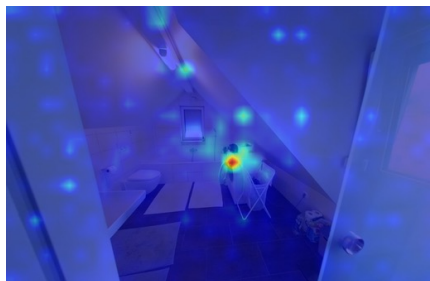
Phase 2.3.1 – Attention Across Layers

Global Attention Layers

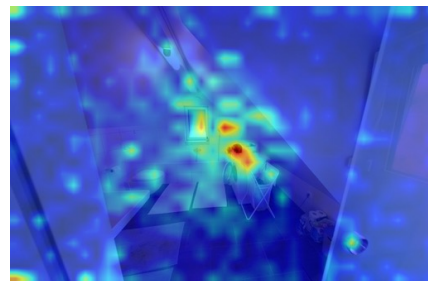
Block 0



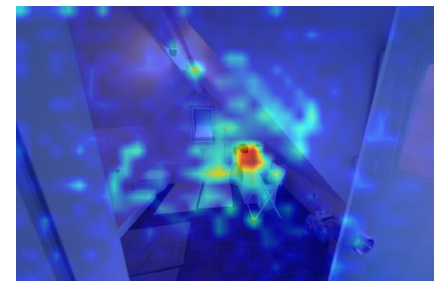
Block 6



Block 12



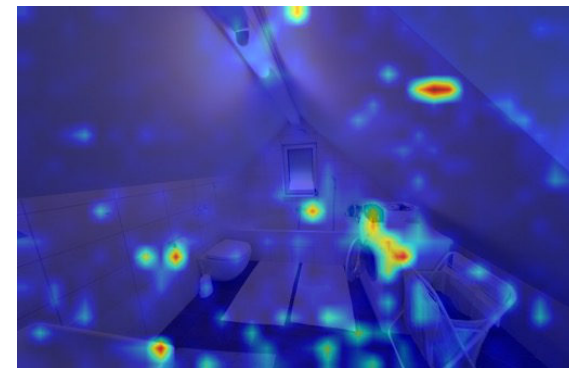
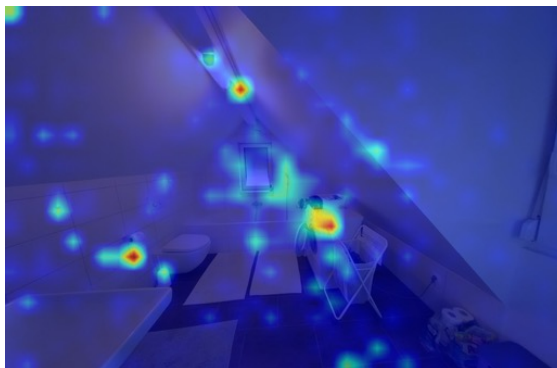
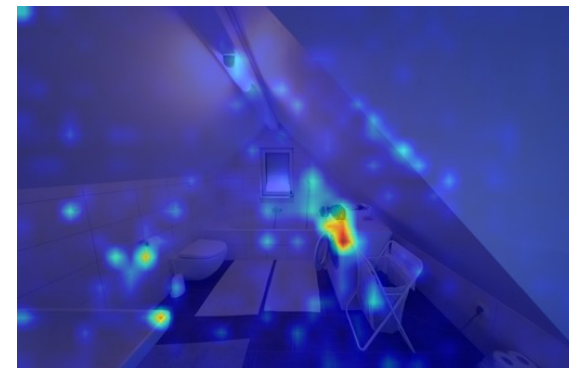
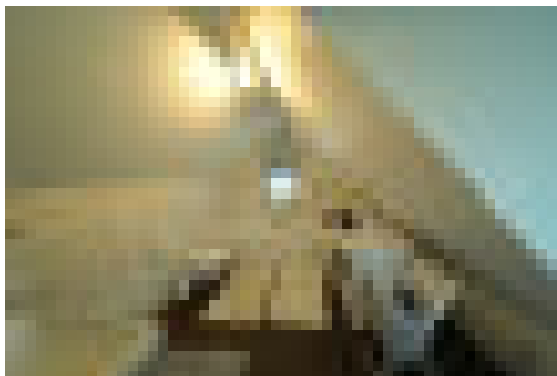
Block 18



Multiview Transformer

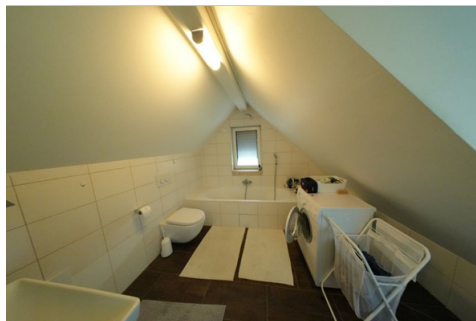
Phase 2.3.1 – Attention Across Views

Landmark Attention: The model attends to similar objects across views



Phase 2.3.1 – Correlation Map

Correlation Maps:
Attention without Softmax



Baseline

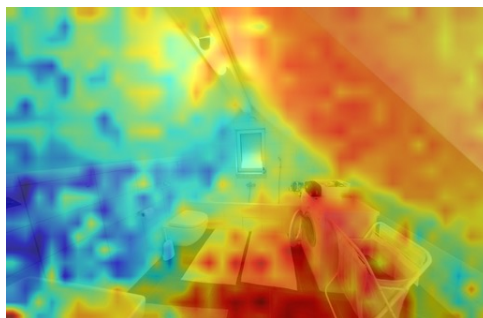
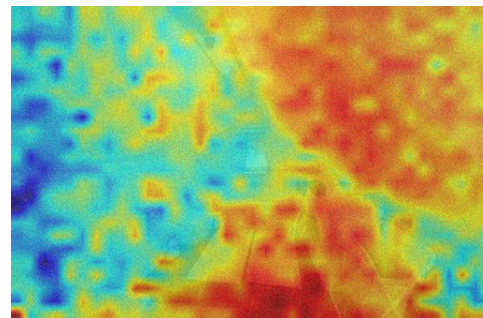
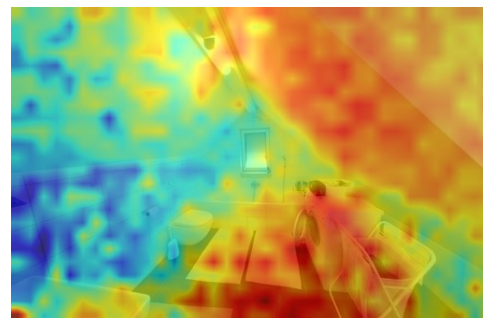


Image Noised



Depth Checkerboard

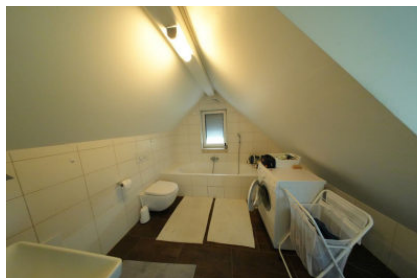


Checkerboard pattern visible

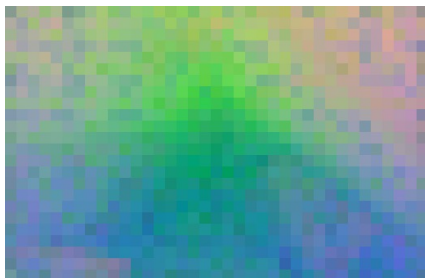
Phase 2.3.1 – Features Across Layers

Global Attention Layers

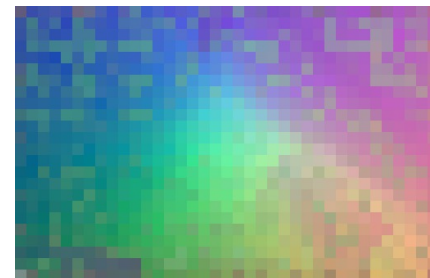
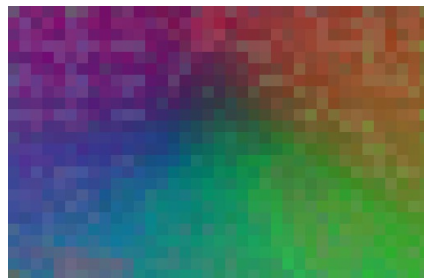
Block 0



Block 8

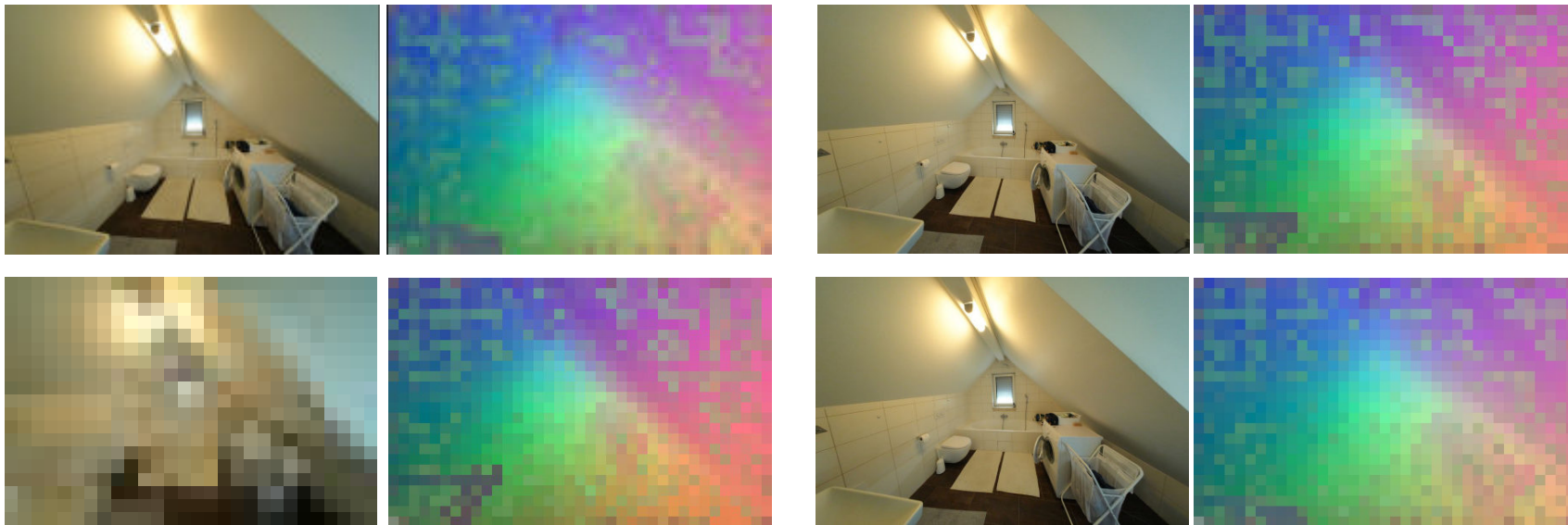


Block 22

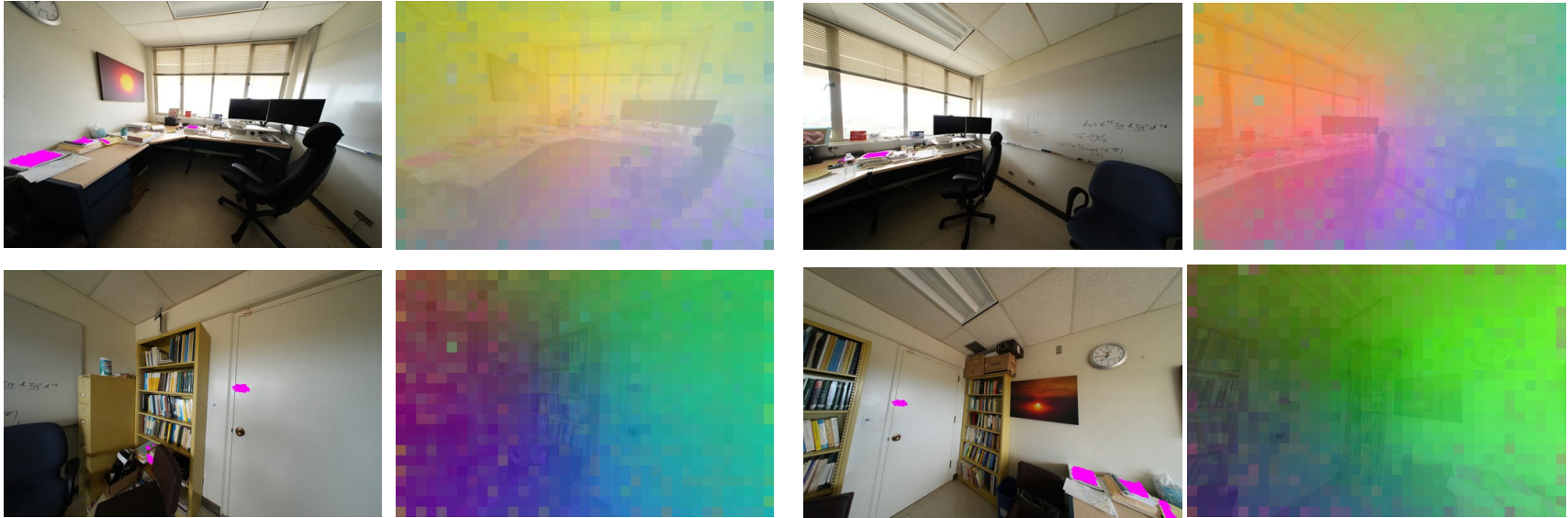


Multiview Transformer

Phase 2.3.2 – Feature Consistency



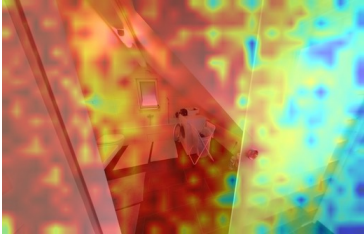
Phase 2.3.2 – Feature Consistency



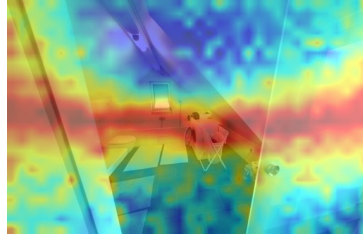
Features Representation: Consistent for across views

Phase 2.3.2 – Gradient CAM – Points

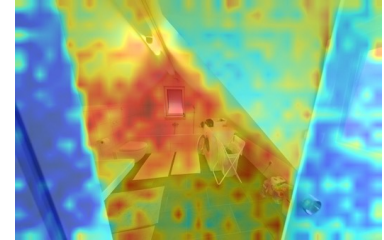
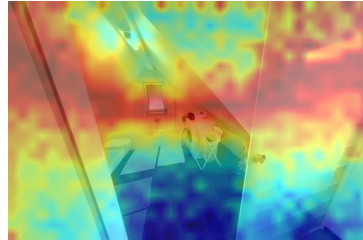
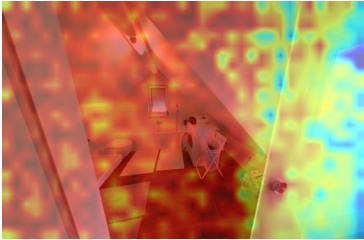
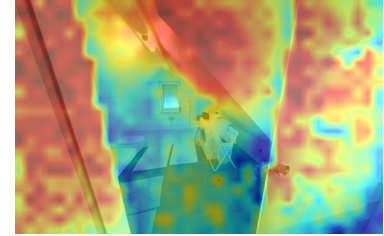
X



Y

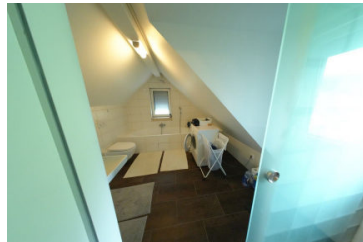


Z

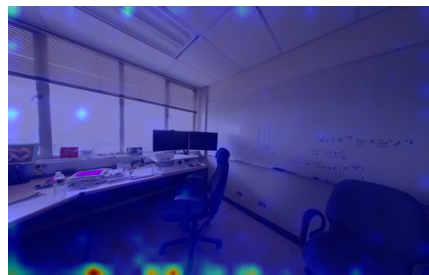


Pattern distinguishable

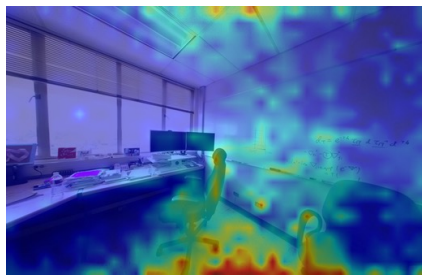
-> Each dimension is dependent on different parts of the view



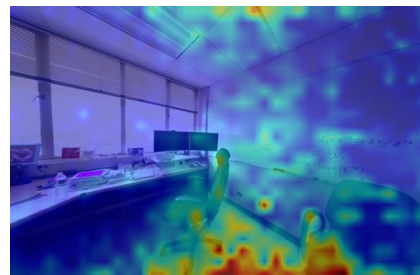
Phase 2.3.4 – Scale Analysis



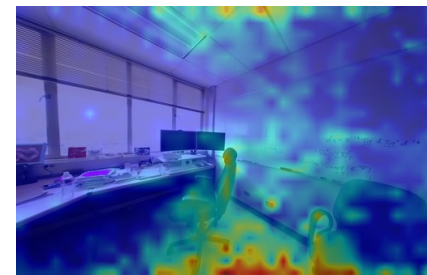
RGB + Pose



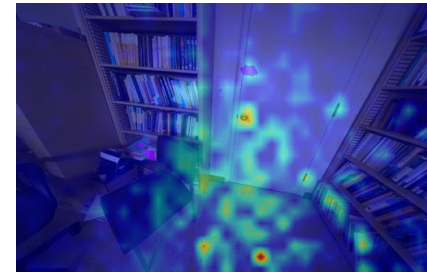
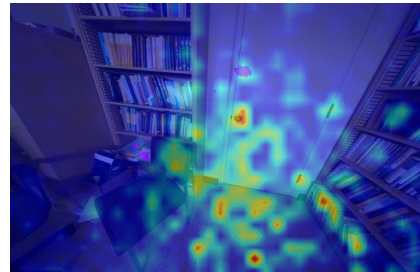
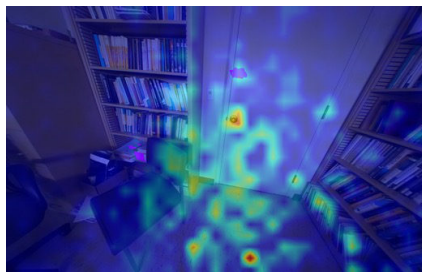
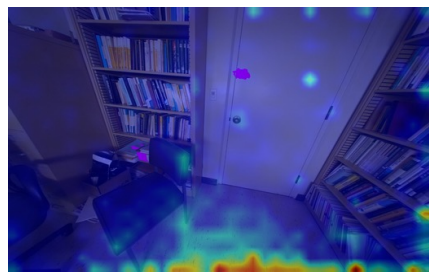
RGB + Intrinsics



RGB + Depth



RGB Only



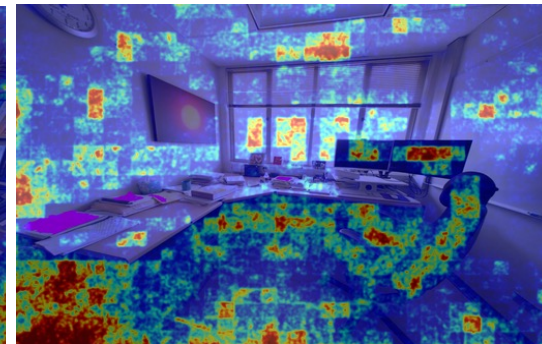
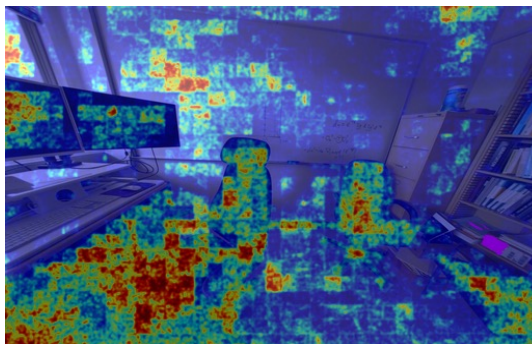
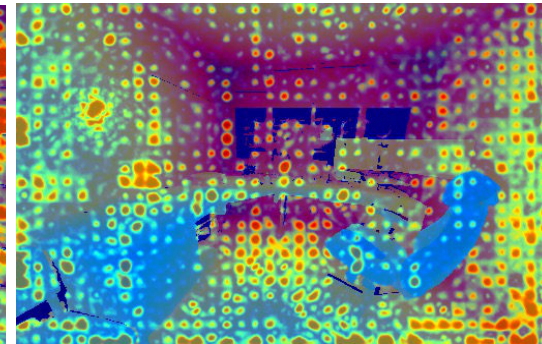
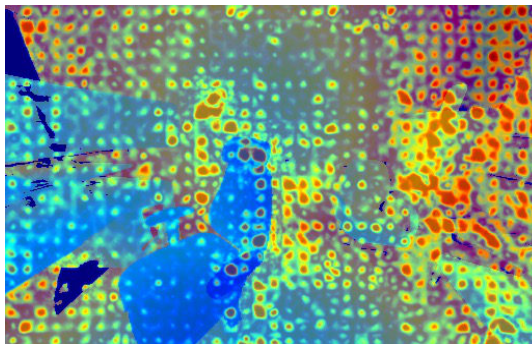
Scale Sensitivity: Scale is only sensitive to RGBs & Poses

Phase 2.3.4 – Depth Saliency Analysis

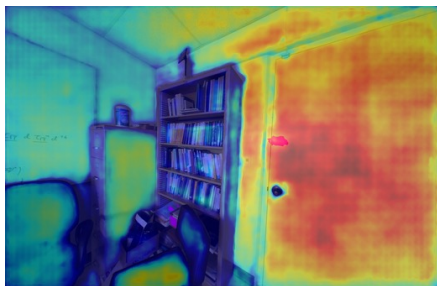
RGB Saliency: Gradient magnitude is significantly higher in **textureless regions**, indicating sensitivity to color homogeneity.

Depth Saliency: Exhibits a contrasting pattern, with magnitudes peaking in areas of **minimal depth variation** (planar surfaces).

Complementary Fusion: The two modalities focus on distinct scene characteristics, providing **complementary signals** that improve overall reconstruction robustness.



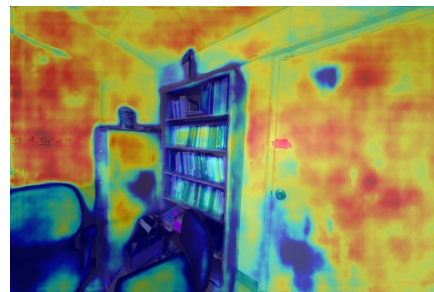
Phase 2.3.4 – Depth Analysis



Pass Through



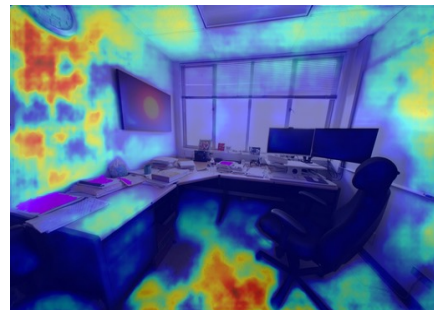
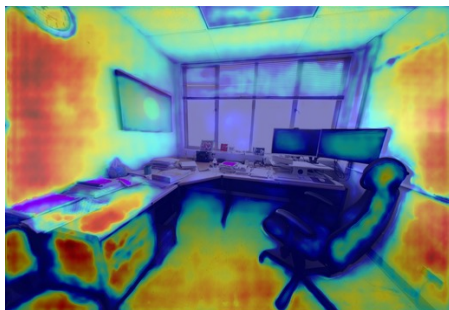
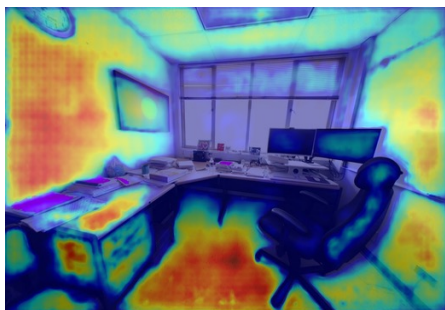
0% Masking



60% Masking



Checkerboard



Confidence Map: Depth confidence maintains same input depth pattern but other modes boost confidence

Lessons Learned



Multi-View Training Generalizes: VGGT/MAPA beating MAST3R on just 2 views proves that training on long sequences builds stronger priors than pairwise training alone.

Photometric Loss is Insufficient: Self-supervised minimizing of pixel error does not guarantee accurate SE(3) pose recovery or geometry.

Complementary Attention: RGB and Depth attend to distinct features (texture vs. planar structure), making their fusion essential for robustness.

The "Blind Trust" Risk: MapAnything lacks a validation mechanism for geometric inputs; providing incorrect intrinsics is worse than providing none at all.

Future Work



Confidence-Aware Gating: Implement mechanisms to estimate uncertainty in geometric inputs (intrinsic/poses) to reject "lying" priors rather than blindly trusting them.

Hybrid Optimization: Investigate using feed-forward predictions (like MapAnything) to warm-start self-supervised methods (like SPFSplat) to solve the pose convergence issue.

Dynamic Scene Analysis: Extend the benchmarking from static ScanNet++ scenes to dynamic video sequences to test temporal consistency.

Non-Pinhole Generalization: Investigate performance on fisheye or highly distorted inputs to determine if the model's "blind trust" in intrinsic leads to failure when the standard pinhole assumption is violated.

Thank you for your Attention

Vy, Osama

Munich, 30. January 2026

