Supplementary Material for
Consistent Yet Wrong: Evidence Insensitivity in Spatial Vision-Language Models

1. Additional Results

We move additional Qwen diagnostic plots here to reduce redundancy in the main Results while keeping representative examples available. Model names follow the main paper: Qwen2-VL [8], InternVL [6], MiniCPM-V [7], LLaVA [5], VILA [4], SpatialRGPT [3], DepthLM [2], and ZoeDepth [1].

ZoeDepth variants in the supplementary tables use different distance/statistics settings: we aggregate ZoeDepth depth per region with a statistic (median or trimmed mean), then compute distance by either depth_diff (absolute difference of region depths) or pseudo3d (unproject region centers with intrinsics and take 3D distance, falling back to depth_diff when intrinsics are missing).

1.1. ViewDiag Construction Overview

ViewDiag is built from Hypersim, ScanNet, and KITTI360 using dataset-provided masks, depth, and camera metadata. We apply region validity gates (minimum area, depth support, depth consistency), pair constraints (IoU/separation), and track stability gates (center drift and area ratio) before keeping tracks with sufficient views. Key parameters are dataset-specific: minimum mask area (1024 px indoor; 128 px KITTI360), minimum valid depth pixels (60 indoor), depth consistency band $\epsilon = 0.1$ m and unimodality ratio $\sigma/\mu \leq 0.40$, pair overlap IoU ≤ 0.20 (indoor), center-std caps (0.20 indoor, 0.22 KITTI360), area-ratio caps (4.0 indoor, 4.5 KITTI360), and minimum views per track (4/3/2 for Hypersim/ScanNet/KITTI360).

1.2. Additional Qwen Diagnostics

These plots elaborate the collapse signatures shown in the main text. The residual-vs-GT diagnostic (log bin count) highlights long-range errors, the GT–prediction density highlights systematic bias, and the histogram and mode profile show concentration around a small set of favored outputs (Figures 1–3).

1.3. Calibration Transfer Across Datasets

We test how a single L2 scale fitted on one dataset transfers to another, highlighting scale instability under domain shifts (Table 1).

Algorithm 1 VIEWDIAG Construction Pipeline

Require: Scene frames with masks, depth, and camera metadata; thresholds; V views; pair budget P
Ensure: VIEWDIAG samples $\mathcal{D}$

$\mathcal{D} \leftarrow \emptyset$
for all scenes s **do**
  $\mathcal{W} \leftarrow \textsc{SelectWindow}(s, V)$
  $\mathcal{W} \leftarrow \textsc{RemapImages}(\mathcal{W})$
  $\mathcal{C} \leftarrow \textsc{BuildCandidates}(\mathcal{W})$
  // min area, min depth pixels, depth band ϵ,
  // unimodality σ/μ, optional semantic purity / allowlist
  $\mathcal{T} \leftarrow \textsc{AssociateTracks}(\mathcal{C})$
  $\mathcal{T} \leftarrow \textsc{FilterTracksByViews}(\mathcal{T})$
  $\Pi \leftarrow \textsc{SamplePairs}(\mathcal{T}, P)$
  for all $(t_i, t_j) \in \Pi$ **do**
    for all views $v \in \mathcal{W}$ **do**
      $(r_i^v, r_j^v) \leftarrow \textsc{FetchRegions}(t_i, t_j, v)$
      if $\textsc{FirstValidView}(t_i, t_j, v)$ **then**
        $\textsc{ApplySpatialFilter}(r_i^v, r_j^v)$
  // IoU and minimum-separation constraints
      end if
    end for
    $\textsc{CheckBBoxStability}(t_i, t_j)$
  // center standard deviation and area-ratio constraints
    $d_{\text{gt}} \leftarrow \textsc{Median3DCenterDistance}(t_i, t_j)$
    $x \leftarrow \textsc{SerializeQA}(t_i, t_j, d_{\text{gt}})$
    $\textsc{OptionallyAttachPixelsAndIntrinsics}(x)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{x\}$
  end for
end for
return $\mathcal{D}$

Large transfer gaps indicate that a single global scale does not generalize cleanly across domains.

1.4. Error vs. Depth Bins

We report raw MAE across ground-truth depth bins to highlight long-range degradation (Table 2).

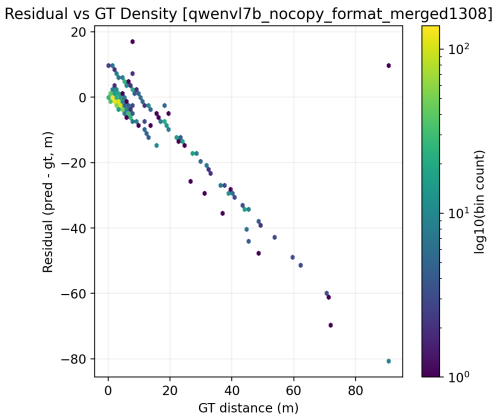


Figure 1. Residual vs GT (log bin count) (Qwen2-VL-7B). Error grows with range, highlighting evidence sensitivity failures at long distances.

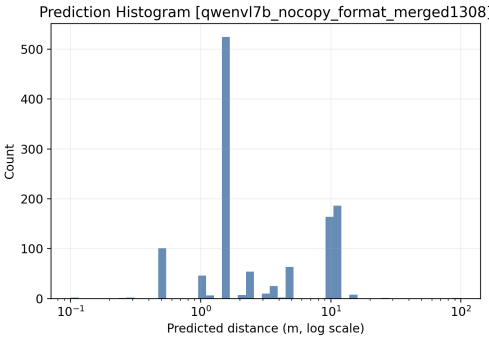


Figure 2. Prediction histogram (log bins) for Qwen2-VL-7B, illustrating mode concentration.

Errors grow sharply with range for all models, highlighting long-distance fragility.

1.5. Per-Dataset MAE and Concentration

We report per-dataset MAE and Top1Mass to show how accuracy and output concentration vary by domain (Table 3).

Indoor datasets are consistently easier, while KITTI360 shows both higher MAE and higher concentration.

1.6. Per-Dataset Consistency vs. Accuracy

We plot MAE against Top1Mass separately for each dataset to visualize collapse behavior under different domains (Figure 4).

1.7. Pair-Level Distributions

A *pair* is the two queried regions (Region [0], Region [1]) within one image; a *pair track* is that same region pair across multiple views. We aggregate predictions by pair (mean over views for each pair_id) to reduce view-level repetition and visualize distributional collapse at the pair level

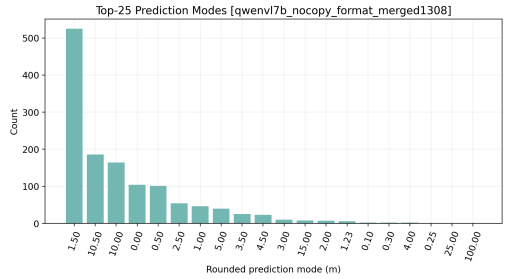


Figure 3. Example prediction mode profile (Qwen2-VL-7B) illustrating output concentration.

Model	Fit dataset	Test dataset	MAE(L2)
Qwen2-VL-7B	Hypersim	KITTI360	14.67
Qwen2-VL-7B	ScanNet	KITTI360	16.83
Qwen2-VL-7B	KITTI360	KITTI360	12.34
InternVL3-8B	Hypersim	KITTI360	37.72
InternVL3-8B	ScanNet	KITTI360	17.56
InternVL3-8B	KITTI360	KITTI360	16.97
MiniCPM-V-4.5	Hypersim	KITTI360	18.32
MiniCPM-V-4.5	ScanNet	KITTI360	17.23
MiniCPM-V-4.5	KITTI360	KITTI360	17.62

Table 1. Calibration transfer to KITTI360 using a single L2 scale fit on a source dataset. Large gaps indicate scale instability under domain shifts.

(Figures 5).

The pair-level plots show that concentration persists even after averaging across views.

1.8. Error vs. View-Range Bins

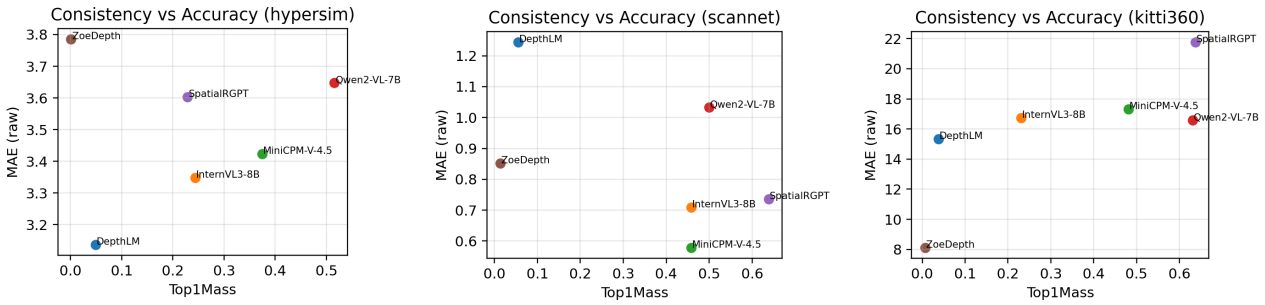
We report raw MAE stratified by frame-index range within each pair track (a coarse proxy for view-change magnitude) in Table 4.

Model	2–3	4–5	6–9	10+
DepthLM	10.47	31.73	7.32	4.95
InternVL3-8B	12.19	29.19	17.94	5.12
Qwen2-VL-7B	12.15	31.05	10.53	5.53
SpatialRGPT	13.51	37.48	12.71	6.49

Table 4. Raw MAE by frame-range bin (view-change proxy). Smaller bins are low-sample; the 10+ bin dominates coverage.

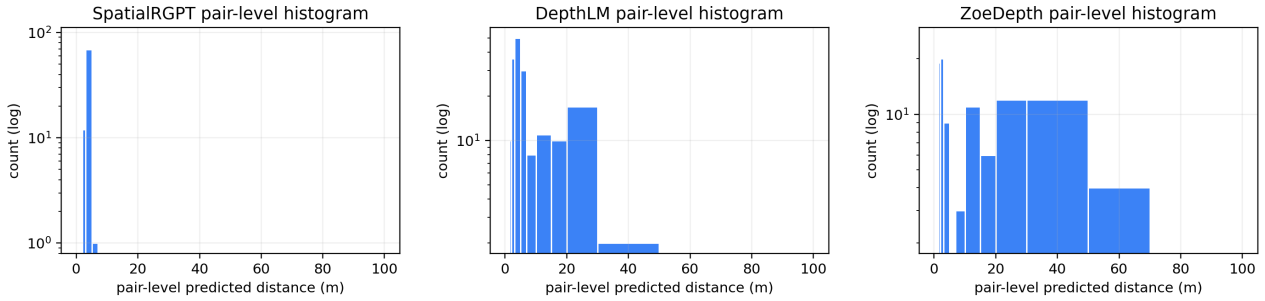
1.9. Held-Out L2 Calibration

We fit s^* on a random 50% split and report MAE on the held-out half (Table 5).



Hypersim ScanNet KITTI360

Figure 4. Per-dataset consistency vs. accuracy (MAE vs Top1Mass). Each point is a model. High Top1Mass with high MAE indicates evidence-insensitive collapse.



SpatialRGPT DepthLM ZoeDepth

Figure 5. Pair-level prediction histograms (log bins). Each pair id contributes one aggregated prediction.

Model	0–2	2–5	5–10	10–20	20–50	50–100
DepthLM	2.10	1.52	4.35	7.19	18.92	60.47
InternVL3-8B	0.68	1.56	6.74	9.85	19.90	60.35
Qwen2-VL-7B	0.94	2.05	4.34	7.31	24.31	61.52
SpatialRGPT	0.52	1.92	4.09	12.12	30.37	70.60

Table 2. Raw MAE by ground-truth depth bin (meters). Errors rise sharply in long-range bins.

Model	N_{train}	N_{test}	MAE(L2 holdout)
DepthLM	653	653	6.53
InternVL3-8B	654	654	6.37
Qwen2-VL-7B	654	654	6.38
SpatialRGPT	654	654	7.25

Table 5. Held-out L2-calibrated MAE using a 50/50 split for scale fitting.

Model	Dataset	N	MAE(raw)	Top1Mass
DepthLM	Hypersim	941	3.1361	0.0489
InternVL3-8B	Hypersim	941	3.3481	0.2434
MiniCPM-V-4.5	Hypersim	941	3.4235	0.3741
Qwen2-VL-7B	Hypersim	941	3.6474	0.5154
SpatialRGPT	Hypersim	941	3.6028	0.2285
ZoeDepth	Hypersim	941	3.7861	0.0011
DepthLM	KITTI360	293	15.3396	0.0375
InternVL3-8B	KITTI360	295	16.7228	0.2305
MiniCPM-V-4.5	KITTI360	295	17.3204	0.4814
Qwen2-VL-7B	KITTI360	295	16.5734	0.6305
SpatialRGPT	KITTI360	295	21.7776	0.6373
ZoeDepth	KITTI360	295	8.1039	0.0068
DepthLM	ScanNet	72	1.2446	0.0556
InternVL3-8B	ScanNet	72	0.7085	0.4583
MiniCPM-V-4.5	ScanNet	72	0.5779	0.4583
Qwen2-VL-7B	ScanNet	72	1.0324	0.5000
SpatialRGPT	ScanNet	72	0.7365	0.6389
ZoeDepth	ScanNet	72	0.8525	0.0139

Table 3. Per-dataset MAE(raw) and Top1Mass for key models. Top1Mass is the fraction of predictions at the most frequent output value.

References

[1] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 1

[2] Zhipeng Cai, Ching-Feng Yeh, Xu Hu, Zhuang Liu, Gregory Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. Depthlm: Metric depth from vision language models. *arXiv preprint arXiv:2509.25413*, 2025. 1

[3] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37: 135062–135093, 2024. 1

[4] Ji Lin, Hongxu Yin, Wei Ping, et al. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2405.04464*, 2024. 1

[5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 2023. 1

[6] InternVL Team. Internvl 2.5: Better than 90% of vlms via dual-scale vit and progressive scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1

[7] MiniCPM Team. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 1

[8] Qwen Team. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1