WHEN VISIBILITY DEGRADES TOGETHER: EVALUATING HUMAN POSE ESTIMATION ACROSS SEVERITY AND COMPOSITIONAL LOW-VISIBILITY CONDITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Human pose estimation (HPE) rarely operates under a single isolated visibility degradation, yet existing HPE evaluations typically treat each low-visibility condition as an independent fixed-difficulty bucket. They either isolate degradations into separate benchmarks or lack explicit low-visibility annotations, making severity-aware and compositional robustness analysis difficult. We introduce LV-POSE, a diagnostic evaluation framework for analyzing 2D HPE under severity-aware and compositional low-visibility conditions. LV-Pose reveals robustness failures that remain hidden under existing evaluation protocols. Individually mild degradations can compound when they co-occur, while performance within a single condition can deteriorate sharply as visibility severity increases. LV-Pose groups images into three difficulty levels (easy, medium, hard). Keypoint and severity annotations are provided for three primary conditions, haze/fog, low-light/nighttime, and precipitation, together with image-level tags for additional co-occurring visibility conditions. LV-Pose therefore supports severity-conditioned and composition-conditioned analysis of HPE robustness, instead of a single aggregate robustness claim under low visibility. Across representative HPE estimators and foundation models such as vision foundation models and vision-language models, LV-Pose reveals strong severity-dependent robustness degradation and compounding failures where co-occurring degradations reduce accuracy beyond their isolated effects. Restoration preprocessing also fails to consistently improve downstream pose robustness across severity levels, and can even degrade strong HPE systems despite visually enhanced inputs.

1 INTRODUCTION

Human pose estimation (HPE), performed by specialized estimators or foundation models, is increasingly used in real-world applications, including surveillance, autonomous driving, healthcare, and workplace safety (Zheng et al., 2024; Fang & López, 2018; Chen et al., 2020b; Park et al., 2024). In practice, these systems must often process images captured under challenging visibility conditions such as haze and fog, nighttime and low-light, glare from oncoming headlights, motion blur, and adverse weather conditions including rain or snow (Dai et al., 2020; Cho et al., 2024; Ai et al., 2024; Chen et al., 2026a). Recent failures of deployed vision systems under low-visibility conditions further highlight the importance of robustness beyond clear daytime conditions (National Highway Traffic Safety Administration, 2026). However, visibility degradation in the real world is neither isolated nor binary. Severity varies continuously, and multiple degradations often co-occur within the same image.

Existing HPE evaluation benchmarks do not fully take real-world challenges into consideration. Standard evaluations such as MPII (Andriluka et al., 2014) and COCO (Lin et al., 2014) are built largely from well-lit images and provide little lighting or weather related visibility-specific analysis. Occlusion and truncation also reduce visibility, but their cause is the scene or the framing, whether a joint is blocked by other people, as in OCHuman (Zhang et al., 2019) and CrowdPose (Li et al., 2019b), or cropped out of the frame (Purkrábek & Matas, 2025). In this paper, we focus on low-visibility conditions that degrade the image through weather, lighting, or image capture, rather than conditions defined by occlusion or truncation. Recent evaluations of these conditions cover one or

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

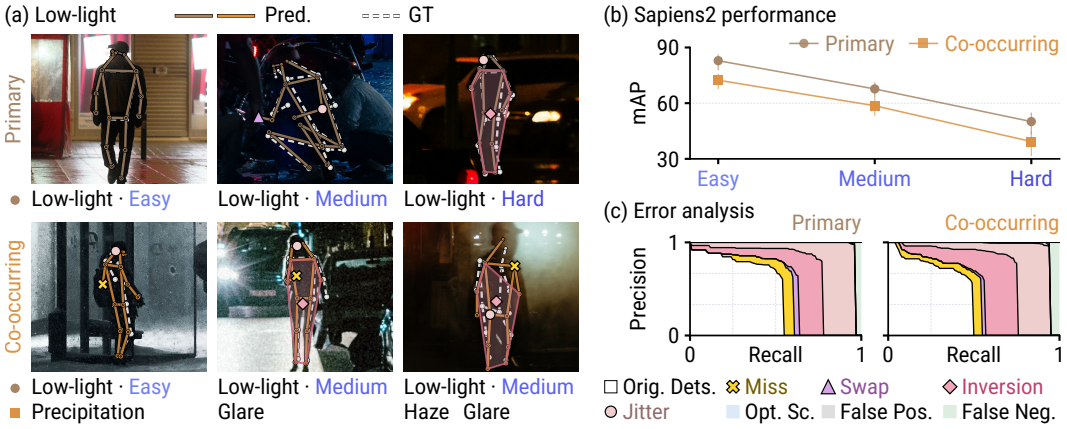


Figure 1: (a) Illustrative low-light examples: primary (top), co-occurring (bottom). (b) Sapiens2 mAP for primary and co-occurring across low-light severity, with 90% intervals. (c) Diagnostic PR curves with OKS@0.9 on medium low-light for primary and co-occurring, with seven corrections (Ronchi & Perona, 2017).

two degradations, such as low-light or motion blur (Lee et al., 2023; Xu et al., 2024a; Cho et al., 2024; Cormier et al., 2024; Su et al., 2025), yet still treat each as an isolated or fixed-difficulty category. As summarized in Figure 2, existing benchmarks may cover individual visibility conditions and their severity, but do not combine severity annotation across multiple conditions with explicit tags for co-occurring degradations. As a result, the following two important robustness failures remain largely unexplored. First, model performance within a single visibility condition can deteriorate sharply as severity increases. Second, individually mild degradations can produce compounding failure effects when they co-occur within the same image.

We introduce LV-POSE, a diagnostic evaluation framework for 2D human pose estimation built from real-world low-visibility images (Figure 1). LV-Pose supports severity-conditioned and composition-conditioned analysis of HPE robustness, instead of a single aggregate robustness claim under low visibility. LV-Pose includes three primary low-visibility conditions, haze/fog, low-light/nighttime, and precipitation, together with additional co-occurring degradation tags including glare/backlight, motion blur, and focus blur. Each annotated visibility condition is assigned one of three difficulty levels (easy, medium, or hard). These degradation tags make co-occurring visibility conditions explicit and enable controlled analysis of compounding and severity robustness failures.

We evaluate representative HPE estimators, including modern Transformer architectures and foundation models, under controlled severity-aware and compositional low-visibility conditions. Grouping visibility conditions into different difficulty levels reveals strong severity-dependent robustness degradation within conditions previously summarized by a single score. Compositional analysis further shows that individually mild degradations can produce compounding failures when they co-occur within the same image. We further evaluate whether image restoration can improve downstream pose robustness by applying multiple image restoration methods across different severity levels. These restoration methods struggle to consistently improve HPE performance, and can even decrease it.

Our contributions are:

- The first systematic compositional evaluation of co-occurring low-visibility degradations for human pose estimation, enabling controlled analysis of compounding robustness failures under real-world conditions.
- A severity-aware diagnostic evaluation framework that groups low-visibility conditions into three difficulty levels, allowing robustness analysis beyond single aggregate performance under low visibility.
- An empirical study across representative HPE models, vision foundation models, and vision-language models, revealing strong severity-dependent robustness degradation, compounding failures across co-occurring degradations, and inconsistent benefits from restoration preprocessing for downstream pose robustness.

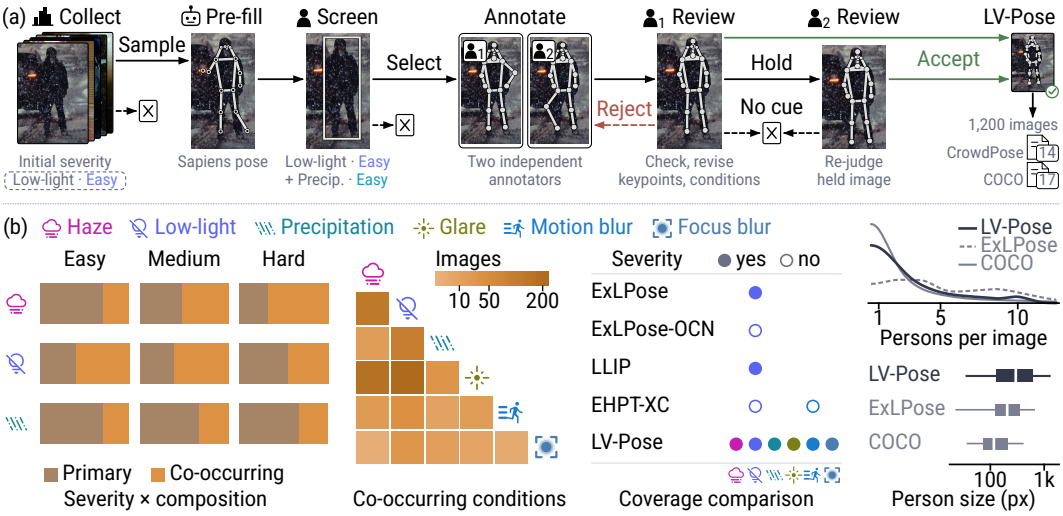


Figure 2: LV-Pose construction and dataset overview. (a) Annotation and review pipeline. (b) Primary-only and co-occurring shares by severity; pairwise co-occurrence counts (darker: more images); condition coverage (filled/hollow: with/without severity labels); and person-count and person-size distributions for LV-Pose, ExLPose (test), and COCO val2017.

2 Evaluation Framework

2.1 Image Source

Our evaluation framework uses only real low-visibility images from publicly available sources. We draw low-light/night candidates from ExDark (Loh & Chan, 2019), ARID (Xu et al., 2021), LMOT (Wang et al., 2024c), and M3Track (Fan et al., 2025), haze/fog candidates from UFDD (Nada et al., 2018), RESIDE-RTTS (Li et al., 2019a), and Fire360 (Tiwari et al., 2025), and precipitation candidates from ACDC (Sakaridis et al., 2021; 2026). We supplement these sources with a curated subset of reLAION-5B (Schuhmann et al., 2022) and publicly released images from Pexels to improve coverage of real-world low-visibility conditions.

This initial candidate image pool is substantially larger than the final accepted ground-truth dataset. We compute condition-specific scores across the pool using whole-image FADE (Choi et al., 2015) for haze/fog, LIME (Guo et al., 2017) for low-light/night, and JSTASR (Chen et al., 2020a) for precipitation. Screeners and reviewers jointly examine candidate images grouped across these score distributions, visually inspect sample images, and set initial condition-specific thresholds separating easy, medium, and hard difficulty. These scores provide initial severity levels. During the screening and reviewing stages, these levels are refined based on actual keypoint visibility and human pose recoverability.

Before full annotation begins, the team uses a small subset of images to establish operational standards for each stage and align judgment criteria among participants within each role. Annotators also receive annotated images and visualizations from COCO (Lin et al., 2014), CrowdPose (Li et al., 2019b), and ExLPose (Lee et al., 2023) as reference examples to guide annotation.

2.2 Annotation Process

Images enter the annotation portal with initial condition and severity labels, candidate per-instance bounding boxes from DINOv3 (Siméoni et al., 2026) or upstream ground truth (GT), and candidate keypoints from Sapiens (Khirodkar et al., 2024). The schema uses 19 keypoints, combining the COCO 17-keypoint convention (Lin et al., 2014) with the `top_head` and `neck` landmarks from CrowdPose (Li et al., 2019b) to maintain compatibility with both evaluation protocols. Each annotated image receives one primary condition chosen from `haze/fog`, `low-light/night`, or `precipitation`, and zero or more secondary degradation

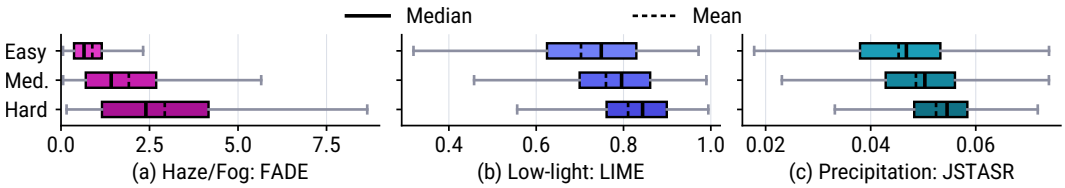


Figure 3: Local visual difficulty around final COCO-17 keypoints: (a) FADE for haze/fog, (b) LIME for low-light/night (the LIME score is inverted, so higher values indicate darker regions), and (c) JSTASR for precipitation. Solid lines indicate the median and dashed lines indicate the mean.

tags from `{haze/fog, low-light/night, precipitation, glare/backlight, motion blur, focus blur}`. Every present condition, whether primary or secondary, receives its own separate difficulty level of `easy`, `medium`, or `hard`.

Each image then proceeds through screening and dual annotation before final review. During screening, screeners filter out unusable images, correct candidate bounding boxes, and refine all primary and secondary condition labels as well as the separate severity level of each condition, based on actual keypoint visibility and human pose recoverability rather than global image appearance. In the dual annotation stage, two independent annotators work on each screened image exclusively to localize, refine, or replace keypoint coordinates and assign each keypoint its `0/1/2` visibility state following COCO (Lin et al., 2014).

In the review stage, reviewers inspect keypoint locations, keypoint visibility states, all primary and secondary condition labels, and every associated severity level to decide whether each image is accepted, revised, or rejected. Ambiguous cases concerning keypoint placement, keypoint visibility states, condition labels, or severity levels are placed on hold for discussion and re-judging by a second reviewer. Final condition and severity assignments reflect keypoint visibility and pose recoverability through this review process. Figure 3 shows that final severity levels correspond to local visibility degradation around keypoints, and Figure 1 provides examples of low-light images with varying severity levels and co-occurring conditions. The annotation pipeline preserves intermediate annotation states throughout screening, dual annotation, and review, and exports final accepted annotations in both COCO and CrowdPose keypoint formats for evaluation. We will release only bounding box, keypoint, visibility-condition, and difficulty annotations together with source identifiers, while original images remain subject to their respective licenses.

2.3 Annotation Quality

Inter-annotator agreement Two independent annotators label every image without seeing each other's pass, a reviewer checks both annotations and resolves disagreements instead of averaging keypoint coordinates, and hard images are held for a second review typically. Before review, 78.6% of COCO-17 keypoints labeled by both annotators agree at $k_s \geq 0.85$, the human error margin of Ronchi & Perona (2017) (21.4% disagreement, against around 30% for COCO annotators).

Sapiens initialization The initial keypoints of Section 2.2 come from Sapiens, and the annotation and reviewing stages limit the influence of the initialization. On 180 images re-annotated without Sapiens initialization, Sapiens2 ranks first under both sets of ground-truth annotations, and the order of the eight models changes only by one swap of two adjacent models (Kendall τ 0.929). Replacing keypoints left unchanged from the Sapiens initialization with their coordinates from re-annotation without initialization changes Sapiens2's AP relative to the other models by -0.09 $[-0.33, +0.19]$, indicating no detectable advantage for Sapiens2 on this audited subset. Details of the initialization audit are in Section G.

Severity and local visibility We measure local visual difficulty around every final COCO-17 keypoint using a patch whose radius is 5% of the person-box diagonal. We evaluate FADE (Choi et al., 2015) for haze/fog, LIME (Guo et al., 2017) for low-light/night, and JSTASR (Chen et al., 2020a) small-scale SIM for precipitation. Patch scores are pooled directly across keypoints within

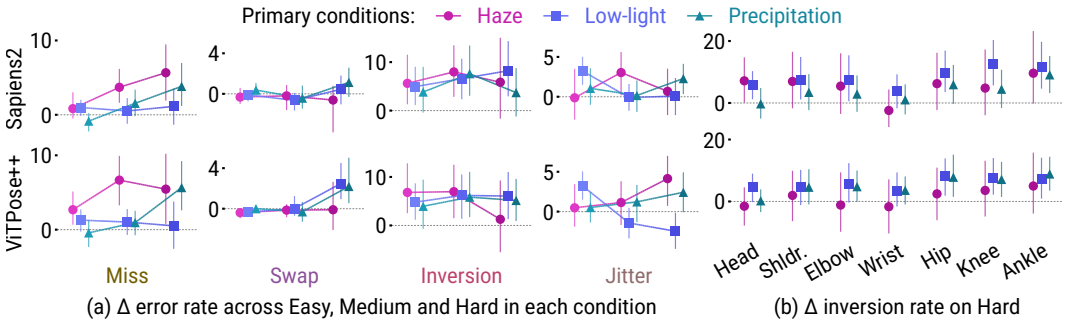


Figure 4: (a) shows co-occurring minus only rates in percentage points across error types and severities. (b) shows hard inversion rate differences across seven parts. Inversion is same-person left-right confusion. Whiskers show pointwise 90% confidence intervals.

each severity group. Score ranges overlap because each condition's severity is assigned at the image level, yet means and medians increase from easy to medium to hard in Figure 3.

3 Experiments

3.1 Experiment Setup

Our evaluation follows the standard COCO evaluation protocol (Lin et al., 2014), reporting mAP (noted as AP in the rest of the paper) over OKS thresholds from 0.50 to 0.95 with a stride of 0.05, consistent with existing HPE evaluations (Lin et al., 2014; Zhang et al., 2019; Li et al., 2019b; Lee et al., 2023; Cho et al., 2024). We compute OKS using `xtcocotools` with the standard 17-keypoint COCO sigmas provided by `pycocotools` and the 14-keypoint CrowdPose sigmas provided by `crowdposetools`, evaluated on the corresponding 17- or 14-keypoint export of our 19-keypoint annotations. For the well-lit (WL) baseline, we report performance on COCO-val2017 (Lin et al., 2014).

For top-down methods, we use ground-truth person bounding boxes during evaluation. Vision foundation models (VFMs) and Vision-language models (VLMs) are evaluated in the same top-down setting. For each ground-truth person box, we crop the surrounding image context, encode the target box within a normalized `xy` coordinate frame, and prompt the model to predict all 17 COCO keypoints in JSON format. For closed-source models, we use API calls. For open-source models, we utilize their official released checkpoints on HuggingFace. All remaining HPE models use their publicly released COCO checkpoints (Lin et al., 2014). Specific backbone variants and parameter counts are listed in Table 2.

3.2 Compounding Effects

Table 1 compares images containing haze/fog alone with images containing haze/fog and at least one additional visibility degradation. The co-occurring subset contains 218 images (41 easy, 84 medium, 93 hard), while the haze-only subset contains 221 images (100 easy, 77 medium, 44 hard). Because these severity distributions differ, AP reflects differences in both severity and composition.

Across the evaluated models, AP is lower on the co-occurring subset, both overall and within each haze/fog severity group. Sapiens2 achieves 67.5 AP on haze-only images versus 46.2 on co-occurring images, with gaps of 9.8, 15.8, and 14.3 AP at easy, medium, and hard severity, respectively. The same direction holds across specialized HPE models, VFMs, and VLMs.

These results support evaluating co-occurring degradations alongside severity: severity grouping alone does not capture the performance differences between haze-only and co-occurring images. The comparison is observational, however, and severity grouping does not eliminate other differences between the subsets. Larger per-condition samples would enable finer-grained analysis of individual secondary degradations.

Table 1: AP under Haze/Fog for co-occurring and single-condition images, pooled over severity and at easy, medium, and hard; see Section 3.2. **Bold** marks the column leader and every model that a paired bootstrap cannot separate from it.

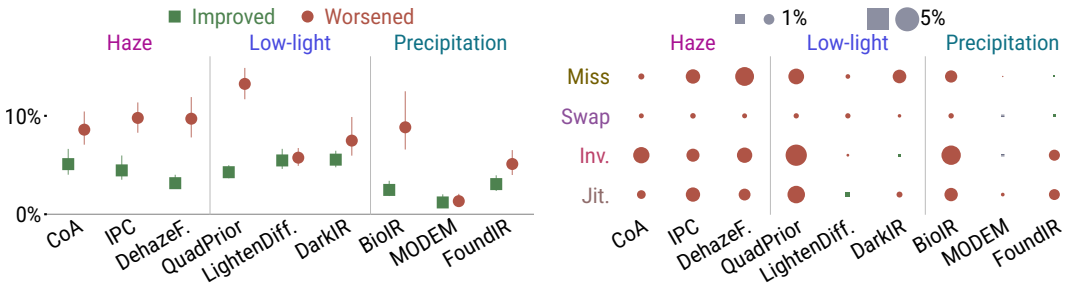
Model	Co-occurring$_{218}$				Haze/Fog only$_{221}$			
	All$_{218}$	Easy$_{41}$	Med.$_{84}$	Hard$_{93}$	All$_{221}$	Easy$_{100}$	Med.$_{77}$	Hard$_{44}$
ViTPose++-H	42.9	**66.4**	42.6	**30.7**	**66.1**	**79.6**	**60.3**	40.4
ProbPose	28.2	53.3	29.3	14.8	54.1	69.2	46.7	27.2
CIGPose-L	33.3	59.0	35.9	17.8	58.9	72.9	52.3	33.9
GroupPose	27.8	51.0	27.2	16.8	50.0	61.4	45.0	30.4
YOLO26x-pose	23.0	45.4	23.6	12.4	45.0	57.5	39.7	25.4
Sapiens2	**46.2**	**68.1**	**48.0**	**32.4**	**67.5**	**77.9**	**63.8**	**46.7**
GKDT-H	36.5	58.8	36.3	26.4	52.4	61.6	50.9	33.2
LocLLM	41.9	60.3	42.8	**31.1**	64.7	75.6	60.4	**43.0**
Rex-Omni	12.9	20.6	16.7	6.2	26.1	31.3	26.8	13.0
SenseNova-Vision	17.6	31.2	19.5	10.1	34.8	43.0	32.7	19.9
GPT-5.6-Sol	4.6	9.7	5.5	1.3	11.3	14.5	11.2	3.9
Claude-Opus-5	7.2	15.5	7.7	3.2	15.5	20.0	12.4	9.9
Gemini-3.8-Flash	7.0	12.4	8.6	2.8	17.7	21.4	17.6	9.4
Qwen-3.8	9.0	17.0	9.5	5.7	18.4	23.7	17.6	7.7

Figure 4 further examines localization errors across the three primary conditions for Sapiens2 and ViTPose++. Both models show higher miss-rate estimates under co-occurring haze. Sapiens2 also shows higher inversion-rate estimates across conditions, while the pattern for ViTPose++ varies with severity. On hard images, inversion-rate differences vary across body parts and conditions. These results show that the performance gaps are accompanied by changes in specific localization errors.

3.3 Overall performance

Table 2: We evaluate representative HPE methods (Xu et al., 2024b; Purkrábek & Matas, 2025; Li et al., 2026; Liu et al., 2023; Jocher et al., 2026), VFMs (Khirodkar et al., 2026; Lu et al., 2026; Wang et al., 2024a; Jiang et al., 2026; Han et al., 2026) and VLMs (OpenAI, 2026; Anthropic, 2026; Doshi & Popa, 2026; Qwen Team, 2026) on our evaluation set: well-lit (WL) baseline plus haze/fog, low-light/night, and precipitation conditions, each split into easy/medium/hard severity strata. Numbers are mAP@.50:.95 (↑). The subscript on each condition label is the number of images in that condition, and on each severity label the number of images in that cell. **Bold** marks the best model in each column and every model that a paired bootstrap cannot separate from it. The intervals themselves are in Table 13.

Model	Params	WL	Haze/Fog$_{411}$			Low-light/Night$_{392}$			Precipitation$_{397}$		
			Easy$_{128}$	Med.$_{148}$	Hard$_{135}$	Easy$_{118}$	Med.$_{148}$	Hard$_{126}$	Easy$_{118}$	Med.$_{148}$	Hard$_{131}$
ViTPose++-H	632M	**82.1**	**77.5**	52.7	34.0	75.1	55.9	39.1	**79.7**	**69.6**	**63.2**
ProbPose	40M	76.5	66.5	38.6	19.0	62.7	45.4	28.7	67.4	56.2	50.6
CIGPose-L	28M	78.4	70.2	45.0	23.4	67.4	48.9	29.8	72.3	62.4	54.3
GroupPose	51M	73.6	59.6	37.5	21.3	44.2	35.0	21.0	47.9	46.8	36.8
YOLO26x-pose	57M	70.8	55.7	33.8	16.9	37.4	27.6	13.5	44.8	42.5	34.0
Sapiens2	5B	78.8	**76.5**	**57.3**	**37.7**	**77.7**	**64.0**	**46.5**	**78.6**	**70.6**	**61.7**
GKDT-H	1.4B	80.5	60.5	45.2	28.7	67.8	47.5	30.6	68.1	61.7	55.0
LocLLM	7B	77.4	72.7	52.2	**35.5**	70.2	55.9	37.5	73.4	63.5	57.0
Rex-Omni	4B	44.9	28.5	20.8	8.5	32.6	23.5	9.6	32.1	31.8	26.0
SenseNova-Vision	14B	58.3	41.3	25.7	13.2	38.7	29.6	12.5	40.4	37.0	32.6
GPT-5.6-Sol	/	18.7	13.7	7.8	2.1	10.9	7.3	3.0	13.6	13.9	14.0
Claude-Opus-5	/	19.0	19.3	9.5	5.4	12.6	8.5	5.1	14.5	14.4	15.5
Gemini-3.8-Flash	/	32.5	19.8	13.5	5.3	13.3	10.2	3.8	15.5	15.1	15.6
Qwen-3.8	27B	29.7	22.7	12.6	6.2	17.3	14.5	5.9	20.2	15.4	16.2



(a) Improved and worsened keypoints

(b) Improved and worsened by error type

Figure 5: Restoration under hard severity for Sapiens2. (a) Transition percentages of improved and worsened joints, with marginal 90% confidence intervals. (b) Improved and worsened percentages by error type. A gray equals sign marks exact equality of the two percentages.

Table 2 reports mAP@.50:.95 on the well-lit (WL) baseline together with haze/fog, low-light/night, and precipitation severity levels. This subsection focuses on severity-aware evaluation, while compounding effects from co-occurring degradations are analyzed separately in Section 3.2.

Among specialized HPE estimators and VFMs, ViTPose++-H leads on WL at 82.1, followed by GKDT-H at 80.5, Sapiens2 at 78.8, and CIGPose-L at 78.4. The general-purpose VLMs perform substantially worse on WL, with Gemini-3.8-Flash at 32.5 and GPT-5.6-Sol at 18.7. This large WL performance gap makes the low-visibility robustness results harder to interpret for VLMs. Their low-visibility scores reflect both baseline pose-estimation capability and performance under degraded visibility. As a result, the VLM rows in LV-Pose function less as direct robustness comparisons and more as a diagnostic check of whether general-purpose vision-language systems can reliably support human pose estimation under challenging low-visibility conditions.

Across specialized HPE estimators and VFMs, grouping haze/fog images into different difficulty levels reveals two robustness patterns that a single overall haze score would hide. First, four representative models remain relatively close at Easy severity (Sapiens2 76.5, LocLLM 72.7, CIGPose-L 70.2, ProbPose 66.5) but separate more at Hard severity (Sapiens2 37.7, LocLLM 35.5, CIGPose-L 23.4, ProbPose 19.0). Methods with similar performance under mild degradation therefore exhibit substantially different performance under severe visibility degradation. Second, the leading model changes across severity levels. ViTPose++-H has the highest reported AP on WL and Easy haze/fog, while Sapiens2 has the highest reported AP on Medium and Hard haze/fog. Evaluating only on WL would therefore miss the change in the leading model under severe haze/fog. Likewise, a single aggregate haze score cannot show how model rankings and performance gaps vary across severity levels. Grouping images by severity therefore changes the evaluation question from "which method performs better under haze?" to "how do model performance and rankings change as visibility becomes harder?"

The severity-dependent decline also extends to low-light/night and precipitation for specialized HPE estimators and VFMs. For example, Sapiens2 decreases from 77.7 at Easy to 46.5 at Hard under low-light/night, and from 78.6 to 61.7 under precipitation. The leading model also depends on the condition: Sapiens2 has the highest reported AP across all low-light/night severity levels, while ViTPose++-H has the highest reported AP at Easy and Hard precipitation and Sapiens2 at Medium. These results motivate reporting performance separately by both visibility condition and severity.

3.4 Low-light/Nighttime

We examine whether restoration preprocessing improves downstream pose estimation under low visibility. Table 3 reports low-light/night results for QuadPrior (Wang et al., 2024b), LightenDiff (Jiang et al., 2024), and DarkIR (Feijoo et al., 2025) across 14 pose estimation models and three severity levels. Subscripts indicate AP changes relative to inference without preprocessing. Figure 5 complements this comparison by examining improved and worsened keypoints for Sapiens2 at Hard severity across haze/fog, low-light/night, and precipitation.

The effect of low-light enhancement depends on both the enhancer and the downstream pose estimator. QuadPrior lowers AP in 39 of its 42 model–severity combinations, with its largest decrease on

Table 3: Pose estimation under low-light with low-light-enhancement preprocessing, stratified by severity. WL is the well-lit baseline. Restoration methods are QuadPrior (Wang et al., 2024b), LightenDiff (Jiang et al., 2024), and DarkIR (Feijoo et al., 2025). Numbers are mAP@.50:.95 (↑); subscripts show the change from the no-preprocessing Low-light/Night score in Table 2 (green: improvement, red: degradation).

Model	WL	QuadPrior			LightenDiff			DarkIR		
		Easy_{118}	Med._{148}	Hard_{126}	Easy_{118}	Med._{148}	Hard_{126}	Easy_{118}	Med._{148}	Hard_{126}
ViTPose++-H	82.1	$63.9_{-11.2}$	$47.7_{-8.2}$	$33.1_{-6.0}$	$74.8_{-0.3}$	$58.1_{+2.2}$	$40.0_{+0.9}$	$76.0_{+0.9}$	$57.2_{+1.3}$	$39.0_{\pm 0.0}$
ProbPose	76.5	$52.8_{-9.9}$	$39.6_{-5.8}$	$27.0_{-1.7}$	$62.1_{-0.6}$	$49.1_{+3.6}$	$33.0_{+4.3}$	$61.9_{-0.7}$	$48.2_{+2.8}$	$32.3_{+3.6}$
CIGPose-L	78.4	$59.6_{-7.8}$	$43.0_{-5.9}$	$29.3_{-0.5}$	$66.7_{-0.7}$	$50.6_{+1.7}$	$34.9_{+5.1}$	$67.9_{+0.5}$	$50.1_{+1.1}$	$33.5_{+3.8}$
GroupPose	73.6	$38.8_{-5.4}$	$29.7_{-5.3}$	$17.2_{-3.8}$	$43.9_{-0.4}$	$34.1_{-0.9}$	$20.9_{\pm 0.0}$	$43.2_{-1.0}$	$34.0_{-1.0}$	$20.4_{-0.6}$
YOLO26x-pose	70.8	$35.7_{-1.7}$	$25.2_{-2.3}$	$16.1_{+2.6}$	$40.3_{+2.9}$	$28.4_{+0.8}$	$17.4_{+4.0}$	$41.1_{+3.7}$	$28.3_{+0.7}$	$16.5_{+3.1}$
Sapiens2	78.8	$64.6_{-13.1}$	$50.3_{-13.8}$	$34.9_{-11.6}$	$77.3_{-0.4}$	$61.6_{-2.5}$	$45.2_{-1.4}$	$77.4_{-0.3}$	$61.5_{-2.6}$	$44.0_{-2.6}$
GKDT-H	80.5	$54.7_{-13.1}$	$38.0_{-9.6}$	$24.2_{-6.4}$	$67.7_{-0.1}$	$47.5_{\pm 0.0}$	$32.6_{+2.0}$	$69.0_{+1.2}$	$48.7_{+1.1}$	$31.7_{+1.2}$
LocLLM	77.4	$60.9_{-9.4}$	$45.3_{-10.6}$	$30.5_{-7.0}$	$68.7_{-1.5}$	$54.5_{-1.4}$	$37.0_{-0.6}$	$69.5_{-0.7}$	$55.6_{-0.3}$	$36.3_{-1.3}$
Rex-Omni	44.9	$21.7_{-10.9}$	$15.2_{-8.3}$	$7.1_{-2.5}$	$30.8_{-1.7}$	$23.9_{+0.4}$	$9.6_{\pm 0.0}$	$29.8_{-2.8}$	$23.1_{-0.4}$	$9.6_{\pm 0.0}$
SenseNova-Vision	58.3	$31.9_{-6.7}$	$22.9_{-6.7}$	$11.0_{-1.5}$	$41.1_{+2.5}$	$31.2_{+1.6}$	$16.0_{+3.5}$	$40.7_{+2.1}$	$29.2_{-0.5}$	$13.9_{+1.4}$
GPT-5.6-Sol	18.7	$9.6_{-1.3}$	$6.4_{-0.9}$	$2.8_{-0.1}$	$13.4_{+2.6}$	$10.1_{+2.8}$	$5.0_{+2.0}$	$12.6_{+1.8}$	$10.2_{+3.0}$	$4.4_{+1.4}$
Claude-Opus-5	19.0	$11.2_{-1.4}$	$7.6_{-0.9}$	$4.4_{-0.7}$	$14.3_{+1.7}$	$11.4_{+2.9}$	$6.2_{+1.1}$	$13.0_{+0.4}$	$10.9_{+2.4}$	$5.2_{\pm 0.0}$
Gemini-3.8-Flash	32.5	$13.6_{+0.3}$	$8.6_{-1.6}$	$4.4_{+0.6}$	$14.6_{+1.2}$	$13.0_{+2.8}$	$6.1_{+2.3}$	$15.7_{+2.4}$	$13.1_{+2.9}$	$5.6_{+1.8}$
Qwen-3.8	29.7	$15.0_{-2.3}$	$10.1_{-4.4}$	$5.2_{-0.6}$	$18.7_{+1.4}$	$13.3_{-1.2}$	$8.8_{+2.9}$	$18.3_{+1.0}$	$15.7_{+1.3}$	$6.0_{+0.1}$

Sapiens2 at Medium severity (-13.8 AP). LightenDiff and DarkIR provide more mixed results, improving AP in 25 and 26 of their 42 combinations, respectively. The benefit also varies with severity. LightenDiff lowers AP for 8 of the 14 models at Easy severity but improves it for 10 at Hard severity. At Hard severity, its largest improvements are on CIGPose-L ($+5.1$ AP), ProbPose ($+4.3$), and YOLO26x-pose ($+4.0$). These results show that an enhancer's downstream benefit cannot be inferred from its performance at a single severity level or with a single pose estimator.

Enhancement also does not consistently benefit models with strong performance on unprocessed images. Sapiens2 and LocLLM lose AP under all three enhancers at every severity level. Without preprocessing, Sapiens2 achieves 77.7, 64.0, and 46.5 AP at Easy, Medium, and Hard severity, respectively. The highest AP achieved with enhancement at each severity remains below these values: 77.4 with DarkIR at Easy, and 61.6 and 45.2 with LightenDiff at Medium and Hard. In contrast, 22 of the 24 VLM results with LightenDiff or DarkIR improve, with gains of up to 3.0 AP. These gains should be interpreted alongside the VLMs' low baseline pose-estimation accuracy: improvement from preprocessing does not by itself establish reliable pose estimation.

Figure 5 further shows that restoration can improve some keypoints while worsening others within the same evaluation. For Sapiens2 at Hard severity, the estimated percentage of worsened keypoints exceeds that of improved keypoints for all nine restoration methods across the three primary conditions. The error-type breakdown shows that these changes are not uniform: worsening is particularly apparent for miss and inversion errors under several restorers, whereas swap-error changes are generally smaller. These keypoint-level transitions complement the AP results by showing that successful corrections can coexist with newly introduced localization errors. The transition percentages are not a decomposition of AP, and the plotted differences should not all be interpreted as statistically significant.

Overall, restoration preprocessing does not provide a consistent improvement in downstream pose estimation. Its effects depend on the restoration method, pose estimator, and visibility severity, motivating evaluation of downstream pose accuracy rather than assuming that image restoration will improve pose robustness.

4 Related Work

Low-visibility human pose estimation Existing 2D HPE evaluations largely assume clear and well-lit conditions. MPII (Andriluka et al., 2014) and COCO (Lin et al., 2014) establish standard protocols

for articulated pose and keypoint detection, while OCHuman (Zhang et al., 2019) and CrowdPose (Li et al., 2019b) focus on occlusion and dense crowds, ProbPose (Purkrábek & Matas, 2025) models keypoints that fall outside the image. Occlusion hides a joint and truncation crops it out of the frame, while LV-Pose focuses on image degradation caused by weather (haze, precipitation), lighting (nighttime, glare), or image capture (motion and focus blur). LV-Pose therefore complements rather than replaces these efforts. A growing line of work extends HPE evaluation to degraded visibility conditions, particularly under low-light settings. Lee et al. (2023) introduce an RGB low-light evaluation split, while LLIP (Su et al., 2025) extends low-light pose evaluation using motion-capture-derived annotations. Thermal and infrared sensing provide additional directions for low-visibility pose estimation. Cormier et al. (2024) incorporate thermal cues for pose estimation in dim scenes, while InfPose (Xu et al., 2024a) targets real-time infrared deployment. Cho et al. (2024) further extend low-visibility evaluation with an event-based setting that combines motion blur and low-light conditions. However, two important limitations remain. First, existing evaluations do not group real low-visibility images by controlled severity levels, making severity-dependent robustness degradation difficult to analyze. Second, most evaluations isolate one or two degradations and do not explicitly analyze co-occurring visibility conditions, preventing systematic study of compounding robustness failures. LV-Pose addresses both limitations through severity-aware grouping and explicit compositional analysis. A comparison with existing benchmarks is provided in the appendix (Table 12).

Image restoration and enhancement Recent image restoration methods have improved visual quality under degraded visibility conditions. For low-light enhancement, Retinexformer (Cai et al., 2023), Diff-Retinex (Yi et al., 2023), and DarkIR (Feijoo et al., 2025) represent recent approaches, with DarkIR additionally handling deblurring. For dehazing, DehazeFormer (Song et al., 2023), RIDCP (Wu et al., 2023), and CoA (Ma et al., 2025) explore Transformer attention, codebook priors, and synthetic-to-real adaptation. For rain and snow removal, BioIR (Cui et al., 2025), MODEM (Wang et al., 2025), FoundIR-v2 (Chen et al., 2026b), and RestoreVAR (Rajagopalan et al., 2026) explore bio-inspired designs, Morton-order degradation estimation, pre-training data mixtures, and visual autoregressive generation. General restoration frameworks such as PromptIR (Potlapalli et al., 2023) and InstructIR (Conde et al., 2024) further target multiple degradations within a unified model. However, improved visual quality does not necessarily translate into improved downstream pose robustness. Recent work reports that image enhancement may introduce artifacts or fail to recover sufficient detail, limiting downstream pose estimation (Ai et al., 2024; Chen et al., 2026a). Whether modern restoration methods consistently improve downstream robustness across different visibility conditions and severity levels remains an open empirical question. We therefore treat restoration as a diagnostic intervention rather than a presumed solution, systematically evaluating whether restoration preprocessing improves downstream pose robustness under severity-aware and compositional low-visibility conditions.

5 Conclusion

We introduced LV-Pose, a diagnostic evaluation framework for analyzing human pose estimation under severity-aware and compositional low-visibility conditions. Unlike existing evaluations that treat low-visibility conditions as isolated fixed-difficulty categories, LV-Pose groups visibility conditions into controlled difficulty levels and explicitly annotates co-occurring degradations on the same image. This enables systematic analysis of how visibility severity and compounded degradations jointly affect downstream pose robustness. Through evaluation across HPE-specialized models, vision-language models, and multimodal foundation models, LV-Pose reveals several robustness failures hidden by conventional aggregate evaluation protocols. First, model rankings can change substantially across visibility severity levels, showing that overall low-visibility scores may hide severe degradation-specific failures. Second, co-occurring degradations produce strong compounding robustness failures beyond isolated visibility conditions alone. Third, restoration preprocessing does not consistently improve downstream pose robustness and can substantially degrade performance for strong HPE models under severe visibility conditions. Overall, our results suggest that robustness under real-world low-visibility conditions cannot be reliably characterized using single-condition or single-score evaluation protocols alone. We hope LV-Pose can provide a useful diagnostic benchmark for studying robustness, failure analysis, and evaluation methodology for future human-centric vision systems operating under challenging real-world environments.

AI USE STATEMENT

In this work, we used generative AI tools to design or provide feedback on research methodology or experiments, to implement methods, to support qualitative and thematic data analysis, and to interpret results. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, to propose or refine hypotheses, to assist with translation, or to clean and reformat the dataset. The remaining tasks with required disclosure do not apply here. LV-POSE is a benchmark built from real photographs only (Section 2.1), so we generated no synthetic data sets, and the paper states no mathematical claims and contains no proofs. Additionally, we used generative AI tools to create or modify scientific figures or images, to suggest experimental parameters, to create or edit software code, to summarize or analyze existing literature, to brainstorm, to source and search for information, to edit the paper to improve readability, to identify relevant literature, to format references, and to propose a title and keywords.

We have reviewed all AI-assisted work. We read and ran every piece of AI-assisted code before using its output. We verified every AI-suggested reference against the original publication record before citing it, and we checked AI-written summaries of prior work against the papers themselves. We also checked AI-drafted readings of our results against the per-image evaluation outputs, and we narrowed or dropped the claims that did not survive that check.

We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

LV-POSE is built on curated subsets of publicly available image sources (Section 2.1), and we evaluate existing models on them. The public release of LV-POSE does not redistribute the original images. We release the annotations we produced, which are bounding boxes, keypoints, visibility-condition labels, and difficulty levels, together with per-image source identifiers. Users obtain the images from their original sources under those sources' licenses, and all external datasets, models, and code assets are cited and used under their respective licenses and terms of use. The annotation schema records no identity, demographic, or biometric attributes.

We introduce and release no models, generative systems, or datasets that pose a high risk of misuse, so no additional safeguards are required for this release. People carry out the screening, dual annotation, and review described in Section 2.2. We used no crowdsourcing, conducted no research with human subjects, and collected no participant data.

This work improves understanding of model robustness under low-visibility conditions, which are common in real-world environments such as nighttime, fog, or adverse weather. Evaluating performance across severity levels and compositional degradations can help identify failure modes in human pose estimation systems and related vision models. This may contribute to safer deployment where reliability under challenging conditions is critical. Better evaluation of low-visibility perception could indirectly support the deployment of vision systems in sensitive contexts such as surveillance, which may raise privacy concerns. Benchmark scores can also be misread as evidence of general reliability. LV-Pose is a diagnostic evaluation set, not a deployment certification. We encourage careful interpretation of our results, particularly in safety-critical applications such as autonomous driving and assistive technologies.

We read and comply with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Sections 2.1 and 2.2 describe image selection and the annotation pipeline. Section 2.3 and Section F report annotation quality, with inter-annotator agreement; Section G audits the Sapiens initialization. All models come from publicly released checkpoints or public APIs. We report 90% BCa intervals from 10,000 whole-image bootstrap resamples, on which all models in a cell are scored. Sections B, C and J give full metric breakdowns, Section H the comparability of the single-condition and co-occurring groups, and Section D pose from predicted person boxes. The supplementary material has 18 LV-POSE images, included only for the review, two per condition-by-severity cell (one single-condition and one co-occurring; Figure 6), with complete COCO-17 and CrowdPose-14 ground truth,

each image's condition-by-severity cell, and a README on the format. The public release contains annotations only: bounding boxes, keypoints, visibility, condition labels and severities, and per-image source identifiers. Users download the images from their upstream sources under their licenses, with our instructions and SHA-256 hashes to check that each image matches the one we annotated.

REFERENCES

Yihao Ai, Yifei Qi, Bo Wang, Yu Cheng, Xinchao Wang, and Robby T. Tan. Domain-adaptive 2d human pose estimation via dual teachers in extremely low-light conditions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, pp. 221–239. Springer, 2024. doi: 10.1007/978-3-031-72970-6_13.

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3686–3693, June 2014.

Anthropic. Introducing Claude Opus 5. Anthropic News, July 2026. URL `https://www.anthropic.com/news/claude-opus-5`.

Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12470–12479, 2023. doi: 10.1109/ICCV51070.2023.01149.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=r35clVtGzw`.

Haopeng Chen, Yihao Ai, Kabeen Kim, Robby T. Tan, Yixin Chen, and Bo Wang. Udapose: Unsupervised domain adaptation for low-light human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13781–13792, June 2026a.

Wei-Ting Chen, Hao-Yu Fang, Jian-Jiun Ding, Cheng-Che Tsai, and Sy-Yen Kuo. JSTASR: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal. In *European Conference on Computer Vision (ECCV)*, pp. 754–770, 2020a. doi: 10.1007/978-3-030-58589-1_45.

Weiming Chen, Zijie Jiang, Hailin Guo, and Xiaoyang Ni. Fall detection based on key points of human-skeleton using openpose. *Symmetry*, 12(5):744, 2020b.

Xiang Chen, Jinshan Pan, Jiangxin Dong, Jian Yang, and Jinhui Tang. FoundIR-v2: Optimizing pre-training data mixtures for image restoration foundation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

Hoonhee Cho, Taewoo Kim, Yuhwan Jeong, and Kuk-Jin Yoon. A benchmark dataset for event-guided human pose estimation and tracking in extreme conditions. In *Advances in Neural Information Processing Systems*, volume 37, pp. 134826–134840. Curran Associates, Inc., 2024. doi: 10.52202/079017-4284.

Lark Kwon Choi, Jaehee You, and Alan Conrad Bovik. Referenceless prediction of perceptual fog density and perceptual image defogging. *IEEE Transactions on Image Processing*, 24(11): 3888–3901, 2015. doi: 10.1109/TIP.2015.2456502.

Marcos V. Conde, Gregor Geigle, and Radu Timofte. InstructIR: High-quality image restoration following human instructions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, pp. 1–21. Springer, 2024. doi: 10.1007/978-3-031-72764-1_1.

Mickael Cormier, Caleb Ng Zhi Yi, Andreas Specker, Benjamin Blaß, Michael Heizmann, and Jürgen Beyerer. Leveraging thermal imaging for robust human pose estimation in low-light vision. In *ACCV Workshops*, Lecture Notes in Computer Science, pp. 69–86. Springer, 2024. doi: 10.1007/978-981-96-2641-0_5.

Yuning Cui, Wenqi Ren, and Alois Knoll. Bio-inspired image restoration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. doi: 10.52202/085713-2689.

Dengxin Dai, Christos Sakaridis, Simon Hecker, and Luc Van Gool. Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. *Int. J. Comput. Vis.*, 128(5): 1182–1204, 2020.

Tulsee Doshi and Raluca Ada Popa. Introducing Gemini 3.8 Flash and 3.8 Flash Cyber. Google Blog, September 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/3-8-flash-and-3-8-flash-cyber/.

Huijie Fan, Yu Qiao, Yihao Zhen, Tinghui Zhao, Baojie Fan, and Qiang Wang. All-day multi-camera multi-target tracking. In *CVPR*, pp. 16892–16901. Computer Vision Foundation / IEEE, 2025.

Zhijie Fang and Antonio M. López. Is the pedestrian going to cross? answering by 2d pose estimation. In *Intelligent Vehicles Symposium*, pp. 1271–1276. IEEE, 2018.

Daniel Feijoo, Juan C. Benito, Álvaro García, and Marcos V. Conde. DarkIR: Robust low-light image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10879–10889, 2025.

Jiayi Fu, Siyu Liu, Zikun Liu, Chun-Le Guo, Hyunhee Park, Ruiqi Wu, Guoqing Wang, and Chongyi Li. Iterative predictor-critic code decoding for real-world image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12700–12709, 2025. doi: 10.1109/CVPR52734.2025.01185.

Xiaojie Guo, Yu Li, and Haibin Ling. LIME: low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993, 2017. doi: 10.1109/TIP.2016.2639450.

Xiaoyang Han, Jianhua Li, Kewang Deng, Zukai Chen, Xuanke Shi, Sihan Wang, Boxuan Li, Linyan Wang, Siyi Xie, Xin You, Jinsheng Quan, Zhongang Cai, Haiwen Diao, Ziwei Liu, Lei Yang, Dahua Lin, and Quan Wang. Vision as Unified Multimodal Generation. *arXiv preprint arXiv:2607.06560*, 2026. doi: 10.48550/arXiv.2607.06560.

Hai Jiang, Ao Luo, Xiaohong Liu, Songchen Han, and Shuaicheng Liu. LightenDiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 161–179, 2024. doi: 10.1007/978-3-031-73195-2_10.

Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25472–25483, June 2026.

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics YOLO26: Unified real-time end-to-end vision models, 2026. URL https://arxiv.org/abs/2606.03748.

Rawal Khirodkar, Timur M. Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *ECCV (4)*, volume 15062 of *Lecture Notes in Computer Science*, pp. 206–228. Springer, 2024.

Rawal Khirodkar, He Wen, Julieta Martinez, Yuan Dong, Zhaoen Su, and Shunsuke Saito. Sapiens2. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=IVAlYCqdvW.

Sohyun Lee, Jaesung Rim, Boseung Jeong, Geonu Kim, Byungju Woo, Haechan Lee, Sunghyun Cho, and Suha Kwak. Human pose estimation in extremely low-light conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 704–714, June 2023.

Bohao Li, Zhicheng Cao, Huixian Li, and Yangming Guo. CIGPose: Causal intervention graph neural network for whole-body pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23484–23494, June 2026.

Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28 (1):492–505, 2019a. doi: 10.1109/TIP.2018.2867951.

Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10863–10872, June 2019b.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV (5)*, Lecture Notes in Computer Science, pp. 740–755. Springer, 2014.

Huan Liu, Qiang Chen, Zichang Tan, Jiang-Jiang Liu, Jian Wang, Xiangbo Su, Xiaolong Li, Kun Yao, Junyu Han, Errui Ding, Yao Zhao, and Jingdong Wang. Group pose: A simple baseline for end-to-end multi-person pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15029–15038, October 2023.

Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019. doi: 10.1016/j.cviu.2018.10.010.

Changsheng Lu, Yuxin Chen, Haokun Gui, Rong Wang, Jie Yang, Harry Yang, Anton van den Hengel, and Jiaya Jia. GKDT: general keypoint detection transformer. In *ECCV (3)*, volume 17003 of *Lecture Notes in Computer Science*, pp. 185–204. Springer, 2026. doi: 10.1007/978-3-032-36846-1_11.

Long Ma, Yuxin Feng, Yan Zhang, Jinyuan Liu, Weimin Wang, Guang-Yong Chen, Chengpei Xu, and Zhuo Su. CoA: Towards real image dehazing via compression-and-adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11197–11206, 2025. doi: 10.1109/CVPR52734.2025.01046.

Hajime Nada, Vishwanath A. Sindagi, He Zhang, and Vishal M. Patel. Pushing the limits of unconstrained face detection: A challenge dataset and baseline results. In *IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–10. IEEE, 2018. doi: 10.1109/BTAS.2018.8698561.

National Highway Traffic Safety Administration. Fsd collisions in reduced roadway visibility conditions. https://static.nhtsa.gov/odi/inv/2026/INOA-EA26002-10023.pdf, 2026.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. OpenAI Blog, July 2026. URL https://openai.com/index/gpt-5-6/.

Minsoo Park, Almo Senja Kulinan, Dai Quoc Tran, Jinyeong Bak, and Seunghee Park. Preventing falls from floor openings using quadrilateral detection and construction worker pose-estimation. *Automation in Construction*, 165:105536, September 2024. ISSN 0926-5805. doi: 10.1016/j.autcon.2024.105536. URL http://dx.doi.org/10.1016/j.autcon.2024.105536.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman H. Khan, and Fahad Shahbaz Khan. PromptIR: Prompting for all-in-one image restoration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Miroslav Purkrábek and Jiri Matas. Probpose: A probabilistic approach to 2d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27124–27133, June 2025.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Sudarshan Rajagopalan, Kartik Narayan, and Vishal M. Patel. RestoreVAR: Visual autoregressive generation for all-in-one image restoration. In *International Conference on Learning Representations (ICLR)*, 2026.

Matteo Ruggero Ronchi and Pietro Perona. Benchmarking and error diagnosis in multi-instance pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 369–378, October 2017.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10745–10755. IEEE, 2021.

Christos Sakaridis, Haoran Wang, Ke Li, René Zurbrügg, Arpit Jadon, Wim Abbeloos, Daniel Olmeda Reino, Luc Van Gool, and Dengxin Dai. ACDC: the adverse conditions dataset with correspondences for robust semantic driving scene perception. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48 (3):2970–2988, 2026. doi: 10.1109/TPAMI.2025.3633063.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Trans. Mach. Learn. Res.*, 2026. URL https://openreview.net/forum?id=2NlGyqNjns.

Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023.

Yong Su, Defang Chen, Meng Xing, Changjae Oh, Xuewei Liu, and Jieyang Li. Coming out of the dark: Human pose estimation in low-light conditions. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1882–1890, 2025. doi: 10.24963/ijcai.2025/210.

Aditi Tiwari, Farzaneh Masoud, Dac Trong Nguyen, Jill Kraft, Heng Ji, and Klara Nahrstedt. Fire360: A benchmark for robust perception and episodic memory in degraded 360° firefighting video. In *NeurIPS*, 2025.

Dongkai Wang, Shiyu Xuan, and Shiliang Zhang. Locllm: Exploiting generalizable human keypoint localization via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 614–623, June 2024a.

Hainuo Wang, Qiming Hu, and Xiaojie Guo. MODEM: A morton-order degradation estimation mechanism for adverse weather image recovery. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Wenjing Wang, Huan Yang, Jianlong Fu, and Jiaying Liu. Zero-reference low-light enhancement via physical quadruple priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26057–26066, 2024b. doi: 10.1109/CVPR52733.2024.02462.

Xinzhe Wang, Kang Ma, Qiankun Liu, Yunhao Zou, and Ying Fu. Multi-object tracking in the dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 382–392, 2024c. doi: 10.1109/CVPR52733.2024.00044.

Ruiqi Wu, Zheng-Peng Duan, Chun-Le Guo, Zhi Chai, and Chongyi Li. RIDCP: Revitalizing real image dehazing via high-quality codebook priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22282–22291, 2023. doi: 10.1109/CVPR52729.2023.02134.

Xin Xu, Xinchao Wei, Yuelei Xu, Zhaoxiang Zhang, Kun Gong, Huafeng Li, and Leibing Xiao. Infpose: Real-time infrared multi-human pose estimation for edge devices based on encoder-decoder CNN architecture. *IEEE Robotics and Automation Letters*, 9(4):3672–3679, 2024a. doi: 10.1109/LRA.2023.3303070.

Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. ARID: A new dataset for recognizing action in the dark. In *Deep Learning for Human Activity Recognition*, Communications in Computer and Information Science, pp. 70–84. Springer Singapore, 2021. doi: 10.1007/978-981-16-0575-8_6.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2): 1212–1230, 2024b. doi: 10.1109/TPAMI.2023.3330016.

Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-Retinex: Rethinking low-light image enhancement with a generative diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12268–12277, 2023. doi: 10.1109/ICCV51070.2023.01130.

Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 889–898, June 2019.

Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Comput. Surv.*, 56 (1):11:1–11:37, 2024.

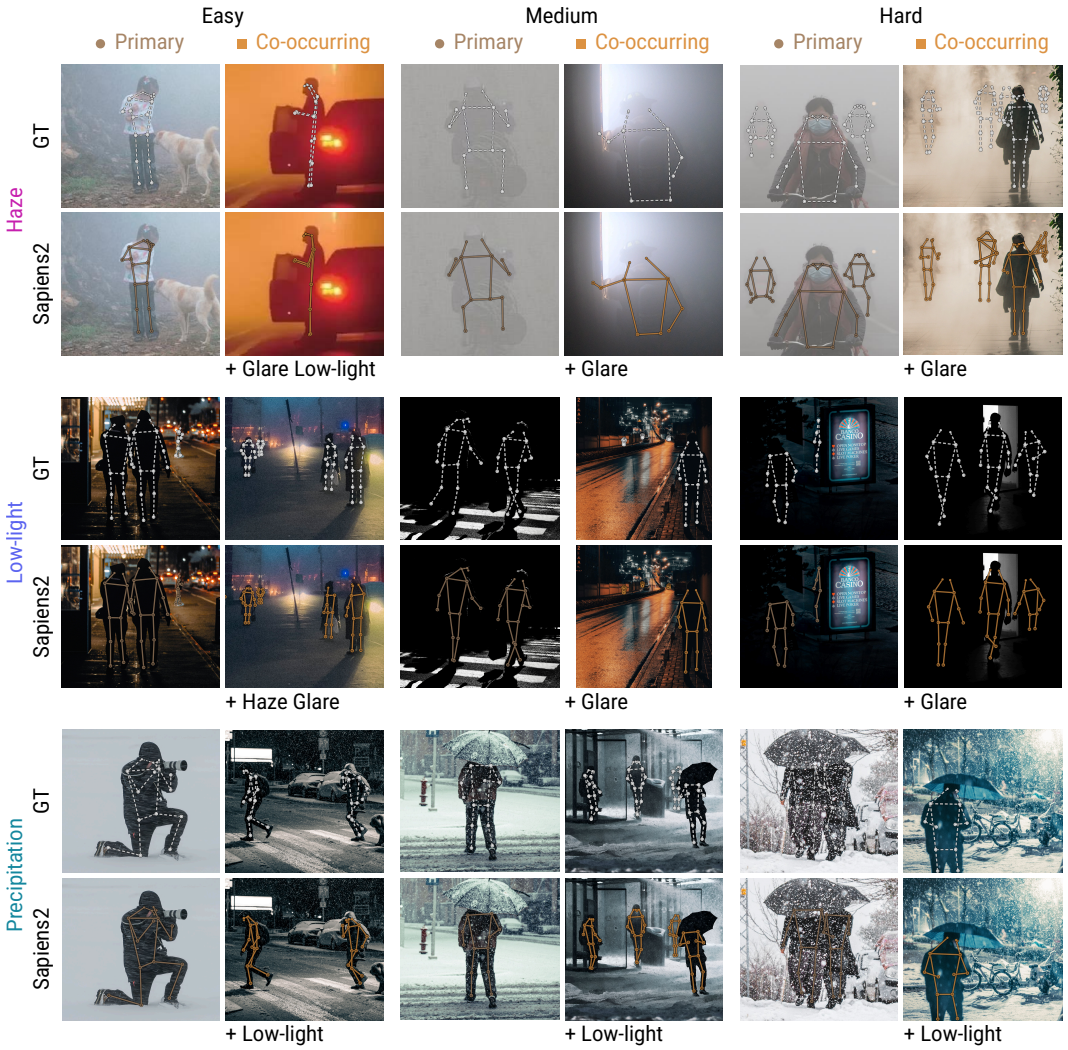


Figure 6: Qualitative results on 18 LV-POSE crops, one single-condition and one co-occurring image per condition-by-severity cell. Each condition block shows ground truth in the upper row and Sapiens2 predictions on ground-truth boxes in the lower row. Columns run Easy, Medium and Hard, single-condition before co-occurring; "+" lists other conditions.

A QUALITATIVE RESULTS

Figure 6 shows ground truth and Sapiens2 predictions on one single-condition and one co-occurring image from each condition-by-severity cell. These 18 images, with corresponding complete ground truth, are the review sample in the supplementary material. Per-cell accuracy of all models is in Table 2, and single-condition versus co-occurring comparisons per condition are in Tables 1, 4 and 5.

B ADDITIONAL COMPOUNDING RESULTS

We provide additional results of compounding effect on Low-light and Precipitation, which show similar trends as discussed in Section 3.2.

Table 4: mAP under Low-light/Night for co-occurring and single-condition images; see Section 3.2. **Bold** marks the column leader and every model that a paired bootstrap cannot separate from it.

Model	Co-occurring$_{258}$				Low-light/Night only$_{201}$			
	All$_{258}$	Easy$_{88}$	Med.$_{111}$	Hard$_{59}$	All$_{201}$	Easy$_{59}$	Med.$_{71}$	Hard$_{71}$
ViTPose++-H	52.7	**71.8**	53.1	32.2	60.2	**80.7**	60.5	42.6
ProbPose	40.3	57.9	42.1	20.8	50.1	69.7	50.2	34.3
CIGPose-L	45.5	65.0	46.6	23.7	51.8	71.9	53.2	34.2
GroupPose	34.0	46.4	34.7	21.4	32.8	48.6	33.8	20.2
YOLO26x-pose	26.3	41.7	26.2	13.0	26.3	40.3	28.3	14.2
Sapiens2	**56.7**	**72.4**	**58.6**	**39.3**	**65.7**	**82.9**	**67.7**	**50.1**
GKDT-H	44.2	63.6	44.8	23.5	52.0	73.8	51.1	34.5
LocLLM	50.2	66.2	52.3	30.2	58.6	76.3	60.6	42.2
Rex-Omni	20.1	27.7	23.1	9.7	23.1	40.3	23.4	10.1
SenseNova-Vision	24.2	31.4	29.0	11.6	28.8	48.5	29.5	13.2
GPT-5.6-Sol	6.8	7.9	8.1	3.9	7.1	14.8	6.7	2.7
Claude-Opus-5	8.4	10.8	8.7	5.6	9.5	16.2	9.2	4.8
Gemini-3.8-Flash	8.7	10.2	10.7	3.9	9.7	17.3	10.2	3.9
Qwen-3.8	11.0	14.2	12.7	5.9	13.7	22.2	15.5	6.2

Table 5: mAP under Precipitation for co-occurring and single-condition images; see Section 3.2. **Bold** marks the column leader and every model that a paired bootstrap cannot separate from it.

Model	Co-occurring$_{128}$				Precipitation only$_{275}$			
	All$_{128}$	Easy$_{35}$	Med.$_{50}$	Hard$_{43}$	All$_{275}$	Easy$_{86}$	Med.$_{100}$	Hard$_{89}$
ViTPose++-H	**61.9**	**77.4**	**65.3**	**51.3**	**73.0**	**80.8**	71.6	**70.0**
ProbPose	49.8	63.8	52.8	40.1	59.7	69.4	57.9	56.5
CIGPose-L	53.0	70.0	57.6	40.9	65.3	73.1	64.9	62.0
GroupPose	37.8	54.2	39.8	28.4	46.0	46.1	50.9	41.4
YOLO26x-pose	32.1	50.1	32.0	24.5	43.5	43.2	48.8	39.3
Sapiens2	**60.5**	**72.8**	**62.7**	**52.8**	**72.8**	**81.2**	**74.6**	66.7
GKDT-H	52.6	71.7	53.9	43.0	64.2	67.1	65.9	61.7
LocLLM	56.1	69.5	58.9	47.6	66.5	75.2	65.9	62.5
Rex-Omni	25.9	37.4	27.8	19.6	31.1	29.9	33.9	29.6
SenseNova-Vision	28.6	38.8	30.6	23.0	39.6	41.2	40.8	38.0
GPT-5.6-Sol	11.6	11.5	12.9	10.8	14.9	14.5	14.3	15.9
Claude-Opus-5	12.0	13.7	12.0	11.5	16.4	14.8	16.0	17.9
Gemini-3.8-Flash	11.0	12.3	11.1	10.6	17.5	16.4	17.2	18.3
Qwen-3.8	13.8	17.3	13.8	12.3	18.2	21.3	16.3	18.5

C Additional restoration results

Table 6 reports mAP@.50:.95 for each model using three dehazing methods—CoA (Ma et al., 2025), IPC (Fu et al., 2025), and DehazeFormer (Song et al., 2023)—under the same Easy/Medium/Hard severity levels used in Table 2. Subscripts indicate the absolute change relative to the corresponding no-preprocessing haze score, where green denotes improvement, red denotes degradation, and neutral denotes unchanged performance.

For HPE-specialized methods (GroupPose, LocLLM, Sapiens2, and so on), dehazing preprocessing rarely improves pose robustness. Among the 72 dehazing results in this group, 68 show performance degradation, while the remaining four gains are both smaller than 2 mAP. The negative effect also becomes substantially larger as visibility conditions become harder. At Easy severity, the largest performance drop is 5.6 mAP, while at Hard severity the degradation becomes much more severe. For example, LocLLM with DehazeFormer drops from 35.6 to 27.4 (-8.2), and Sapiens2 drops from 37.7 to 29.8 (-7.9). These results suggest that image restoration does not reliably transfer into downstream pose robustness for HPE-specialized models, especially under severe low-visibility conditions where restoration is expected to help the most.

Table 6: Pose estimation under haze with dehazing preprocessing, stratified by severity. WL is the well-lit baseline. Restoration methods are CoA (Ma et al., 2025), IPC (Fu et al., 2025), and DehazeFormer (Song et al., 2023). Numbers are mAP@.50:.95 (↑); subscripts show the change from the no-preprocessing Haze/Fog score in Table 2 (green: improvement, red: degradation).

Model	WL	CoA			IPC			DehazeFormer		
		Easy_{128}	Med._{148}	Hard_{135}	Easy_{128}	Med._{148}	Hard_{135}	Easy_{128}	Med._{148}	Hard_{135}
ViTPose++-H	82.1	$73.1_{-4.5}$	$47.6_{-5.1}$	$28.2_{-5.8}$	$72.0_{-5.6}$	$47.5_{-5.2}$	$30.4_{-3.6}$	$75.0_{-2.6}$	$47.7_{-5.0}$	$25.3_{-8.7}$
ProbPose	76.5	$63.9_{-2.6}$	$37.0_{-1.6}$	$18.3_{-0.7}$	$63.5_{-3.0}$	$38.3_{-0.3}$	$20.5_{+1.5}$	$63.5_{-3.1}$	$36.9_{-1.8}$	$16.6_{-2.4}$
CIGPose-L	78.4	$67.8_{-2.5}$	$44.2_{-0.8}$	$23.8_{+0.4}$	$67.8_{-2.4}$	$41.4_{-3.6}$	$24.5_{+1.1}$	$68.7_{-1.5}$	$40.8_{-4.2}$	$22.1_{-1.3}$
GroupPose	73.6	$58.1_{-1.4}$	$34.7_{-2.8}$	$17.5_{-3.8}$	$57.0_{-2.5}$	$35.5_{-2.0}$	$19.6_{-1.7}$	$58.6_{-0.9}$	$32.9_{-4.7}$	$16.8_{-4.5}$
YOLO26x-pose	70.8	$52.1_{-3.6}$	$30.9_{-2.8}$	$15.1_{-1.8}$	$54.3_{-1.4}$	$33.5_{-0.3}$	$16.7_{-0.2}$	$53.9_{-1.9}$	$29.9_{-3.8}$	$13.9_{-3.0}$
Sapiens2	78.8	$73.7_{-2.7}$	$53.3_{-4.0}$	$32.3_{-5.4}$	$73.1_{-3.3}$	$48.8_{-8.6}$	$29.9_{-7.8}$	$74.9_{-1.6}$	$51.7_{-5.6}$	$29.8_{-7.9}$
GKDT-H	80.5	$61.6_{+1.1}$	$44.5_{-0.8}$	$26.2_{-2.5}$	$60.0_{-0.5}$	$41.6_{-3.7}$	$25.3_{-3.4}$	$56.5_{-4.0}$	$38.5_{-6.7}$	$22.7_{-6.0}$
LocLLM	77.4	$69.6_{-3.1}$	$47.7_{-4.5}$	$28.7_{-6.8}$	$67.3_{-5.4}$	$44.4_{-7.8}$	$30.4_{-5.1}$	$71.2_{-1.5}$	$46.8_{-5.4}$	$27.4_{-8.2}$
Rex-Omni	44.9	$33.4_{+4.8}$	$23.7_{+2.9}$	$10.8_{+2.2}$	$34.5_{+5.9}$	$23.9_{+3.1}$	$11.1_{+2.6}$	$29.1_{+0.6}$	$19.6_{-1.2}$	$7.2_{-1.3}$
SenseNova-Vision	58.3	$42.6_{+1.3}$	$27.9_{+2.1}$	$14.9_{+1.7}$	$43.4_{+2.1}$	$28.6_{+2.8}$	$14.8_{+1.6}$	$40.7_{-0.6}$	$26.0_{+0.3}$	$13.3_{+0.1}$
GPT-5.6-Sol	18.7	$13.3_{-0.4}$	$8.9_{+1.1}$	$4.2_{+2.1}$	$14.2_{+0.4}$	$9.5_{+1.7}$	$4.7_{+2.6}$	$13.5_{-0.3}$	$7.4_{-0.4}$	$2.4_{+0.2}$
Claude-Opus-5	19.0	$21.8_{+2.5}$	$12.5_{+3.0}$	$6.6_{+1.1}$	$21.0_{+1.7}$	$11.3_{+1.8}$	$7.2_{+1.7}$	$19.1_{-0.2}$	$11.5_{+2.0}$	$6.0_{+0.6}$
Gemini-3.8-Flash	32.5	$20.9_{+1.1}$	$13.5_{\pm 0.0}$	$7.4_{+2.1}$	$20.4_{+0.6}$	$11.8_{-1.7}$	$6.5_{+1.1}$	$19.8_{\pm 0.0}$	$12.9_{-0.6}$	$5.0_{-0.3}$
Qwen-3.8	29.7	$23.1_{+0.5}$	$13.8_{+1.2}$	$6.7_{+0.6}$	$23.1_{+0.4}$	$15.2_{+2.5}$	$7.2_{+1.1}$	$23.0_{+0.3}$	$12.7_{\pm 0.0}$	$5.9_{-0.3}$

The behavior differs for foundation and frontier VLMs at Easy severity. All Easy-level results across Rex-Omni, Qwen-3.8, and Gemini-3.8-Flash show improvements after dehazing preprocessing. Gemini-3.1 Pro with CoA improves by 5.0 mAP, while even the smallest gain (DehazeFormer on Rex-Omni) remains positive at $+0.8$. However, these gains largely disappear at Medium and Hard severity levels, where the changes become small and inconsistent, often slightly negative. Unlike HPE-specialized methods, the VLM performance changes remain relatively limited across severity levels, typically within ± 2 mAP. This suggests that restoration mismatch behaves differently across model classes: HPE-specialized models become increasingly sensitive to restoration preprocessing as severity increases, while VLMs exhibit smaller and more stable changes.

These results further support the main motivation of LV-Pose. A single aggregate haze evaluation can hide important robustness differences across both severity levels and model classes. For example, Easy-level VLM results may suggest that dehazing helps low-visibility pose estimation, while Hard-level HPE-specialist results show severe robustness degradation after preprocessing. Without severity-aware evaluation, these contradictory behaviors can be collapsed into a single overall conclusion that fails to reflect actual model behavior under challenging visibility conditions.

The supportable conclusion from Table 6 is therefore more limited and conditional. Dehazing preprocessing may provide small benefits for some VLMs under mild visibility degradation, but it becomes unreliable as severity increases and can substantially harm HPE-specialized models under severe haze conditions. More broadly, these experiments show that visually improved restoration outputs do not necessarily translate into improved downstream pose robustness. This finding is consistent with the restoration mismatch discussed in the introduction and related work, where restoration outputs may deviate from the training distributions expected by downstream pose-estimation models.

Table 7 reports 42 results each for BioIR (Cui et al., 2025), MODEM (Wang et al., 2025), and FoundIR-v2 (Chen et al., 2026b) across 14 models and three severity levels.

BioIR lowers pose accuracy at every severity. It lowers AP in 40 of its 42 results, including all 30 results of the HPE methods and VFMs. Its largest drops are at Hard: ViTPose++-H falls from 63.2 to 52.7 (-10.5), GKDT-H from 55.0 to 45.5 (-9.5), and Sapiens2 from 61.7 to 53.0 (-8.7). BioIR costs more AP at Hard than at Easy for 12 of the 14 models.

MODEM barely changes pose accuracy, and FoundIR-v2 gives mixed results. Every MODEM change lies between -1.1 and $+2.0$ mAP (22 gains, 19 losses, 1 unchanged). FoundIR-v2 has 15 gains, 25 losses, and 2 unchanged results, and its largest losses are at Easy: Sapiens2 (-6.2), Qwen-3.8 (-5.0), and ViTPose++-H (-4.7).

Table 7: Pose estimation under precipitation with restoration preprocessing, stratified by severity. WL is the well-lit baseline. Restoration methods are BioIR (Cui et al., 2025), MODEM (Wang et al., 2025), and FoundIR-v2 (Chen et al., 2026b). Numbers are mAP@.50:.95 (↑); subscripts show the change from the no-preprocessing Precipitation score in Table 2 (green: improvement, red: degradation).

Model	WL	BioIR			MODEM			FoundIR-v2		
		Easy_{118}	Med._{148}	Hard_{131}	Easy_{118}	Med._{148}	Hard_{131}	Easy_{118}	Med._{148}	Hard_{131}
ViTPose++-H	82.1	$77.4_{-2.3}$	$62.9_{-6.7}$	$52.7_{-10.5}$	$79.1_{-0.6}$	$70.4_{+0.8}$	$62.4_{-0.8}$	$75.0_{-4.7}$	$68.9_{-0.7}$	$61.3_{-1.9}$
ProbPose	76.5	$64.5_{-2.9}$	$52.0_{-4.2}$	$44.0_{-6.6}$	$67.9_{+0.4}$	$57.3_{+1.1}$	$51.0_{+0.4}$	$66.5_{-0.9}$	$57.8_{+1.6}$	$51.1_{+0.5}$
CIGPose-L	78.4	$68.8_{-3.5}$	$56.5_{-6.0}$	$47.0_{-7.2}$	$71.7_{-0.6}$	$63.5_{+1.1}$	$54.8_{+0.5}$	$68.5_{-3.7}$	$63.1_{+0.6}$	$54.1_{-0.2}$
GroupPose	73.6	$44.7_{-3.2}$	$42.4_{-4.4}$	$32.1_{-4.6}$	$47.9_{\pm 0.0}$	$47.8_{+1.0}$	$36.8_{+0.1}$	$47.9_{\pm 0.0}$	$47.9_{+1.0}$	$37.9_{+1.1}$
YOLO26x-pose	70.8	$40.9_{-4.0}$	$38.4_{-4.1}$	$30.2_{-3.8}$	$44.8_{-0.1}$	$43.9_{+1.4}$	$34.8_{+0.8}$	$46.7_{+1.9}$	$45.4_{+2.8}$	$35.4_{+1.4}$
Sapiens2	78.8	$75.8_{-2.8}$	$63.9_{-6.6}$	$53.0_{-8.7}$	$78.1_{-0.5}$	$70.8_{+0.2}$	$61.4_{-0.3}$	$72.4_{-6.2}$	$68.1_{-2.4}$	$59.9_{-1.8}$
GKDT-H	80.5	$64.0_{-4.1}$	$54.0_{-7.7}$	$45.5_{-9.5}$	$67.6_{-0.5}$	$62.8_{+1.1}$	$54.4_{-0.5}$	$64.3_{-3.9}$	$60.9_{-0.8}$	$53.7_{-1.3}$
LocLLM	77.4	$70.4_{-3.1}$	$59.3_{-4.3}$	$49.1_{-7.9}$	$72.6_{-0.8}$	$64.3_{+0.8}$	$57.3_{+0.3}$	$68.8_{-4.6}$	$62.7_{-0.9}$	$56.0_{-1.0}$
Rex-Omni	44.9	$28.4_{-3.7}$	$27.0_{-4.8}$	$20.6_{-5.3}$	$31.8_{-0.3}$	$31.7_{-0.1}$	$25.4_{-0.6}$	$33.5_{+1.4}$	$32.3_{+0.5}$	$27.2_{+1.2}$
SenseNova-Vision	58.3	$36.9_{-3.5}$	$31.1_{-5.9}$	$25.8_{-6.8}$	$40.0_{-0.4}$	$36.8_{-0.3}$	$32.9_{+0.3}$	$39.4_{-1.0}$	$37.2_{+0.2}$	$31.4_{-1.2}$
GPT-5.6-Sol	18.7	$13.6_{\pm 0.0}$	$10.7_{-3.1}$	$12.3_{-1.7}$	$12.5_{-1.1}$	$13.4_{-0.5}$	$13.7_{-0.3}$	$12.5_{-1.2}$	$14.6_{+0.7}$	$13.0_{-1.0}$
Claude-Opus-5	19.0	$14.9_{+0.3}$	$14.1_{-0.4}$	$12.3_{-3.2}$	$15.4_{+0.9}$	$16.0_{+1.6}$	$15.9_{+0.3}$	$13.7_{-0.8}$	$16.0_{+1.6}$	$14.7_{-0.8}$
Gemini-3.8-Flash	32.5	$13.5_{-2.0}$	$11.7_{-3.3}$	$14.1_{-1.5}$	$16.6_{+1.1}$	$16.2_{+1.1}$	$16.5_{+0.9}$	$14.4_{-1.1}$	$14.0_{-1.1}$	$14.8_{-0.8}$
Qwen-3.8	29.7	$19.3_{-0.8}$	$13.9_{-1.5}$	$14.4_{-1.7}$	$20.0_{-0.2}$	$17.4_{+2.0}$	$15.7_{-0.5}$	$15.2_{-5.0}$	$17.7_{+2.4}$	$16.2_{\pm 0.0}$

Precipitation restoration either lowers pose accuracy (BioIR, most at Hard) or changes it little (MODEM), and the method and the severity decide which. No restored result exceeds the best unprocessed score of its column in Table 2 by more than 0.2 mAP; the one result above it is Sapiens2 with MODEM at Medium (70.8 against 70.6). The effect of restoration on pose depends on the restorer, the severity, and the pose model, so preprocessing should be judged per severity rather than by one aggregate score.

D Pose estimation with predicted person boxes

The evaluator is the same as in Table 2: xtcocotools COCOeval for keypoints, `use_area=False`, person category only, maxDets=20. The three detectors have different designs: DINOv3 ViT-7B (Siméoni et al., 2026) with a COCO DETR head is a frozen-backbone foundation detector with a closed COCO vocabulary, SAM 3 (Carion et al., 2026) is open-vocabulary and is prompted with the text "person", and YOLO26x (Jocher et al., 2026) is a real-time single-stage detector. Each detector runs at its own released evaluation protocol: DINOv3 resizes the short side to 2048 with the aspect ratio preserved as stated in Section D.9 of its paper, SAM 3 resizes to 1008 by 1008 without preserving the aspect ratio, and YOLO26x runs at 640 letterboxed. Only the box source changes. Each model keeps its own cropping, flipping, scoring and non-maximum suppression. No absolute score threshold is applied to the detector boxes, following the mmpose top-down default. Each image's 60 highest-scoring boxes are passed to the pose model, a cap that is a compromise with compute and time and stays as close to the top-down convention as possible. SenseNova-Vision receives the top 20 boxes instead; its instance score is the box score and it applies no non-maximum suppression, so under maxDets=20 boxes beyond the twentieth are never scored and the cut changes nothing. A model with keypoint scores uses its own default COCO rule for fusing them with the box score where it defines one, and the mmpose default otherwise. A model without keypoint scores uses its own rule for turning the box score into an instance score where it defines one; otherwise the box score is the instance score. Rows marked † (Rex-Omni, SenseNova-Vision) emit no keypoint confidence. For LocLLM (‡), the ground-truth-box score uses its official instance score, the plain mean of its seventeen keypoint scores. Its official detection path ignores the box score, so here it uses the mmpose default, the box score times the mean of the keypoint scores above 0.2. Its change therefore also includes this scoring change. LV-Pose's ground-truth boxes were pre-filled by this same DINOv3 checkpoint and then adjusted by annotators. The reproduction rate quantifies this caveat: it is the fraction of ground-truth boxes a detector reproduces at IoU at or above 0.99, measured at a

budget of 28 detections per image matched across the three detectors. DINOv3 at 2048 reproduces 11.4%, the same DINOv3 checkpoint at 1536, the resolution the pre-fill ran at, reproduces 24.8%, YOLO26x reproduces 2.0% and SAM 3 reproduces 0.26%. The drop from 24.8% to 11.4%, for the same model at a different resolution, shows that these matches are inherited annotation geometry rather than detector skill. The DINOv3 block is therefore optimistic. SAM 3 and YOLO26x have no such relationship to the ground truth.

Each block of Table 8 reports the detector's box AP in its first row, followed by eight pose models with their change from ground-truth boxes. The cost of predicted boxes is bounded rather than a collapse. For the five models that score with their own keypoint confidence (ViTPose++-H, ProbPose, CIGPose-L, Sapiens2 and GKDT-H), the mean change over the nine cells lies between $+0.1$ and -4.0 AP across all three detectors. Detector quality shows across the three blocks. The mean change is largest with YOLO26x boxes for all eight models; for Sapiens2 it is -0.7 on DINOv3, -1.2 on SAM 3 and -3.4 on YOLO26x. Pose losses follow detection losses. Box AP falls from Easy to Hard under Haze/Fog and Low-light/Night, where SAM 3 scores 66.7, 57.1 and 39.5 at night, but far less under Precipitation, where SAM 3 scores 66.0, 69.4 and 67.9. Among these five models, the two largest drops are both at Low-light/Night Hard on YOLO26x boxes, where Sapiens2 loses 8.4 AP and ViTPose++-H loses 7.3 AP.

Predicted boxes are not uniformly worse. On DINOv3 boxes, ProbPose gains 2.5 AP at Haze/Fog Hard and Sapiens2 gains 1.0 AP at Haze/Fog Medium. The larger absolute drops of stronger models track their baselines. At Low-light/Night Hard on YOLO26x boxes, Sapiens2, ViTPose++-H, CIGPose-L and GKDT-H each lose 18 to 20% of their ground-truth-box AP (46.5 to 38.1, 39.1 to 31.8, 29.8 to 24.4 and 30.6 to 24.6). ProbPose loses 11%, from 28.7 to 25.5. The data give no sign that stronger models depend more on accurate boxes. The DINOv3 block of Table 8 is optimistic, since LV-Pose's ground-truth boxes were pre-filled by the same DINOv3 checkpoint. The change in the † and ‡ rows mixes the box effect with a scoring change, so those rows are not read as box effects alone.

Table 8: This table reports pose AP in each of the nine cells of Table 2 when each top-down model receives person boxes predicted by DINOv3 (Siméoni et al., 2026), SAM 3 (Carion et al., 2026) or YOLO26x (Jocher et al., 2026) instead of ground-truth boxes. The italic first row of each block, labeled with the detector's name, reports that detector's box AP (person category only) in the same mAP@.50:.95 metric. Numbers are mAP@.50:.95 (↑); subscripts in the pose rows show the change from the ground-truth-box score of the same model and cell (green: improvement, red: degradation). Rows marked † and ‡ change their scoring too; see Section D for details.

Model	Haze/Fog$_{411}$			Low-light/Night$_{392}$			Precipitation$_{397}$		
	Easy$_{128}$	Med.$_{148}$	Hard$_{135}$	Easy$_{118}$	Med.$_{148}$	Hard$_{126}$	Easy$_{118}$	Med.$_{148}$	Hard$_{131}$
DINOv3 ViT-7B	*70.0*	*72.3*	*58.5*	*76.7*	*68.1*	*48.5*	*72.4*	*78.8*	*78.3*
ViTPose++-H	76.0$_{-1.6}$	53.6$_{+0.9}$	34.7$_{+0.7}$	73.2$_{-2.0}$	55.3$_{-0.6}$	36.0$_{-3.1}$	76.6$_{-3.1}$	70.7$_{+1.1}$	62.4$_{-0.8}$
ProbPose	66.2$_{-0.3}$	39.2$_{+0.6}$	21.6$_{+2.5}$	61.4$_{-1.2}$	45.4$_{\pm 0.0}$	27.0$_{-1.7}$	67.2$_{-0.3}$	57.5$_{+1.3}$	50.8$_{+0.2}$
CIGPose-L	68.7$_{-1.6}$	46.0$_{+0.9}$	24.0$_{+0.6}$	66.4$_{-0.9}$	48.2$_{-0.7}$	27.1$_{-2.7}$	70.4$_{-1.8}$	62.3$_{-0.1}$	53.0$_{-1.2}$
Sapiens2	74.3$_{-2.1}$	58.3$_{+1.0}$	38.3$_{+0.6}$	77.5$_{-0.2}$	62.5$_{-1.6}$	42.8$_{-3.8}$	77.7$_{-0.9}$	71.0$_{+0.5}$	61.7$_{\pm 0.0}$
GKDT-H	58.8$_{-1.7}$	46.0$_{+0.8}$	26.4$_{-2.3}$	66.1$_{-1.7}$	45.7$_{-1.9}$	27.4$_{-3.2}$	64.3$_{-3.8}$	60.5$_{-1.2}$	53.8$_{-1.2}$
LocLLM‡	70.7$_{-2.0}$	50.6$_{-1.6}$	32.9$_{-2.6}$	70.1$_{-0.2}$	56.1$_{+0.2}$	32.2$_{-5.3}$	70.4$_{-3.0}$	61.4$_{-2.1}$	55.6$_{-1.4}$
Rex-Omni†	26.5$_{-2.0}$	17.3$_{-3.4}$	5.9$_{-2.6}$	24.6$_{-7.9}$	18.4$_{-5.1}$	7.1$_{-2.5}$	24.5$_{-7.6}$	26.4$_{-5.4}$	21.6$_{-4.3}$
SenseNova-Vision†	34.6$_{-6.8}$	23.4$_{-2.3}$	11.1$_{-2.2}$	32.0$_{-6.7}$	20.9$_{-8.7}$	9.7$_{-2.8}$	33.4$_{-7.0}$	31.1$_{-6.0}$	26.7$_{-5.9}$
SAM 3	*66.1*	*64.0*	*52.6*	*66.7*	*57.1*	*39.5*	*66.0*	*69.4*	*67.9*
ViTPose++-H	76.1$_{-1.5}$	53.0$_{+0.4}$	33.6$_{-0.4}$	72.9$_{-2.2}$	52.2$_{-3.7}$	36.3$_{-2.7}$	75.7$_{-4.0}$	68.8$_{-0.8}$	62.0$_{-1.2}$
ProbPose	66.2$_{-0.4}$	40.0$_{+1.4}$	19.7$_{+0.7}$	61.0$_{-1.7}$	44.5$_{-0.9}$	27.5$_{-1.2}$	66.5$_{-0.9}$	56.4$_{+0.2}$	49.4$_{-1.3}$
CIGPose-L	68.9$_{-1.3}$	44.7$_{-0.3}$	24.0$_{+0.6}$	65.7$_{-1.6}$	48.0$_{-0.9}$	27.3$_{-2.5}$	69.5$_{-2.8}$	61.4$_{-1.0}$	51.8$_{-2.4}$
Sapiens2	74.6$_{-1.9}$	57.3$_{\pm 0.0}$	38.7$_{+1.0}$	76.1$_{-1.6}$	61.8$_{-2.2}$	44.2$_{-2.4}$	76.4$_{-2.2}$	70.8$_{+0.2}$	60.2$_{-1.5}$
GKDT-H	57.5$_{-3.1}$	45.1$_{-0.1}$	25.9$_{-2.8}$	64.4$_{-3.3}$	42.9$_{-4.6}$	25.8$_{-4.8}$	62.5$_{-5.6}$	58.3$_{-3.4}$	52.3$_{-2.7}$
LocLLM‡	70.4$_{-2.2}$	48.9$_{-3.3}$	31.9$_{-3.6}$	68.2$_{-2.0}$	53.2$_{-2.7}$	33.3$_{-4.2}$	70.6$_{-2.9}$	61.2$_{-2.4}$	54.9$_{-2.1}$
Rex-Omni†	23.8$_{-4.8}$	16.7$_{-4.1}$	6.8$_{-1.7}$	25.7$_{-6.8}$	18.5$_{-4.9}$	6.5$_{-3.1}$	24.3$_{-7.9}$	26.4$_{-5.4}$	21.9$_{-4.0}$
SenseNova-Vision†	32.7$_{-8.6}$	23.7$_{-2.0}$	8.7$_{-4.5}$	31.0$_{-7.6}$	21.0$_{-8.6}$	9.8$_{-2.7}$	33.1$_{-7.4}$	30.7$_{-6.3}$	26.7$_{-5.9}$
YOLO26x	*68.5*	*62.1*	*42.5*	*68.3*	*56.9*	*39.0*	*67.6*	*68.4*	*62.5*
ViTPose++-H	75.5$_{-2.1}$	50.4$_{-2.3}$	30.7$_{-3.2}$	70.8$_{-4.3}$	50.0$_{-5.9}$	31.8$_{-7.3}$	74.1$_{-5.7}$	68.6$_{-1.0}$	58.8$_{-4.4}$
ProbPose	66.6$_{\pm 0.0}$	39.5$_{+0.9}$	19.6$_{+0.6}$	58.1$_{-4.6}$	41.4$_{-4.0}$	25.5$_{-3.3}$	64.7$_{-2.8}$	55.4$_{-0.8}$	47.3$_{-3.3}$
CIGPose-L	68.8$_{-1.5}$	44.3$_{-0.8}$	23.0$_{-0.4}$	63.3$_{-4.1}$	44.5$_{-4.4}$	24.4$_{-5.4}$	69.3$_{-3.0}$	59.7$_{-2.7}$	49.9$_{-4.4}$
Sapiens2	74.6$_{-1.9}$	55.0$_{-2.4}$	34.9$_{-2.8}$	74.3$_{-3.3}$	58.5$_{-5.5}$	38.1$_{-8.4}$	76.0$_{-2.6}$	69.9$_{-0.7}$	58.8$_{-2.9}$
GKDT-H	58.2$_{-2.3}$	44.8$_{-0.5}$	24.2$_{-4.5}$	63.1$_{-4.7}$	42.6$_{-4.9}$	24.6$_{-6.0}$	63.2$_{-4.9}$	58.7$_{-3.0}$	50.5$_{-4.5}$
LocLLM‡	69.2$_{-3.5}$	48.6$_{-3.6}$	28.5$_{-7.0}$	64.5$_{-5.7}$	49.8$_{-6.1}$	28.5$_{-9.0}$	67.7$_{-5.8}$	60.0$_{-3.6}$	50.8$_{-6.2}$
Rex-Omni†	26.6$_{-1.9}$	16.7$_{-4.1}$	6.7$_{-1.8}$	22.8$_{-9.8}$	17.9$_{-5.5}$	5.8$_{-3.8}$	24.1$_{-8.1}$	26.2$_{-5.6}$	21.2$_{-4.7}$
SenseNova-Vision†	34.8$_{-6.5}$	21.4$_{-4.3}$	9.7$_{-3.5}$	29.9$_{-8.8}$	21.6$_{-8.1}$	9.6$_{-2.9}$	32.0$_{-8.4}$	29.0$_{-8.1}$	24.8$_{-7.8}$

E LIMITATION

LV-Pose focuses on diagnostic evaluation under real-world low-visibility conditions rather than exhaustive coverage of all deployment scenarios. First, although LV-Pose includes ~~seven~~ six visibility-condition tags, the current benchmark size does not support fine-grained analysis for every individual secondary condition or every possible combination of co-occurring degradations. Our compositional analysis therefore focuses on the overall impact of compounded visibility degradation rather than condition-specific causal attribution.

Second, the current evaluation focuses on static-image 2D human pose estimation. Temporal robustness in video settings, multi-view pose estimation, 3D pose estimation, and multi-person tracking under low-visibility conditions remain outside the scope of this work. Third, while LV-Pose evaluates representative HPE models, vision-language models, and multimodal foundation models, the rapidly evolving landscape of foundation models means that future architectures may exhibit different robustness behaviors under severe low-visibility conditions. Our conclusions should therefore be interpreted as diagnostic observations on the evaluated model classes rather than universal claims about all future multimodal systems.

Finally, LV-Pose emphasizes real-image evaluation and human-annotated visibility difficulty levels. Although this design improves realism and supports severity-aware analysis, it also limits dataset scale compared with automatically generated or synthetically rendered benchmarks. Expanding low-visibility evaluation coverage while preserving annotation quality remains an important direction for future work.

F ANNOTATION AGREEMENT AND ADJUDICATION

Two independent annotators label every image without seeing each other's pass, and a reviewer checks both annotations and resolves disagreements instead of averaging keypoint coordinates. This is how individual-annotator noise and bias are kept out of the ground truth. Before review, the two passes agree within one COCO standard deviation (1 SD) as often as COCO annotators do (76.1% versus 75.0%; Section F.1). Reviewer intervention grows with disagreement (Section F.2): in the third of images with the lowest pre-review agreement, 56.4% have keypoint edits, against 17.0% in the highest third, and when the two passes are more than 2 SD apart, reviewers take the midpoint for only 15.5% of keypoints. Hard images are held for a second reviewer more often (18.6% versus 4.1% of easy images; Section F.3), and for keypoints more than 2 SD apart, reviewers take the midpoint less often in them (11.3% for hard versus 28.1% for easy images).

This appendix covers the 1,200 LV-Pose images and 3,855 persons of the ground truth. The keypoint similarity k_s follows Ronchi & Perona (2017). Comparisons between the two annotators use the mean of their two box areas, and comparisons involving a ground truth use the ground-truth box. Thresholds of $k_s \geq 0.88$ and $k_s \geq 0.61$ correspond to one and two COCO human standard deviations (SD) (Ronchi & Perona, 2017), written as within 1 SD and within 2 SD. Displacement $u = d/(\sigma_j s)$ is measured in COCO standard-deviation units. The 19 keypoints are the COCO-17 keypoints plus top of head and neck; k_s uses only the COCO-17 keypoints, and every other annotator statistic uses all 19. Normalized localization error (NLE) is the distance between the two annotators' placements divided by the mean diagonal of their two person boxes.

F.1 AGREEMENT BETWEEN THE TWO ANNOTATORS

Following Ronchi & Perona (2017), agreement is measured before review, between the two annotator passes, on the 49,889 COCO-17 keypoints both annotators labeled. At 1 SD, LV-Pose agreement matches the COCO reference: 76.1% of these keypoints lie within 1 SD ($k_s \geq 0.88$), against 75.0% for COCO annotators. Of the same keypoints, 88.3% lie within 2 SD ($k_s \geq 0.61$) and 92.8% within 3 SD ($k_s \geq 0.32$), against 95.0% and 98.7% for COCO annotators.

Localization differences are small for most keypoints. Over the 57,453 jointly labeled keypoints of all 19 keypoints, the median NLE is 0.0136; 80.2% lie within 0.05 of the box diagonal and 90.6% within 0.10. The two annotators give the same visibility state to 80.3% of the 73,834 keypoint slots, and both label 77.8% of them.

F.2 Adjudication between the two annotators

Every LV-Pose image was reviewed. Of the 1,200 images, 1,074 (89.5%) passed by one reviewer and 126 (10.5%) were held for a second reviewer, who continues from the first reviewer's draft.

The base is the version of the person boxes and keypoints that the reviewer loads and edits. Reviewers choose from four versions: the screened version (boxes checked at screening, keypoints as initialized by Sapiens before annotation), the first annotator's pass, the second annotator's pass, and the average of the two passes, which is the default. Reviewers chose the average for 737 images (61.4%) and one annotator's pass for 427 (35.6%); the remaining 36 images, all held, continued from the first reviewer's draft.

Of the images and persons with a chosen base, 37.1% of images and 20.6% of persons have at least one keypoint edit. Among keypoints labeled in both the base and the final ground truth, 4.0% were moved and 1.1% changed between occluded and visible. Reviewers removed 1.8% of labeled keypoints and added 0.17%. Most moves are small, with a median displacement of $u = 0.64$ (under 1 SD).

Adjudication follows the disagreement between the two passes. Images are split into thirds by pre-review agreement, the image mean k_s over the COCO-17 keypoints both annotators labeled. In the lowest third (image mean k_s up to 0.86), reviewers started from one annotator's pass for 55.7% of images, edited keypoints in 56.4% and held 17.0% for a second reviewer. In the highest third (above 0.96), these shares are 16.2%, 17.0% and 0.5%. Per keypoint (Table 9), reviewers mostly keep one pass when the two passes lie within 1 SD (69.6%). As the passes move apart, the midpoint share falls (37.2% at 1–2 SD and 15.5% beyond 2 SD) and new placements rise (3.1% within 1 SD, 13.4% at 1–2 SD and 26.1% beyond 2 SD). Keypoints labeled by only one annotator are left unlabeled in 60.8% of cases.

Table 9: Final keypoint placement by annotator disagreement. Rows group COCO-17 keypoints labeled by at least one annotator, on persons with both passes, by the distance between the two passes. One pass kept means the final keypoint equals one of the two passes, New placement is anywhere else, and Removed means left unlabeled. Row shares sum to 100.

Two passes	Keypoints	One pass kept	Midpoint	New placement	Removed
Within 1 SD	37,772	69.6	26.8	3.1	0.5
1–2 SD apart	5,947	47.1	37.2	13.4	2.2
More than 2 SD apart	5,760	53.2	15.5	26.1	5.2
Labeled by one pass only	7,906	34.5	–	4.7	60.8

F.3 Handling of hard cases

Hard images contain more hard cases and need more reviewer involvement. The share of images held for a second reviewer is 4.1% for easy, 8.6% for medium and 18.6% for hard images. The share of images with at least one keypoint edit is 27.8% for easy, 41.7% for medium and 40.8% for hard images.

When the first reviewer cannot settle an image, it is held for a second reviewer, who continues from the first reviewer's draft; the 126 held images were reviewed this way. Reviewers start from one annotator's pass instead of the default average on 21.4% of easy, 39.6% of medium and 44.1% of hard images. For keypoints where the two passes are more than 2 SD apart, reviewers take the midpoint for 28.1% in easy, 14.8% in medium and 11.3% in hard images (all severities: Table 9). They give a new placement to about a quarter of these keypoints at every severity: 28.2% for easy, 23.8% for medium and 27.0% for hard images.

G Sapiens initialization and evaluation

The annotation pipeline limits the influence of the Sapiens initialization, and the final ground truth gives the evaluated Sapiens2 no higher score than a ground truth re-annotated without initialization. Two independent annotators refine every initial keypoint without seeing each other's pass, and a

reviewer resolves their disagreements instead of averaging keypoint coordinates. Where both re-annotation passes without initialization place a keypoint far from the initial keypoint and agree with each other, the final ground truth keeps the initial keypoint for 4 of 118 keypoints (Section G.2). On a re-annotation of 180 images without initialization, Sapiens2 ranks first under both sets of ground-truth annotations and scores no higher on the final ground truth (64.3 against 65.9 AP); the order of the eight models changes only by one swap of two adjacent models (Kendall τ 0.929; Section G.3). Moving the 70 kept initial keypoints that could carry a Sapiens-specific imprint to their re-annotated positions changes Sapiens2's AP relative to the other seven models by -0.09 $[-0.33, +0.19]$. Conclusions that rest on Sapiens2's lead over the other models therefore hold under both ground truths, and conclusions that hold for the non-Sapiens models cannot be produced by a Sapiens-specific effect.

G.1 Initialization and re-annotation audit

The initialization B is the output of Sapiens (Khirodkar et al., 2024) (the 2B model) on candidate person boxes, which are the source dataset's boxes where available and DINOv3 (Siméoni et al., 2026) detections otherwise. Every initial keypoint starts as visible ($v = 2$). In the pipeline of Section 2.2, both annotators start from the same initial keypoints and refine them without seeing each other's pass. A reviewer resolves their disagreements into the final ground truth C, starting from one of four versions: the initial keypoints, either annotator's pass, or the average of the two passes (Section F.2).

The re-annotation audit covers 180 selected images with 272 persons. The same two annotators re-annotated these persons starting from a fixed template pose instead of the Sapiens initialization, giving passes G and H. A different reviewer reviewed them into F. The re-annotation was carried out after all images had been annotated. It used the earliest-annotated batch of images and came after a sufficiently long interval. The keypoint-level analysis covers the 207 persons whose original annotation started from the initialization and whose left and right labels match between the two ground truths C and F.

We compare keypoint positions by the per-keypoint OKS similarity k and the scale-free distance u:

$$k_j(X, Y) = \exp\left(-\frac{\|X - Y\|^2}{2\,(2\sigma_j s)^2}\right) = \exp\left(-\frac{u^2}{8}\right), \qquad u = \frac{\|X - Y\|}{\sigma_j s}. \tag{1}$$

The scale is $s = \sqrt{0.53\,wh}$ of the final ground-truth box. The per-keypoint OKS similarity k uses the COCO keypoint sigmas σ_j. u divides the distance by $\sigma_j s$, the scale that OKS uses in the evaluation, so it counts COCO standard deviations (SD) of the keypoint.

G.2 How the pipeline limits the initialization

The pipeline edits the initialization where independent annotators place the keypoint elsewhere. In the 207-person cohort, a strict screen, $k(B, G) < 0.50$, $k(B, H) < 0.50$ and $k(G, H) \geq 0.90$, selects 120 keypoints. This screen selects keypoints where both re-annotation passes place the keypoint far from the initial keypoint ($k < 0.50$ means more than 2.35 SD) and agree with each other. Of the 118 labeled in C, 4 keep the initial keypoint exactly. When both annotators moved an initial keypoint during annotation, the reviewer never restored it (0 of 932). On the full benchmark, among persons whose original annotation started from the initialization (1,083 images), the final ground truth marks 10.0% of the initial keypoint slots as occluded, a change that only an annotator could make.

The initial keypoints that remain sit within the precision of human placement. Where the final ground truth C keeps the initial keypoint exactly, the kept position lies a median of 0.51 SD from h, the midpoint of the two re-annotation passes G and H. The two re-annotation passes lie a median of 0.65 SD apart. Under independent click noise, a third independent annotator would be expected about $0.866 \times 0.65 \approx 0.56$ SD from their midpoint, so the kept initial keypoints sit closer to the re-annotated placements than a third annotator would.

G.3 Evaluation ranking

On the 180 audit images with 272 persons, the model ranking barely changes between the two ground truths (Table 10). Kendall τ over the eight main-table models is 0.929. Only ViTPose++-H and

LocLLM swap places, and Sapiens2 is first under both C and F. Sapiens2 scores 64.3 AP on C and 65.9 on F, so the final ground truth gives it no higher AP than the re-annotated one.

Table 10: AP$\times$100 and ranks of the eight models evaluated on the final ground truth C and the ground truth F re-annotated without initialization across the 180 audit images (272 persons).

Model	AP on C	AP on F	Rank on C	Rank on F
Sapiens2	64.3	65.9	1	1
ViTPose++-H	61.3	59.4	2	3
LocLLM	60.9	59.9	3	2
GKDT-H	59.2	57.8	4	4
CIGPose-L	55.6	53.6	5	5
ProbPose	48.4	46.9	6	6
GroupPose	46.0	44.9	7	7
YOLO26x-pose	41.6	39.9	8	8

We also ask whether the kept initial keypoints could favor Sapiens2 in the evaluation. A kept initial keypoint can carry a Sapiens-specific imprint only where it disagrees both with the re-annotation midpoint h and with the other models, taken as the median prediction of the five non-Sapiens top-down models (ViTPose++-H, LocLLM, CIGPose-L, GKDT-H and ProbPose), with $k < 0.90$ to both. In the 207-person cohort, 71 kept initial keypoints meet this condition. Moving the 70 of them that are labeled in F to their positions in F and recomputing AP changes Sapiens2 relative to the mean of the other seven models by -0.09 AP (90% BCa interval $[-0.33, +0.19]$ from a whole-image bootstrap). The ranking of the eight models stays the same as on C (Kendall $\tau = 1.000$).

H Comparability of single-condition and co-occurring images

Co-occurring images score lower than single-condition images of the same condition and severity, and in eight of the nine condition-by-severity cells the image covariates do not account for this. The covariate differences mostly favor the co-occurring images, which hold larger persons at higher resolution and show the lower local visual difficulty on their primary condition (Section H.1). On the covariates alone, co-occurring images are predicted to score higher in seven of the nine cells (predicted gaps -0.9 to -5.0 OKS points), yet they score lower than predicted in eight of the nine cells (Section H.2). Across the 14 models and nine cells, co-occurring AP is lower in 114 of the 126 comparisons, and in none does the interval of the gap lie entirely below zero (Section H.3).

The analysis uses the exact grouping of the compounding tables in Section 3.2 (Tables 1, 4 and 5). Within each primary condition and severity, single-condition images carry only the primary condition tag and co-occurring images also carry one or more secondary condition tags. No image is re-selected, reweighted or dropped. All results use the final ground truth and the 14 models in the compounding tables.

Each covariate is computed per image, with person-level values averaged over the labeled persons in the image. The five image covariates are persons, scale, resolution, crowding and truncation. Persons is the log number of labeled persons. Scale is the mean log person scale, where scale is the OKS scale $\sqrt{0.53\,wh}$ of the ground-truth box in pixels. Resolution is the log image resolution. Crowding is the CrowdPose crowd index (Li et al., 2019b), the keypoints of other persons inside a person's box divided by the person's own labeled keypoints. Truncation is the share of persons whose box touches the image border. Degradation is reported separately and measures the local visual difficulty of the primary condition around visible keypoints, with the patch metrics of Figure 3: FADE (Choi et al., 2015) for haze/fog, LIME (Guo et al., 2017) for low-light/night (inverted, so higher values are darker), and the JSTASR small-scale SIM (Chen et al., 2020a) for precipitation. Because AP is computed over a set of images and cannot be assigned to single persons, the covariate model predicts per-person OKS, while the test uses AP, the metric of the compounding tables.

H.1 How the two groups differ

The covariate differences mostly favor the co-occurring images. The standardized mean difference (SMD) is the single-condition mean minus the co-occurring mean over the pooled standard deviation.

Co-occurring images hold larger persons (scale SMD -0.26 to -1.09 in seven of the nine cells) at higher resolution (SMD -0.12 to -0.90 in eight of the nine cells). The covariate model of Section H.2 weighs all covariate differences together. On the local visual difficulty of the primary condition, single-condition images are the harder group in eight of the nine cells (SMD 0.29 to 1.12), so the lower scores of co-occurring images do not come with a harder primary condition.

H.2 Covariate-predicted and observed gaps

Table 11: Observed, covariate-predicted and residual OKS gaps. The outcome is per-person OKS ($\times 100$) to the model's best-matching detection, averaged over the 14 models. A linear model fitted on single-condition persons predicts it from the image covariates and severity fixed effects. All columns are single-condition minus co-occurring; positive values mean co-occurring images score lower. Residual is observed minus predicted. A residual in bold has a 90% percentile interval above zero, from 2,000 whole-image bootstrap resamples that refit the model.

Condition	Severity	Observed	Predicted	Residual
Haze/Fog	Easy	6.6	-0.9	**7.6**
	Med.	10.7	-2.6	**13.3**
	Hard	10.4	-3.7	**14.0**
Low-light/Night	Easy	4.9	-1.9	**6.8**
	Med.	2.5	-2.9	**5.4**
	Hard	5.1	-3.1	**8.3**
Precipitation	Easy	-0.1	-5.0	**4.9**
	Med.	6.4	4.5	1.9
	Hard	11.3	2.4	**8.9**

A linear model of per-person OKS ($\times 100$, to the model's best-matching detection, averaged over the 14 models) is fitted on the single-condition persons of each condition over all three severities and applied to both groups. Its inputs are person scale, crowd index, truncation, largest box IoU with another person, person count, image resolution and severity fixed effects. The predicted gap is the gap that the covariates alone imply. The covariates predict co-occurring images to score higher in seven of the nine cells, with predicted gaps of -0.9 to -5.0 OKS points. Observed gaps are positive in eight of the nine cells, from 2.5 to 11.3. The residual, observed minus predicted, ranges from 1.9 to 14.0, and its 90% interval from 2,000 whole-image bootstrap resamples, with the model refitted, lies above zero in eight of the nine cells (Table 11). Co-occurring images therefore score lower than their covariates predict in eight of the nine cells. With fixed effects for the image source added, the residual interval stays above zero in the same eight cells.

H.3 Consistency of the gap across models

For each of the 14 models in each of the nine cells (126 comparisons), the gap is single-condition AP minus co-occurring AP. Its two-sided 90% BCa interval is computed from the 10,000 whole-image bootstrap resamples behind the compounding tables (Tables 1, 4 and 5) and the leave-one-image-out jackknife of each group. Co-occurring AP is lower in 114 comparisons, the interval excludes zero in 77, and no interval lies entirely below zero.

I Comparison with existing evaluation settings

Table 12 compares LV-Pose with nine representative 2D HPE evaluation settings. Widely used benchmarks such as COCO val2017 (Lin et al., 2014), CrowdPose (Li et al., 2019b), and OCHuman (Zhang et al., 2019) primarily evaluate crowding, occlusion, and general pose estimation performance under relatively well-lit conditions. Recent low-visibility evaluations, including ExLPose and ExLPose-OCN (Lee et al., 2023), WIHPD (Xu et al., 2024a), LLVIP-Pose (Cormier et al., 2024), LLIP (Su et al., 2025), and EHPT-XC (Cho et al., 2024), extend coverage to one or two degradations such as low-light or motion blur, and typically analyze each visibility condition on its own. As summarized in Table 12, existing benchmarks may cover individual visibility conditions and their severity, but

Table 12: 2D HPE evaluations compared by which low-visibility conditions they cover, whether co-occurring degradations are tagged on the same image, and which severity levels real images are grouped into. ✓/✗ mark presence or absence. A / marks none or not applicable. † EHPT-XC is video-based, so frame count is not directly comparable to image-based evaluations.

Evaluation	Compositional degradations	Low visibilities	Severity levels (real images)	# Person /image	# Images
COCO-val2017	/	/	/	2.71	2,346
CrowdPose	/	/	/	3.64	8,000
OCHuman	/	/	/	1.72	2,500
ExLPose	✗	Low-light	Normal, Hard, Extreme	5.68	491
ExLPose-OCN	✗	Low-light	/	2.75	360
WIHPD	✗	Low-light	/	3.65	400
LLVIP-Pose	✗	Low-light	/	2.17	3,462
LLIP	✗	Low-light	Challenging, Difficult, Extreme, Formidable	1.00	4,814
EHPT-XC †	✗	Low-light, motion blur	/	2.48	/
Ours	✓	Haze/fog, low-light/night, precipitation, glare/backlight, motion blur, focus blur	Easy, Medium, Hard	3.21	1,200

do not combine severity annotation across multiple conditions with explicit tags for co-occurring degradations. As a result, it remains unclear whether robustness persists across increasing severity levels, under co-occurring degradations, or only within relatively easy visibility conditions.

Rather than maximizing dataset scale, LV-Pose focuses on diagnostic coverage across severity-aware and compositional low-visibility conditions. LV-Pose covers six visibility conditions on real low-visibility images: three primary conditions, haze/fog, low-light/night, and precipitation, and additional co-occurring degradation tags for glare/backlight, motion blur, and focus blur. Each annotated visibility condition is assigned one of three difficulty levels (easy, medium, or hard) based on keypoint visibility and human pose recoverability. The framework also explicitly tags co-occurring degradations on the same image, which enables controlled analysis of compounding robustness failures. The unit of evaluation therefore shifts from a single aggregate robustness score under "low visibility" to severity-conditioned and composition-conditioned robustness analysis. LV-Pose focuses on severity-aware and compositional low-visibility evaluation under the six annotated visibility conditions, rather than temporal robustness or visibility conditions outside the annotated categories.

J Additional AP@.75 results

We provide AP@.75 with 90% confidence intervals for the haze compounding and overall performance results, which show similar trends as discussed in Sections 3.2 and 3.3.

Table 13: AP@.75 with two-sided 90% BCa intervals (†: percentile interval) across the nine condition-by-severity cells plus a well-lit column; see Section 3.3.

Model	WL	Haze/Fog$_{411}$			Low-light/Night$_{392}$			Precipitation$_{397}$		
		Easy$_{128}$	Med.$_{148}$	Hard$_{135}$	Easy$_{118}$	Med.$_{148}$	Hard$_{126}$	Easy$_{118}$	Med.$_{148}$	Hard$_{131}$
ViTPose++-H	88.9	86.0 [82.8, 89.8]	56.3 [50.2, 61.4]	35.2 [28.0, 41.0]	82.2 [77.0, 86.0]	60.3 [55.0, 64.4]	42.5 [37.7, 47.6]	86.7 [82.0, 90.1]	76.2 [71.1, 79.6]	66.9 [61.5, 72.2]
ProbPose	83.7	72.7 [67.3, 77.4]	40.2 [34.0, 45.8]	19.3 [14.7, 24.5]	68.2 [62.8, 73.6]	46.8 [41.1, 51.0]	28.2 [23.2, 32.8]	72.5 [66.7, 77.5]	58.4 [54.1, 62.9]	52.4 [47.0, 57.6]
CIGPose-L	84.9	79.1 [74.6, 83.6]	48.7 [42.4, 54.1]	22.6 [17.8, 27.6]	72.8 [67.7, 77.3]	50.4 [44.9, 54.7]	29.8 [25.1, 33.9]	79.0 [74.3, 83.8]	67.7 [62.0, 72.1]	57.1 [51.4, 63.2]
GroupPose	80.5	66.6 [60.9, 72.0]	39.9 [34.4, 45.4]	22.0 [17.1, 27.2]	48.5 [42.4, 54.2]	37.7 [33.9, 41.7]	19.4 [16.5, 22.4]	52.2 [45.4, 58.8]	50.8 [46.1, 55.7]	39.7 [34.5, 45.3]
YOLO26x-pose	77.2	60.4 [53.9, 65.8]	37.0 [31.4, 42.4]	16.4 [12.0, 21.3]	40.2 [33.6, 46.1]	27.2 [23.2, 31.2]	12.1 [9.2, 14.8]	47.4 [40.4, 54.0]	44.2 [38.8, 49.1]	35.9 [30.8, 41.8]
Sapiens2	86.3	85.3 [80.5, 88.5]	60.9 [55.0, 65.9]	39.6 [31.2, 46.2]	83.2 [77.7, 86.9]	68.3 [63.2, 72.0]	51.1 [45.6, 55.9]	83.2 [78.2, 87.2]	75.7 [71.5, 80.0]	63.9 [58.0, 68.9]
GKDT-H	87.4	67.0 [62.2, 72.5]	46.3 [40.0, 51.7]	28.4 [21.7, 34.2]	74.4 [69.0, 78.7]	50.2 [44.9, 54.7]	28.7 [24.3, 32.6]	73.8 [67.6, 79.1]	67.3 [62.5, 72.1]	58.2 [52.9, 64.0]
LocLLM	84.4	81.0 [75.2, 85.4]	55.2 [49.9, 61.0]	37.1 [30.9, 43.4]	77.0 [72.1, 80.8]	59.7 [54.2, 63.7]	37.2 [32.1, 41.6]	80.5 [74.3, 85.9]	70.1 [66.2, 74.3]	61.3 [55.6, 66.5]
Rex-Omni	50.5	32.7 [27.7, 36.6]	23.8 [18.8, 27.7]	9.9 [5.9, 12.9]	37.6 [32.7, 41.6]	26.7 [22.8, 29.7]	10.9 [8.9, 13.9]	37.6 [31.7, 44.6]	34.7 [29.7, 38.6]	31.7 [27.7, 36.6]
SenseNova-Vision	67.3	45.5 [40.6, 50.5]	29.7 [24.8, 33.7]	15.8 [12.9, 20.8]	44.6 [38.6, 49.5]	33.7 [29.7, 36.6]	11.9 [9.9, 13.9]†	45.5 [39.6, 51.5]	42.6 [37.6, 47.5]	37.6 [33.7, 42.6]
GPT-5.6-Sol	17.8	12.9 [9.9, 16.8]	6.9 [4.0, 8.9]	2.0 [1.0, 4.0]†	9.9 [7.9, 12.9]	5.9 [5.0, 8.9]	3.0 [2.0, 5.0]	11.9 [8.9, 14.9]	11.9 [8.9, 13.9]	13.9 [10.9, 15.8]
Claude-Opus-5	18.8	21.8 [17.8, 27.7]	9.9 [7.9, 13.9]	6.9 [5.0, 9.9]	12.9 [9.9, 16.8]	7.9 [6.9, 9.9]	5.9 [5.0, 8.9]	15.8 [12.9, 20.8]	15.8 [12.9, 19.8]	17.8 [13.9, 21.8]
Gemini-3.8-Flash	37.6	21.8 [17.8, 27.7]	15.8 [12.9, 19.8]	6.9 [5.0, 9.9]	14.9 [11.9, 18.8]	11.9 [9.9, 14.9]	4.0 [3.0, 5.0]†	16.8 [12.9, 20.8]	17.8 [14.9, 21.8]	16.8 [12.9, 19.8]
Qwen-3.8	32.7	24.8 [20.8, 30.7]	13.9 [10.9, 17.8]	6.9 [5.0, 8.9]†	19.8 [15.8, 22.8]	15.8 [12.9, 18.8]†	5.9 [5.0, 7.9]†	21.8 [17.8, 25.7]	15.8 [12.9, 18.8]	17.8 [13.9, 20.8]

Table 14: AP@.75 with two-sided 90% BCa intervals (†: percentile interval) under Haze/Fog for co-occurring and single-condition images; see Section 3.2.

Model	Co-occurring$_{218}$				Haze/Fog only$_{221}$			
	All$_{218}$	Easy$_{41}$	Med.$_{84}$	Hard$_{93}$	All$_{221}$	Easy$_{100}$	Med.$_{77}$	Hard$_{44}$
ViTPose++-H	46.2 [40.1, 50.9]	72.0 [63.5, 79.8]	45.4 [37.1, 53.4]	33.7 [24.3, 41.5]	72.0 [68.2, 76.3]	88.2 [83.8, 91.3]	65.2 [57.9, 71.8]	39.7 [29.6, 47.9]
ProbPose	29.5 [24.9, 34.5]	58.4 [47.6, 69.2]	30.0 [22.5, 38.0]	14.7 [9.7, 20.2]	58.2 [53.9, 62.9]	75.0 [69.1, 79.6]	49.6 [41.4, 57.6]	28.2 [19.5, 36.2]
CIGPose-L	34.9 [29.9, 39.6]	65.3 [54.3, 74.6]	37.6 [30.0, 45.7]	17.0 [12.1, 22.4]	64.7 [60.6, 69.5]	82.1 [77.9, 86.8]	57.3 [49.5, 64.9]	33.1 [23.8, 41.3]
GroupPose	29.7 [25.2, 34.3]	58.1 [47.2, 68.3]	27.5 [21.4, 35.4]	17.8 [12.6, 23.9]	54.7 [50.3, 59.0]	68.4 [62.1, 74.0]	49.0 [42.1, 55.4]	30.9 [21.8, 40.0]
YOLO26x-pose	24.5 [20.1, 29.2]	50.8 [38.6, 62.0]	24.3 [17.5, 32.5]	12.7 [8.3, 18.0]	47.9 [43.4, 52.3]	62.0 [55.4, 67.9]	43.8 [37.1, 50.4]	23.5 [15.1, 32.3]
Sapiens2	49.4 [43.7, 54.9]	75.6 [65.9, 82.9]	50.4 [42.0, 58.2]	33.7 [23.3, 41.3]	74.4 [71.0, 78.7]	86.6 [81.7, 90.1]	68.9 [61.8, 74.6]	50.2 [39.0, 60.1]
GKDT-H	37.4 [31.4, 42.2]	64.7 [53.9, 73.2]	35.8 [27.9, 43.8]	26.9 [18.4, 34.4]	56.0 [51.4, 59.8]	68.4 [63.3, 74.1]	53.0 [44.9, 60.2]	32.4 [23.9, 41.6]
LocLLM	43.2 [37.7, 48.0]	66.9 [54.4, 76.7]	41.8 [34.4, 48.8]	32.0 [23.8, 38.9]	71.7 [67.9, 75.9]	84.5 [78.6, 88.8]	66.1 [58.4, 72.0]	45.5 [35.1, 54.1]
Rex-Omni	15.8 [10.9, 17.8]	22.8 [14.9, 30.7]	20.8 [15.8, 25.7]	7.9 [5.0, 12.9]	29.7 [26.7, 34.7]	35.6 [30.7, 40.6]	28.7 [22.8, 34.7]	13.9 [8.9, 19.8]
SenseNova-Vision	20.8 [17.8, 24.8]	32.7 [25.7, 39.6]	22.8 [17.8, 28.7]	12.9 [8.9, 17.8]	39.6 [35.6, 42.6]	47.5 [41.6, 52.5]	37.6 [31.7, 43.6]	22.8 [16.8, 28.7]
GPT-5.6-Sol	5.0 [4.0, 6.9]	8.9 [5.9, 14.9]	5.9 [4.0, 9.9]	1.0 [0.0, 3.0]	9.9 [7.9, 11.9]†	12.9 [8.9, 15.8]	8.9 [5.9, 12.9]	4.0 [1.0, 5.9]
Claude-Opus-5	7.9 [5.0, 9.9]	15.8 [9.9, 22.8]	8.9 [5.9, 14.9]	4.0 [2.0, 6.9]	16.8 [12.9, 19.8]	22.8 [17.8, 28.7]	11.9 [7.9, 14.9]	11.9 [6.9, 18.8]
Gemini-3.8-Flash	7.9 [5.9, 10.9]	11.9 [6.9, 17.8]	9.9 [6.9, 14.9]	3.0 [1.0, 5.9]	19.8 [15.8, 22.8]	23.8 [19.8, 29.7]	19.8 [14.9, 24.8]	11.9 [6.9, 17.8]
Qwen-3.8	9.9 [6.9, 11.9]	17.8 [11.9, 25.7]	9.9 [5.9, 13.9]	6.9 [5.0, 10.9]	19.8 [16.8, 23.8]	25.7 [21.8, 32.7]	18.8 [13.9, 24.8]	7.9 [4.0, 10.9]