They Can See It but Not Say It: Iterative Self Knowledge Re-expression in Visual Reasoning Relative Pose Identification

Mengyu Wang **Ken Deng** **Yftah Ziser**

Shay B. Cohen **Tiejun Ma**

University of Edinburgh University of Oxford Nvidia Research

Abstract

The task of visual reasoning relative pose identification (VRRPI) has exposed critical weaknesses in the multi-view 3D spatial reasoning capabilities of modern vision-language models (VLMs). We present a systematic analysis of VLMs' behavior in VRRPI, uncovering stable output biases where VLMs disproportionately predict specific motion patterns regardless of visual evidence. To mitigate these inherent biases, we propose self knowledge re-expression with iterative debiasing (SKR-ID), an annotation-free pipeline that adapts VLMs for VRRPI using only randomly extracted unannotated image pairs. SKR-ID transitions VLMs' output mechanism from generic next-token prediction to task-optimized classification heads. The pipeline proceeds in two stages: (1) a cold-start initialization using debiasing object-centric prompts to anchor motion inference in 2D visual evidence; and (2) a recursive logit-ranking mechanism that enforces statistical parity by using the median logit value as a dynamic threshold. Experimental evaluation across frontier VLMs on standard benchmarks demonstrates that SKR-ID significantly mitigates systematic biases. Our method improves open-source VLMs' Macro-F1 scores from 35.0–45.4% to 53.7–58.0%, without using any annotated data. These results show that SKR-ID can effectively improve 3D reasoning performance from a severely collapsed baseline. The code is publicly available at this repository.

1 Introduction

VLMs have demonstrated exceptional ability in 2D perception and high-level semantic reasoning, excelling in tasks ranging from object detection [16, 42] to complex image understanding [39]. However, these models are primarily optimized for 2D information processing, often lacking a fundamental grasp of 3D spatial geometry. This deficiency has been demonstrated in the task of visual reasoning relative pose identification (VRRPI), which requires inferring the relative view change between two camera viewpoints [8].

Despite the geometric clarity of VRRPI, recent benchmarks reveal that even frontier VLMs struggle to generalize beyond shallow 2D reasoning [9, 26, 32, 7]. Advanced models often perform even below random chance, significantly lagging behind human performance. To investigate the cause of this failure, we conducted a systematic analysis of VLM behavior across various architectures. We uncover a robust and stable output bias: models exhibit a systematic preference for specific motion

Correspondence: 1306206684wmy@gmail.com

These authors jointly supervised this work

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

patterns or selections regardless of the actual visual parallax or prompt configuration. Our analysis reveals that these biases reflect frontier VLMs' inherent multi-view 3D spatial reasoning weaknesses, rather than being a surface-level artifact of instruction following.

However, given the inherent correlation between 3D geometric transformations and the pixel-level shifts involved in 2D object detection and tracking, tasks where modern VLMs already demonstrate good performance [3, 28, 22], we argue that these models are not inherently ignorant of 3D geometry. Instead, we believe some 3D spatial reasoning potential is latent within their parameters but remains underdeveloped because standard vision-encoder training often ignores 3D spatial information. This suggests that the performance gap is not only a lack of knowledge, but also a failure to correctly extract and express existing knowledge.

To use this latent potential, we build upon the framework of self-knowledge re-expression (SKR) [30]. SKR posits that many deficiencies in LLMs are *expression bottlenecks*, where the model possesses the requisite parametric knowledge but is constrained by the next-token prediction (NTP) output paradigm. By *re-expressing* latent knowledge through task-optimized output heads, we can activate the model's internal geometric reasoning without requiring human annotations or external solvers.

Motivated by this perspective, we propose the self knowledge re-expression with iterative debiasing (SKR-ID) method to adapt VLMs for VRRPI in an annotation-free setting. Our method relies solely on randomly extracted image pairs from unlabeled videos and proceeds in two stages. First, we use a debiasing prompt that guides the model to follow a reasoning chain, prioritizing 2D object-centric motion before inferring 3D camera-centric pose. Using this prompt for self-annotation, we implement the SKR process to transition the VLM from token generation to a task-specific classification head.

Second, we further address the model's inherent statistical bias through a logit-ranking mechanism. After the first round of SKR, the model provides task-specific prediction logits. We use these logits to re-calibrate the pseudo-labeling process. Instead of applying a fixed threshold, which would preserve the model's skewed distribution, we rank all self-annotations by their logits and set a dynamic threshold at the median. This ensures a balanced 50/50 label distribution for subsequent rounds of training, forcing the model to learn from balanced data and thereby mitigating bias in its spatial reasoning. This process of using SKR-adapted model to activate the next SKR iteration can be run recursively to further improve performance.

Our experimental results across three different frontier open-source VLMs, LLaVA-OneVision-7B, Qwen3-VL-8B, and VST-RL-7B, demonstrate that SKR-ID substantially reduces output collapse. The Macro-F1 scores are improved by 8.3–23.0 percentage points. Our detailed analysis also reveals that models adapted via SKR-ID successfully mitigate large-scale output biases.

In summary, our contributions are as follows:

1. Analysis of Systematic Spatial Reasoning Biases in VLMs. We provide a rigorous analysis of VLM behavior on VRRPI and identify inherent stable output biases that persist across model architectures and prompting strategies.

2. The SKR-ID Method: We propose an annotation-free method that enables VLMs to improve their 3D relative pose identification by re-expressing their own latent knowledge through task-specific heads and iterative logit-based debiasing.

3. Performance Gains: We demonstrate that SKR-ID consistently improves VRRPI performance across multiple frontier open-source VLMs, substantially mitigating output imbalance by increasing Macro-F1 from 35.0–45.4% to 53.7–58.0%.

2 Related Work

Spatial grounding in VLMs Spatial grounding, binding language to locations, relations, and viewpoints, is a key stress-test for VLMs. Many benchmarks reveal systematic failures that persist across model scales and architectures [24, 13, 41]. Models that perform well on standard recognition tasks often drop to near-chance performance on cross-image correspondence and viewpoint-shifted localization [18, 34, 37, 8]. Recent studies indicate that VLM spatial encodings lack invariance to viewpoint changes [15], leading to performance drops when reasoning moves beyond the ego-centric frame [8]. Existing solutions inject geometric priors or view-invariant objectives during pre-training [12, 5, 38, 14]. Our work identifies that VRRPI failures manifest as stable output biases.

3D spatial reasoning and relative camera pose Recovering relative camera pose from image pairs [11] is a classical problem in computer vision. Traditional methods rely on establishing visual correspondences (e.g., keypoints or dense matches) and recover motion using epipolar geometry or known camera intrinsics [19, 27, 23]. Recent work reformulates this problem in VLMs, where relative motion is expressed as discrete language descriptions [8, 37, 34, 36]. Unlike geometric methods, VLMs rely solely on image inputs and implicit reasoning, without explicit correspondence modeling or camera calibration. Existing approaches augment VLMs with depth estimation or 3D-aware representations [12, 5, 21, 35]. Large-scale supervised datasets have also been used to inject 3D spatial knowledge [4, 33]. However, a gap persists between semantic understanding and geometric reasoning. While VLMs can often identify shared objects across views, they still struggle to infer consistent view change [8, 40]. We show that these limitations are driven by response biases that persist even after fine-tuning, motivating our annotation-free debiasing approach.

Self-training and knowledge re-expression Self-training uses model predictions as pseudo-labels for adaptation [1, 31, 30]. Self-knowledge re-expression (SKR) further takes a complementary approach by replacing the autoregressive generation head with a task-specific expression head [30]. But they remain vulnerable to bias amplification when initial predictions are skewed. Collaborative filtering and contrastive objectives have been proposed to suppress imbalanced labels in annotation-free settings [29, 31, 1]. We extend SKR with an object-centric cold-start and a median-logit-ranking mechanism to enforce statistical parity, preventing distributional collapse during iterative refinement.

3 Task Formulation

We use the standard VRRPI-Bench dataset to develop experiments. It casts relative pose reasoning as a binary-choice visual question answering problem. Each example contains two views of the same scene, denoted by (I_1, I_2), and asks the model to identify the camera motion that transforms the first view into the second. The benchmark discretizes this continuous transformation into two mutually exclusive natural-language options, "*Move left while yawing right*" and "*Move right while yawing left*". Although this task seems easy, studies have demonstrated that even frontier VLMs perform very poorly, with most VLMs performing near or even below random level [8].

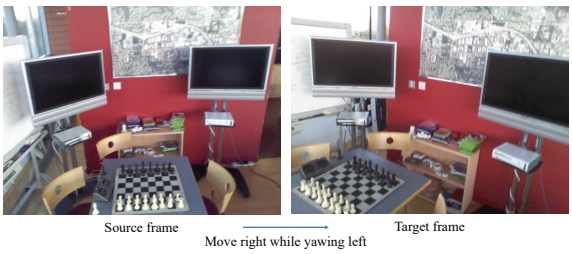

Figure 1: Example instance of the VRRPI task.

Formally, using the annotations $\langle$*left*$\rangle$ and $\langle$*right*$\rangle$ to represent the two options "*Move left ...*" and "*Move right ...*" respectively, each instance is written as $x_i = (I_1^i, I_2^i, p)$, where $p = \text{prompt}(\langle \textit{left} \rangle, \langle \textit{right} \rangle)$ is the prompt template that includes the task instructions and the two options. Given a VLM M, the model is asked to output $\hat{y}_i \in \{\langle \textit{left} \rangle, \langle \textit{right} \rangle\}$ based on the input x_i to identify the change in camera motion.

4 Bias Analysis of VLMs on VRRPI

In this section, we conduct carefully designed evaluations on VRRPI-Bench to demonstrate that frontier VLMs do not fail this task randomly. Instead, they collapse onto a small set of motion descriptions, exhibiting a stable output bias. Specific options are selected far more frequently than their symmetric counterparts, even when the input order is permuted. This phenomenon reveals that VLM failures in 3D spatial reasoning are driven by systematic response pathologies rather than stochastic noise, and these biases remain consistent across different models.

4.1 Evaluation Design

To characterize these biases, we evaluate each model under three controlled perturbations: (1) Base Inputs (I_1, I_2, p), (2) Option-swap Inputs (I_1, I_2, p_r), where $p_r = \text{prompt}(\langle \textit{right} \rangle, \langle \textit{left} \rangle)$ reverses the order of choices in p, and (3) Image-swap Inputs (I_2, I_1, p), which exchanges the source and target

Table 1: Prediction ratios under controlled perturbations. Each cell reports the ratio of predicted labels $\langle \textit{left} \rangle$ vs. $\langle \textit{right} \rangle$. In all experimental settings, the ground-truth distribution is balanced (50.0% / 50.0%). Base, Swap Opt., and Swap Img. denote the base inputs, the option-swap inputs, and the image-swap inputs settings respectively.

	Base	Swap Opt.	Swap Img.
$\hat{y} \in \{\langle \textit{left} \rangle, \langle \textit{right} \rangle\}$			
LLaVA-OV	0.0% / 100.0%	100.0% / 0.0%	0.0% / 100.0%
Qwen3-VL	10.6% / 89.4%	30.8% / 69.2%	13.9% / 86.1%
VST-RL	13.8% / 86.1%	13.8% / 86.1%	18.4% / 81.5%
$\hat{y} \in \{\langle \textit{left} \rangle, \langle \textit{right} \rangle, \langle \textit{neither} \rangle\}$			
LLaVA-OV	0.0% / 100.0%	95.6% / 0.6%	0.0% / 100.0%
Qwen3-VL	11.9% / 85.4%	26.5% / 70.4%	15.9% / 79.7%
VST-RL	6.6% / 92.7%	24.4% / 75.2%	8.8% / 91.1%

frames. These settings isolate visual reasoning from the input configuration. In a well-calibrated model, swapping the option order should preserve the semantic answer, while swapping the image order should logically invert the answer because the underlying camera motion is reversed. Any persistent deviation from this behavior indicates that the model is relying on an indexing prior (a preference for a specific option position) or a motion prior (a preference for a specific movement type) rather than reasoning over the visual content.

Additionally, to verify whether the binary-choice constraint itself induces these failures, we introduce an extra option $\langle \textit{neither} \rangle :=$ "*Neither of them*" and repeat the perturbation experiments. This serves as a control to ensure the model is not merely guessing due to forced choice.

4.2 Problematic Outputs from VLMs

We evaluate three frontier VLMs representing diverse architectures: LLaVA-OneVision, Qwen3-VL, and VST-RL (see Section 6.2 for detailed descriptions).

Table 1 reports the prediction ratios for $\hat{y} = \langle \textit{left} \rangle$ and $\hat{y} = \langle \textit{right} \rangle$ under the three settings. Because the benchmark is balanced by construction, a calibrated model should maintain an approximately 50.0% / 50.0% prediction distribution.

The results reveal a catastrophic breakdown in reasoning. LLaVA-OV exhibits a profound indexing bias, showing a tendency to always select the second option regardless of the option content or the input image order. This preference persists even after the introduction of a third option, confirming a complete collapse on the task.

Qwen3-VL and VST-RL exhibit answer biases, consistently preferring selection $\langle \textit{right} \rangle$ across all configurations. The addition of the $\langle \textit{neither} \rangle$ option does not solve this behavior, proving that these models are genuinely biased toward specific spatial outcomes rather than being constrained by the selection structure.

These problematic outputs may naturally lead to the assumption that frontier VLMs possess fundamental knowledge deficits in 3D spatial reasoning. However, given the inherent correlation between 3D geometric transformations and the 2D visual cues, which VLMs can understand correctly, it is unlikely that these models are entirely ignorant of 3D spatial principles. Therefore, we posit that a primary cause of these failures is an expression bottleneck, where the model fails to correctly map its internal representations to the task's required outputs. This insight motivates our use of SKR-ID, an annotation-free method that relies on the models' own parametric knowledge to bridge this gap.

5 Resolving the Bias: SKR with Iterative Debiasing

We introduce SKR-ID, an annotation-free pipeline designed to transition VLMs from generic language generators into specialized view change reasoners. Our method extends SKR by formalizing

the adaptation process as a recursive optimization loop that progressively decouples latent spatial knowledge from inherent model biases.

5.1 Preliminaries

The self knowledge re-expression (SKR) method is a fully local process that enables task adaptation using only unannotated data $\mathcal{D}_x$ [30]. The core premise of SKR is that LLMs/VLMs often possess much required parametric knowledge to solve a specific task T, but are constrained by the next-token prediction output paradigm. SKR decouples this latent knowledge from its token expression by transitioning the model from a generic mechanism E_{ntp} to a task-specific expression E_T. The process executes in two steps:

In the first step, **self-annotation**, the model uses its native generative mechanism, E_{ntp}, to perform "knowledge extraction". Given a task T and unannotated inputs $\mathcal{D}_x$, the model generates textual outputs to complete the task and the outputs are processed into task-specific pseudo-labels $\hat{\mathcal{D}}_y$, such as numerical values or classification labels:

$$\mathcal{D}_x \xrightarrow[\text{Task Query}]{M_{\theta,\,E_{ntp}}} \hat{\mathcal{D}}_y, \tag{1}$$

where $M_{\theta,E_{\text{ntp}}}$ represents the base model, with parameters θ and expression E_{ntp}, and $\hat{\mathcal{D}}_y$ represents the generated pseudo labels.

In the second step **knowledge re-expression**, the model is fine-tuned to map its internal hidden states directly to a task-adapted expression E_T, such as a classification head or regression layer. This bypasses the inefficiencies and errors of sequential token generation. This step is formalized as:

$$M_{\theta,E_{ntp}} \xrightarrow[(\mathcal{D}_x,\hat{\mathcal{D}}_y),\text{ Loss of } E_T]{\text{Fine-tune}} M_{(\theta,\Delta\theta_T),E_T}, \tag{2}$$

where θ denotes the base parameters and $\Delta\theta_T$ represents the task-specific adapters and head.

5.2 SKR-ID Stage 1: Cold-start via Debiasing Prompts

Directly applying SKR to VRRPI is challenging because some base model's self-annotations $\hat{\mathcal{D}}_y$ are heavily biased, as discussed in Section 4. However, we observe that while VLMs struggle with 3D pose reasoning, they excel at 2D object grounding and tracking.

We therefore design an object-centric reasoning prompt p_{debias} to anchor the self-annotation process in visual evidence. When the base model's outputs are highly biased, instead of directly querying the view change, the prompt guides the model through a causal chain: (i) identify a shared reference object in both frames, (ii) judge the object's movement, and (iii) infer the relative view change. This approach uses the model's 2D proficiency to generate more balanced labels.

It is important to note that this prompting strategy is not a direct solution to the task. Due to the models' inherent geometric reasoning weakness, performance under this setup is still not good enough, which will be discussed in the ablation study Section 7.2. Additionally, since SKR-ID uses only randomly sampled unannotated data for training, the object-centric prompt provides noisy, directionally informative pseudo-labels by grounding the initial prediction in the apparent motion of shared objects. However, its value lies in providing a "cold-start" mechanism: it guides the VLM to generate a significantly more balanced distribution of initial pseudo-labels than direct querying would allow. This relative parity is essential to activate the SKR adaptation without immediately inheriting a collapsed output distribution.

Once these labels, represented by $\hat{\mathcal{D}}_y^0$, are generated, we implement SKR by replacing the model's final language modeling layer, which typically outputs a probability distribution over the entire vocabulary, with a task-specific binary classification head E_{VRRPI}, whose outputs correspond directly to the two VRRPI motion classes. We train this head, together with the LoRA adapters, using the pseudo-labels and binary classification objective. Following the SKR framework, the whole process can be formalized as:

$$\mathcal{D}_x \xrightarrow[p_{debias}]{M_{\theta,\,E_{ntp}}} \hat{\mathcal{D}}_y^0, \quad M_{\theta,E_{ntp}} \xrightarrow[(\mathcal{D}_x,\hat{\mathcal{D}}_y^0),\text{CE Loss}]{\text{Fine-tune}} M^0_{(\theta,\Delta\theta^0_{VRRPI}),E_{VRRPI}}. \tag{3}$$

The SKR model $M^0_{(\theta,\Delta\theta^0_{VRRPI}),E_{VRRPI}}$ now functions as a standard classifier, providing task-specific logits that represent the probability of the candidate view changes. In the following section, we represent the SKR-adapted models as $M^0_{(\theta,\Delta\theta^0_{VRRPI})}$, omitting the explicit mechanism E_{VRRPI} as the architecture remains identical.

5.3 SKR-ID Stage 2: Iterative SKR with Logit-based Debiasing

The initial round of SKR yields the model $M^0_{(\theta,\Delta\theta^0_{VRRPI})}$. It performs much better than the base model on VRRPI. Transitioning to a task-specific head enables M^0 to output continuous class logit distributions, which is far more informative than generated tokens. Both benefits suggest that M^0 has the potential to generate higher-quality self-annotations for the unannotated dataset $\mathcal{D}_x$, enabling a recursive process where each SKR model supports the next turn's SKR for better performance.

To implement this iteration, we define the turn index as i ($i > 0$), where Stage 1 initializes the process with M^0. Using the class logits $L^i_x = \{\ell^i_1, \ell^i_2, \ldots, \ell^i_n\}$ generated by the previous fine-tuned SKR model $M^{i-1}_{(\theta,\Delta\theta^{i-1}_{VRRPI})}$, we implement a median logit-ranking strategy for self-annotation. We operate under a viewpoint symmetry prior: in a sufficiently large set of random changed viewpoints, the distribution of opposing directions (left vs. right) should be approximately uniform. Therefore, instead of using an absolute threshold 0.5 to decide labels, we can use the logits to force the self-annotations follow the balanced distribution. We rank the logits L^i_x for the entire set $\mathcal{D}_x$ and define the threshold T^i_{med} at the median value: $T^i_{\text{med}} = \text{median}(L^i_x)$. We then assign refined pseudo-labels $\hat{\mathcal{D}}^i_y$ by partitioning the data at this threshold. This ensures an exact 50/50 label distribution, effectively forcing the model to learn under a balanced supervision. Each iteration i is formalized as:

$$\mathcal{D}_x \xrightarrow{M^{i-1}_{(\theta,\Delta\theta^{i-1}_{VRRPI})}} L^i_x \xrightarrow{T^i_{\text{med}}} \hat{\mathcal{D}}^i_y, \quad M_\theta \xrightarrow[(\mathcal{D}_x,\hat{\mathcal{D}}^i_y),\text{CE Loss}]{\text{Fine-tune}} M^i_{(\theta,\Delta\theta^i_{VRRPI})}. \tag{4}$$

The iterated fine-tuned model $M^i_{(\theta,\Delta\theta^i_{VRRPI})}$ demonstrates further improved performance and enhanced spatial reasoning. The process terminates upon reaching a maximum iteration count or when the performance on the validation set (which is also split from the unannotated data) fails to improve for a set number of iterations. The complete pipeline is summarized in Algorithm 1.

Algorithm 1 SKR-ID

Input: Unlabeled image pairs $\mathcal{D}_x$, base VLM M_θ with parameters θ, debiasing prompt p_{debias}, max iterations $I_{\max}$
Output: Optimized view change reasoner $M^i_{(\theta,\Delta\theta^i_{VRRPI})}$
Stage 1: Cold-start Initialization ($i = 0$)
$\hat{\mathcal{D}}^0_y \leftarrow$ Query $M_{\theta,E_{\text{ntp}}}$ on $\mathcal{D}_x$ using p_{debias}
$M^0_{(\theta,\Delta\theta^0_{VRRPI})} \leftarrow$ Fine-tune M_θ on $(\mathcal{D}_x, \hat{\mathcal{D}}^0_y)$ via cross-entropy Loss
Stage 2: Iterative Logit-Ranking Refinement
for $i = 1$ **to** $I_{\max}$ **do**
    Extract logits $L^i_x \leftarrow M^{i-1}_{(\theta,\Delta\theta^{i-1}_{VRRPI})}(\mathcal{D}_x)$
    Compute threshold $T^i_{\text{med}} \leftarrow \text{median}(L^i_x)$
    Update pseudo-labels $\hat{\mathcal{D}}^i_y$: $\hat{y}^i_j = 1$ if $\ell^i_j \geqslant T^i_{\text{med}}$ else 0
    $M^i_{(\theta,\Delta\theta^i_{VRRPI})} \leftarrow$ Fine-tune M_θ on $(\mathcal{D}_x, \hat{\mathcal{D}}^i_y)$ via cross-entropy Loss
    if validation accuracy stagnates **then break**
end for
return $M^i_{(\theta,\Delta\theta^i_{VRRPI})}$

6 Experimental Settings

6.1 Dataset and Evaluation

Dataset We evaluate our methods on VRRPI-Bench [8], broken down by the 7 Scenes [10, 25] and ScanNet [6] subsets. Image pairs are grouped by approximate viewpoint separation into four difficulty levels: 15°, 30°, 45°, and 60°. The ground truth labels follow the binary classification objective detailed in Section 3. For SKR-ID, we use only random unlabeled image pairs sampled from ScanNet (no overlaps with VRRPI-Bench). The VRRPI-Bench dataset is used only for evaluation. This setup strictly follows an annotation-free protocol while evaluating on a standard labeled benchmark.

Metric We report Macro-F1 as the evaluation metric. Accuracy can be misleading when predictions are imbalanced. A model with heavily biased outputs can obtain seemingly high accuracy by outputting meaningless predictions. Macro-F1 treats classes equally by computing the F1-score for each class and averaging them: $\text{Macro-F1} = \frac{1}{2}\left(\text{F1}_{\langle \textit{left} \rangle} + \text{F1}_{\langle \textit{right} \rangle}\right)$. For balanced binary labels, random guessing yields an expected Macro-F1 near 0.5. Heavily biased predictions result in scores below 0.5. Macro-F1 therefore serves as a better metric reflecting both accuracy and biases. All SKR-ID results are averaged over three runs, and the improvements are statistically significant under a t-test at the 0.5% level.

6.2 Models and Baselines

To evaluate the effectiveness of SKR-ID, we select three frontier VLMs as backbones:

LLaVA-OneVision (LLaVA-Onevision-Qwen2-7B-ov-hf) [17]: A unified framework designed for image and video understanding, recognized for its superior performance among open-source models.

Qwen3-VL (Qwen3-VL-8B-Instruct) [2]: A recently released general-purpose VLM with enhanced perceptual capabilities, representing the state-of-the-art in open-source multimodal architectures.

VST-RL (VST-7B-RL) [33]: A specialized model fine-tuned using the Visual Spatial Tuning framework, which specifically targets human-like spatial perception and reasoning.

These backbones represent a diverse spectrum of current multimodal research. By adapting these models via SKR-ID and comparing with them, we aim to demonstrate that even specialized spatial models struggle in the task of VRRPI and can benefit from our SKR-ID method.

6.3 Implementation Details

For model structure, we replace the vocabulary layer with a lightweight classification head attached to the last token's final hidden state. This head consists of a LayerNorm and a two-layer MLP with a 1,024-dimension bottleneck and GELU activation. The final output is a scalar probability generated by a Sigmoid function, mapping internal representations directly to the target labels.

We apply parameter-efficient fine-tuning via LoRA on all LLMs' attention projection layers (W_q, W_k, W_v, W_o) with hyperparameters $r = 4$, $\alpha = 8$, and a dropout rate of 0.1. Optimization is conducted using the AdamW optimizer (learning rate of 10^{-5}) and a binary cross-entropy (BCE) loss objective.

We implement a recursive adaptation process with a maximum of 5 iterations. For the initialization stage ($i = 0$), we use the object-centric debiasing prompt for LLaVA-OneVision and Qwen3-VL to address severe inherent biases, while VST-RL uses direct querying. Each iteration involves 10 epochs of training with a batch size of 8. Validation occurs every 200 steps to preserve optimal checkpoints.

All experiments are conducted using 4 NVIDIA A100 GPUs. The SKR training process uses 2 GPUs for LLaVA-OneVision and 4 GPUs for the other two models. Inference is performed on a single GPU for all models. Experimental prompts are detailed in Appendix A.

7 Results and Discussion

We introduce detailed experimental results and analysis in this section. Section 7.1 presents the performance of the backbone models, representing state-of-the-art open-sourced VLMs, compared to

Table 2: Results (Macro-F1 in %) on VRRPI-Bench.

Model	7 Scenes					ScanNet					All
	15°	30°	45°	60°	Avg.	15°	30°	45°	60°	Avg.	
LLaVA-OV	37.6	33.9	44.7	35.8	37.6	38.0	36.6	31.0	33.9	34.8	35.0
↳SKR-ID	54.1	57.9	61.0	61.6	58.6	61.9	60.0	46.1	63.0	57.7	**58.0**
Qwen3-VL	30.6	41.1	36.5	53.9	38.9	37.3	40.3	39.2	42.6	39.9	39.5
↳SKR-ID	42.7	52.1	58.0	64.0	54.2	36.0	47.3	65.8	67.9	54.2	**53.8**
VST-RL	43.6	47.0	41.8	59.3	46.7	37.2	35.5	46.6	57.6	44.2	45.4
↳SKR-ID	47.7	52.4	53.2	61.5	53.7	43.5	52.7	52.6	69.4	54.5	**53.7**

our SKR-ID method. Section 7.2 presents the performance of different stages of our SKR-ID method, analyzing the ablated contribution of our designs. Section 7.3 demonstrates the sensitivity of SKR-ID to different class-distribution priors, showing the effectiveness of median-thresholding. Section 7.4 uses LLaVA-OneVision as an example to clearly demonstrate the change in confusion matrices across our pipeline, highlighting the systematic debiasing effect.

7.1 VRRPI Results: Comparison with Backbones

Table 2 summarizes the performance of frontier VLMs compared to those adapted via our SKR-ID method on the VRRPI-Bench. Appendix B reports the complementary evaluation metrics for these results, including accuracy, balanced accuracy, per-class F1, prediction ratios, and AUROC. The results demonstrate a consistent and significant performance improvement across all evaluated backbones.

These base VLMs, due to their highly biased outputs on the VRRPI task, often fall below even the random-guess baseline (50.0%). Following SKR-ID adaptation, all three models exhibit a Macro-F1 improvement of at least 8.3 percentage points. LLaVA-OV achieves the most dramatic recovery, with its average Macro-F1 score improving by 23.0 absolute points. Even for the specialized VST-RL model, which is pre-trained for spatial reasoning and initially shows a more balanced distribution, SKR-ID provides a clear boost from 45.4% to 53.7%. These substantial gains suggest that by transitioning from token prediction to a task-specific head and enforcing label balance, the model can successfully decouple its latent representations from its biased pre-training priors.

Overall, the consistent gains across different backbones validate SKR-ID as a robust, annotation-free method for adapting VLMs to the VRRPI tasks. The recovery of heavily biased models to much better performance confirms that current VLM spatial reasoning failures are related to the "expression bottlenecks" rather than a total absence of multi-view 3D spatial reasoning knowledge.

7.2 Ablation Study: Contributions of Individual SKR-ID Components

Figure 2 illustrates the performance at each stage of the SKR-ID pipeline across the three backbone models. The ablation demonstrates that the transition from a base state to the final improved model is driven by the integration of object-centric prompting, knowledge re-expression, and recursive refinement. We also add a comparison with the standard supervised fine-tuning (SFT), which uses the self-annotations generated by the base model (the same as stage 1 initialization), but conducts SFT to fine-tune the base model rather than SKR. Appendix C provides the extended component ablation and additional baselines on LLaVA-OneVision.

The transition from base model to using special prompt (Stage 1 initialization) highlights the effectiveness of the object-centric prompt in debiasing VLMs' 3D spatial understanding. The performance of all three backbones improved after applying the debiasing reasoning. Models with the strongest initial bias benefited the most, LLaVA-OneVision jumped from 35.0% to 49.1%, demonstrating the prompt's critical role in anchoring 3D view change inference in 2D visual evidence.

The subsequent SKR method provides consistent improvements across all models. The first round of SKR, which re-expresses model outputs to classification information, further improves results. Importantly, the SKR models' inputs use the same direct query prompt as the base model, rather than the debiasing prompt. This confirms that the SKR mechanism can effectively unearth and formalize the model's latent spatial abilities that were originally only accessible through specialized prompting.

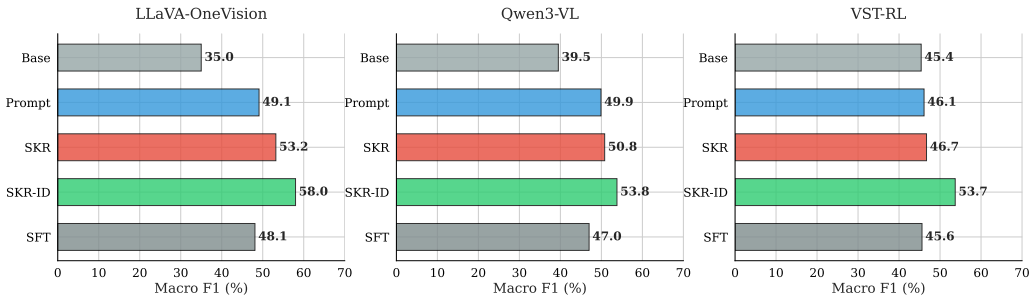

Figure 2: Comparison of VRRPI performance across adaptation phases, from base models to the final SKR-ID models. SFT denotes supervised fine-tuning using self-annotations from the base model.

Finally, SKR-ID pushes all backbones to their peak performance (58.0% for LLaVA-OneVision, 53.8% for Qwen3-VL, and 53.7% for VST-RL). By using predicted logits to enforce statistical parity through median thresholding, SKR-ID successfully unearths the models' potential for VRRPI.

Additionally, SFT results are clearly weaker than those of SKR-ID. While SFT outperforms the base model by partially mitigating biases via the debiasing prompt, it significantly underperforms the final SKR-ID. SFT remains constrained by the NTP paradigm and cannot provide task-specific information like class logits to support granular logit-ranking debiasing.

7.3 Logit-Based Debiasing Analysis: Baselines and Prior Sensitivity

Stage 2 of SKR-ID uses median-based thresholding to mitigate output bias, based on the balanced class prior expected for randomly sampled pairs in our symmetric binary camera-motion task. To assess this design, we compare median thresholding across alternative base-model score constructions and examine its sensitivity to different class-distribution priors.

We first compare SKR-ID with median thresholding applied directly to scores from frozen VLMs, using three common score construction methods: (1) prompting the model to generate class probabilities, (2) aggregating token scores for class-relevant keywords, and (3) computing length-normalized sequence log-probabilities for the candidate answers. These controls use no task-specific classification head, fine-tuning, or iterative self-training. Table 3 shows that median thresholding of these scores does not reproduce the performance of SKR-ID. Its effect varies by score construction and backbone, while SKR-ID performs best for all three models. This comparison indicates that changing the output balance alone is insufficient to obtain the gains of the full method.

We next examine the influence of the class-distribution prior. Because SKR-ID uses no annotated training data, we cannot construct a training set with a ground-truth class distribution. Therefore, we compare two approximate imbalance settings to assess the sensitivity of our prior: (1) constructing a 70/30 training subset using pseudo-labels from the object-centric prompt, and (2) replacing the median threshold with a 0.7-quantile threshold. The first setting does not guarantee a true 70/30 ground-truth distribution since the pseudo-labels may be noisy and some image pairs may be ambiguous. As shown in Table 3, the pseudo-label-based 70/30 setting moderately reduces Macro-F1 relative to standard SKR-ID, while the 0.7-quantile threshold causes a larger decrease. The median threshold performs best in these experiments, supporting the reasonableness of our balanced prior and the effectiveness of the method designed around this assumption.

Table 3: Macro-F1 (%) for median-threshold controls and class-distribution prior sensitivity on VRRPI. The 70/30 subset is constructed using noisy pseudo-labels from the object-centric prompt.

Model	Base	Prompt prob.	Keyword token	Sequence log-prob.	70/30 subset	0.7-quantile	Median
LLaVA-OV	35.0	32.3	34.3	43.4	56.7	47.2	**58.0**
Qwen3-VL	39.5	46.3	36.7	37.8	51.9	46.3	**53.8**
VST-RL	45.4	48.9	43.4	34.3	52.4	42.1	**53.7**

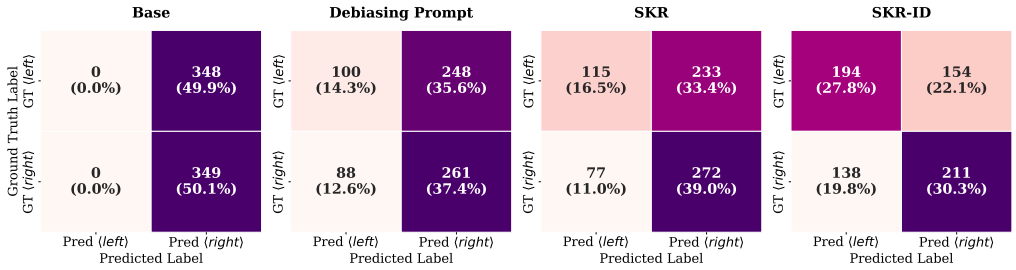

Figure 3: Confusion matrices change of LLaVA-OneVision model through the SKR-ID pipeline

We also evaluate SKR-ID on a test subset with a 70/30 class distribution, with results reported in Appendix D. These results further demonstrate SKR-ID's effectiveness under the tested class-distribution shift, rather than only when the test and training data have similar class distributions.

7.4 Bias Analysis: Evolution of the Confusion Matrices

Figure 3 visualizes the evolution of prediction distributions for the LLaVA-OneVision model through each stage of the SKR-ID pipeline. This qualitative analysis reveals how the heavily biased outputs are progressively rectified.

As shown in the first matrix, the base model exhibits a total collapse of the output distribution. As analyzed in Section 4, LLaVA-OneVision's direct outputs are heavily biased to label $\langle right \rangle$. This confirms that the base model is not performing meaningful geometric reasoning but is instead trapped in a deep-seated language prior. The introduction of the object-centric debiasing prompt effectively activates the model's reasoning ability, anchored in 2D object tracking. While this breaks the initial collapse, expanding the prediction space to cover both labels, the results remain heavily skewed.

The transition to the recursive refinement loop marks a decisive change in performance. The first round of SKR (replacing the NTP head with a classification head) shows a clear trend of shifting the confusion matrix toward a diagonal structure. The following recursive SKR process, using logit-ranking debiasing, achieves a clear diagonal shape. This transition confirms that SKR-ID has successfully decoupled the model's visual representations from its pre-training textual biases, transforming the VLM into a specialized VRRPI identifier capable of reasoning 3D spatial change.

8 Conclusions, Limitations, and Future Work

We have demonstrated that VLMs possess latent spatial knowledge that is often obscured by systematic biases and the limitations of token generation. By using an iterative SKR method with median-logit ranking (SKR-ID), we successfully force the model to learn from a balanced representation of its own knowledge, effectively debiasing the spatial reasoning process for VRRPI.

Despite the effectiveness of our method, our work has two primary limitations. First, we limited our evaluation to three specific backbones. However, our selection includes the two best-performing models of differing architectures from the VRRPI benchmark and the latest spatial-data fine-tuned VST-RL model. Therefore, they represent the competitive frontier of open-source VLM performance on this task. Second, the randomly extracted unannotated image pairs are heavily noised, including motion blur and geometric inconsistencies relative to the VRRPI-Bench standards. We intentionally retained these noises to prove that SKR-ID has minimal requirements for training data and that raw data can be extracted with marginal human effort. However, this also indicates significant potential for further performance improvement through including quality filtering.

This work opens a new perspective on VLM deficiencies by identifying stable output biases that likely affect a wide range of multimodal tasks beyond VRRPI. Our SKR-ID method shows strong potential for further improvement and for extension to a broader range of vision tasks.

Acknowledgments

This work was supported by the UKRI Centre for Doctoral Training (CDT) in Natural Language Processing through UKRI grant EP/S022481/1. We appreciate the comments and feedback by the reviewers and the area chair. We are grateful for funding provided by G-Research and St Catherine's College, Oxford to disseminate this study. We acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023] [20]. We also thank the Edinburgh Language Model team in their support. The authors declare no competing interests.

References

[1] Hikaru Asano, Tadashi Kozuno, and Yukino Baba. Self iterative label refinement via robust unlabeled learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] Wentao Bao, Kai Li, Yuxiao Chen, Deep A Patel, Martin Renqiang Min, and Yu Kong. Exploiting vlm localizability and semantics for open vocabulary action detection. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 8280–8290, 2025.

[4] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14455–14465, June 2024.

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

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[7] Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7395–7408, 2025.

[8] Ken Deng, Yifu Qiu, Yoni Kasten, Shay B Cohen, and Yftah Ziser. Lost in space? vision-language models struggle with relative camera pose estimation. *arXiv preprint arXiv:2601.22228*, 2026.

[9] Mengfei Du, Binhao Wu, Zejun Li, Xuan-Jing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–355, 2024.

[10] Ben Glocker, Shahram Izadi, Jamie Shotton, and Antonio Criminisi. Real-time RGB-D camera relocalization. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 173–179, October 2013. doi: 10.1109/ISMAR.2013.6671777.

[11] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[12] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

[13] Xinmiao Huang, Qisong He, Zhenglin Huang, Boxuan Wang, Zhuoyun Li, Guangliang Cheng, Yi Dong, and Xiaowei Huang. Spatial-dise: A unified benchmark for evaluating spatial reasoning in vision-language models, 2026. URL https://arxiv.org/abs/2510.13394.

[14] Haoyi Jiang et al. Spa3r: Predictive spatial field modeling for 3d visual reasoning. *arXiv preprint arXiv:2602.21186*, 2026.

[15] Raphi Kang et al. Linear mechanisms for spatiotemporal reasoning in vision language models. *arXiv preprint arXiv:2601.12626*, 2026.

[16] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. Open-vocabulary object detection upon frozen vision and language models. In *The Eleventh International Conference on Learning Representations*.

[17] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*.

[18] Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025.

[19] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[20] Simon McIntosh-Smith, Sadaf R Alam, and Christopher Woods. Isambard-ai: a leadership class supercomputer optimised specifically for artificial intelligence, 2024. URL https://arxiv.org/abs/2410.11199.

[21] Kevin Qu et al. Loc3r-vlm: Language-based localization and 3D reasoning with vision-language models. *arXiv preprint arXiv:2603.18002*, 2026.

[22] Rahul Raja, Arpita Vats, Omkar Thawakar, and Tajamul Ashraf. Object tracking: A comprehensive survey from classical approaches to large vision-language and foundation models. *Available at SSRN 5541079*, 2025.

[23] Prateeth Rao and Sachit Rao. Geometry-aware epipolar correspondence graphs for monocular visual odometry, 2026. URL https://arxiv.org/abs/2604.04554.

[24] Fatemeh Shiri, Xiao-Yu Guo, Mona Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21440–21455, 2024.

[25] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2937, 2013.

[26] Ilias Stogiannidis, Steven McDonagh, and Sotirios A Tsaftaris. Mind the gap: Benchmarking spatial reasoning in vision-language models. In *Greeks in AI Symposium 2025*.

[27] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021.

[28] Yiming Sun, Fan Yu, Shaoxiang Chen, Yu Zhang, Junwei Huang, Yang Li, Chenhui Li, and Changbo Wang. Chattracker: Enhancing visual tracking performance via chatting with multimodal large language model. *Advances in Neural Information Processing Systems*, 37: 39303–39324, 2024.

[29] Mengyu Wang, Yijia Shao, Haowei Lin, Wenpeng Hu, and Bing Liu. Cmg: A class-mixed generation approach to out-of-distribution detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 502–518. Springer, 2022.

[30] Mengyu Wang, Xiaoying Zhi, Zhiyi Li, Robin Schmucker, Shay B. Cohen, Tiejun Ma, and Fran Silavong. Self-knowledge re-expression: A fully local method for adapting llms to tasks using intrinsic knowledge. *arXiv preprint arXiv:2604.22939*, 2026.

[31] Qian-Wei Wang et al. Fine-tuning pre-trained vision-language models in a human-annotation-free manner. *arXiv preprint arXiv:2602.04337*, 2026.

[32] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025.

[33] Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025.

[34] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025.

[35] Xiaoda Yang et al. A progressive training strategy for vision-language models to counteract spatio-temporal hallucinations in embodied reasoning. *arXiv preprint arXiv:2604.10506*, 2026.

[36] Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 12000–12008, 2026.

[37] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025.

[38] Jiang Zhang et al. Spatialstack: Layered geometry-language fusion for 3d vlm spatial reasoning. *arXiv preprint arXiv:2603.27437*, 2026.

[39] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8): 5625–5644, 2024.

[40] Xuejun Zhang et al. Predicting camera pose from perspective descriptions for spatial reasoning. *arXiv preprint arXiv:2602.06041*, 2026.

[41] Yiming Zhang et al. Revsi: Rebuilding visual spatial intelligence evaluation for accurate assessment of vlm 3d reasoning. *arXiv preprint arXiv:2604.24300*, 2026.

[42] Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 8954–8975, 2024.

A Prompts

We show the prompt variants mentioned in Section 4 (For each prompt template, we instruct VLMs in two ways: 1. to output the index of option; 2. to output the option itself. We report the best performance under these two output-formats.):

(1) Base inputs prompt (the same as image-swap inputs prompt):

Prompt for base inputs.

```
You are given two images from the same scene but different
    perspectives. The first image is the source image (src) and
    the second image is the target image (tgt). Based on these two
    images, determine if the camera movement from src to tgt is
    Move left while yawing right or Move right while yawing left.
    Please directly output your option. Do not output anything
    else.
```

(2) Option-swap inputs prompt to test positional bias:

Prompt for option-swap inputs.

```
You are given two images from the same scene but different
    perspectives. The first image is the source image (src) and
    the second image is the target image (tgt). Based on these two
    images, determine if the camera movement from src to tgt is
    Move right while yawing left or Move left while yawing right.
    Please directly output your option. Do not output anything
    else.
```

(3) Extended prompt with an additional $\langle neither \rangle$ option:

Prompt for base inputs (with extra option).

```
You are given two images from the same scene but different
    perspectives. The first image is the source image (src) and
    the second image is the target image (tgt). Based on these two
    images, determine if the camera movement from src to tgt is
    Move left while yawing right or Move right while yawing left
    or Neither of them. Please directly output your option. Do not
    output anything else.
```

We also present the debiasing object-centric prompt here:

Debiasing object-centric prompt

```
You are given two images from the same scene but different
    perspectives. The first image is the source image (src) and
    the second image is the target image (tgt). First, identify a
    common object that appears in both images. Second, determine
    if this common object moved left or moved right from the src
    image to the tgt image.
```

B Complete Evaluation Results on VRRPI-Bench

Table 4 supplements the Macro-F1 results in Table 2 with accuracy (Acc), balanced accuracy (Bal Acc), per-class F1 (C1 F1, C2 F1), prediction ratios (C1 Ratio, C2 Ratio), and AUROC. The prediction ratios show the degree of output collapse in the base models. The balanced accuracy, per-class F1, and AUROC provide complementary measures of class discrimination. Across all three backbones, SKR-ID improves these measures and produces less collapsed predictions.

Table 4: Complete evaluation results on the balanced VRRPI-Bench test set. All metrics are reported as percentages.

Backbone	Method	F1	Acc	Bal Acc	C1 F1	C2 F1	C1 Ratio	C2 Ratio	AUROC
LLaVA-OV	Base	35.0	50.4	50.4	66.7	3.4	98.7	1.3	50.4
	SKR-ID	**58.0**	**58.1**	**58.1**	57.1	59.1	47.6	52.4	**60.4**
Qwen3-VL	Base	39.5	48.1	48.0	16.2	62.9	10.3	87.2	49.4
	SKR-ID	**53.8**	**55.8**	**55.8**	44.2	63.4	29.3	70.7	**57.7**
VST-RL	Base	45.4	46.6	46.6	37.4	53.5	35.3	64.7	46.6
	SKR-ID	**53.7**	**54.1**	**54.1**	**49.2**	**58.1**	40.5	59.5	**57.3**

C Detailed Ablation Study of SKR-ID

To study SKR-ID in greater detail, we conduct an extended ablation study on LLaVA-OneVision. We compare the following settings:
1. **B**: Directly query the base model.
2. **B + P**: Query the base model using the object-centric prompt.
3. **B + P + SKR**: Use the object-centric prompt to generate pseudo-labels for one round of SKR.
4. **B + P + SKR + I (without M)**: Iteratively apply SKR using the prompt-generated pseudo-labels, without median-based balancing.
5. **B + SKR + M + I (without P)**: Initialize SKR using the base model's outputs, then iteratively refine the pseudo-labels using median-based balancing.
6. **Full SKR-ID**: Use the object-centric prompt for initialization, followed by iterative SKR with median-based balancing.
7. **Random guess**: Randomly predict one of the two selections.
8. **Random guess SKR**: Initialize SKR with randomly generated balanced labels.
9. **Random guess SKR-ID**: Initialize with randomly generated balanced labels, then apply the full iterative SKR-ID procedure.
10. **Prompt SC3**: Apply self-consistency to three independently generated responses to the object-centric prompt.
11. **Prompt SC5**: Apply self-consistency to five independently generated responses to the object-centric prompt.
12. **Probing**: Train a probing layer on the model's final representations without adding adapters.
13. **Probing with M + I**: Iteratively train the probing layer using median-based label balancing.

Results are reported in Table 5. Iterative SKR without median thresholding gives only a limited improvement over one-round SKR, while omitting the object-centric prompt causes the initialization to collapse to one class. Random balanced labels can break this initial collapse, but random-label SKR and SKR-ID remain near the balanced-random result, indicating that class balance alone does not supply informative image-label correspondences. Prompt self-consistency provides only limited improvement over a single prompted response. The probing baselines also remain below the full SKR-ID result, including when combined with iterative median-based refinement.

Table 5: Extended ablation study on LLaVA-OneVision. All metrics are reported as percentages. SC denotes self-consistency; probing trains a linear classification layer on the final multimodal representations without adapters.

Setting	F1	Acc	Bal Acc	C1 F1	C2 F1	C1 Ratio	C2 Ratio	AUROC
B	35.2	51.7	50.2	67.9	2.4	98.9	1.1	50.2
B + P	49.1	51.8	51.8	37.3	60.8	27.0	73.0	51.8
B + P + SKR	53.2	55.5	55.5	42.6	63.7	27.6	72.5	59.3
B + P + SKR + I (without M)	54.0	55.2	55.2	46.6	61.5	33.9	66.1	57.4
B + SKR + M + I (without P)	33.3	49.9	50.0	66.7	0.0	100.0	0.0	50.0
Full SKR-ID	**58.0**	**58.1**	**58.1**	**57.1**	59.1	47.6	52.4	**60.4**
Random guess	50.3	50.3	50.3	49.0	51.0	49.9	50.1	50.3
Random guess SKR	47.6	52.1	52.0	32.4	62.9	21.0	79.1	55.3
Random guess SKR-ID	50.8	51.7	51.6	44.5	57.2	37.2	62.8	51.6
Prompt SC3	49.6	50.5	50.5	56.4	42.8	64.0	36.0	50.5
Prompt SC5	49.8	50.8	50.8	56.7	42.9	63.8	36.2	50.8
Probing	34.2	50.2	50.2	1.7	66.7	0.7	99.3	60.5
Probing with M + I	50.2	52.1	52.1	40.6	59.9	30.7	69.3	54.6

D Evaluation on Imbalanced Data

Table 6 reports results on a VRRPI-Bench test subset with a 70/30 class distribution. SKR-ID consistently improves performance over the corresponding base model. These results further demonstrate SKR-ID's effectiveness under the tested class-distribution shift, rather than only when the test and training data have similar class distributions.

Table 6: Evaluation on the 70/30 imbalanced VRRPI-Bench test subset. All metrics are reported as percentages.

Backbone		F1	Acc	Bal Acc	C1 F1	C2 F1	C1 Ratio	C2 Ratio	AUROC
LLaVA-OV	Base	41.9	**69.8**	50.1	**82.1**	1.6	99.3	0.7	50.1
	SKR-ID	**53.2**	55.5	**55.6**	63.5	**43.0**	52.0	48.0	**57.0**
Qwen3-VL	Base	30.0	32.0	47.1	16.3	43.4	9.8	87.6	48.3
	SKR-ID	**46.2**	**46.5**	**51.8**	**50.2**	42.2	37.5	62.5	**54.4**
VST-RL	Base	40.0	40.0	46.7	41.2	38.8	32.0	68.0	46.7
	SKR-ID	**44.3**	**44.3**	**52.6**	**44.4**	**44.1**	30.3	69.8	**56.4**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction clearly states the main contributions: 1. a comprehensive analysis on 3D spatial bias; 2. an annotation-free method SKR-ID, which helps mitigate the bias and re-express VLM's knowledge. The experimental results in Section 4 and Section 7 support these claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to Section 8.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not contain formal theoretical results, theorems, or proofs. Our claims are supported through empirical evaluation and analysis.

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please refer to Section 6 for details about the experimental settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Please refer to Section 3 for the dataset and to our GitHub repository for the released code and reproduction instructions: `https://github.com/astudentuser/Codes_of_SKR-ID`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to Section 6 for details about the experimental settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Please refer to Section 6 for details about the experimental settings and Section 7 for the results. We claim that all results are statistically significant and averaged over three runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to Section 6 for details about computing resources used in the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read the NeurIPS Code of Ethics and confirm that our research conforms to it in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to Section 8.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not have such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the assets used in this paper are properly credited and the license and terms of use are explicitly mentioned and properly respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code for SKR-ID has been released on GitHub: `https://github.com/astudentuser/Codes_of_SKR-ID`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: This paper focuses on enhancing the spatial understanding of vision Large Language Models (LLMs).

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.