Gaze-Regularized Vision-Language-Action Models for Robotic Manipulation

Supplementary Material

This supplementary document provides extended methodological details, additional ablations and implementation clarifications to support the claims made in the main paper. The structure is as follows:

- Appendix A: Notation Table
- Appendix B: Expanded Methodological Clarifications
- Appendix C: Additional Attention - Gaze Alignment Evidence
- Appendix D: Synthetic Gaze Reliability and Ablations
- Appendix E: Other Experiments
- Appendix F: Pseudo-code and Reproducibility Details
- Appendix G: Summary of Additions and Discussion

A. Notation and Symbol Table

To improve clarity and provide a quick reference for readers, we summarize the key notations used throughout the paper and supplementary material. These symbols cover visual tokens, patch grids, gaze heatmaps, attention matrices, and their corresponding distributions.

B. Expanded Methodological Clarifications

In this section, we provide additional details on how gaze supervision is integrated into the VLA architecture. We first clarify how spatial attention is extracted and regularized, then discuss the properties and reliability of the predicted gaze signals used throughout our experiments. These clarifications are intended to make the connection between model internals, gaze priors, and action prediction more explicit than in the main paper.

Constructing a Singular Global Query from Language Tokens. To obtain a unified representation of the instruction, we collapse the sequence of language embeddings into a single global query vector. This can be implemented through simple pooling, a learned linear projection, or a lightweight attention-based aggregator; in our implementation, a simple projection maps the full language-token sequence $\{X_l^{(1)}, \ldots, X_l^{(N_l)}\}$ into a compact semantic vector Q_{lang}. This vector captures the dominant intent of the instruction and serves as a query over the visual scene.

Detailed Attention Extraction Our approach introduces gaze-guided supervision into the VLA model by regularizing its *internal spatial attention* during training. Since robots do not possess an innate mechanism analogous to human eye-gaze, the goal is to endow the policy with a learned surrogate of gaze i.e a structured prior that encourages the transformer to focus on task-relevant regions during manipulation.

The spatial attention regularized in our framework emerges from the interaction between the vision and language streams in the final transformer layer of the VLA backbone. The language encoder first produces a sequence of instruction tokens $X_l \in \mathbb{R}^{N_l \times d}$, which are aggregated through a learned projection to form a global query vector (as mentioned in the previous paragraph) $Q_{\text{lang}}^{(l)}$. This query functions as a compact representation of the semantics of the task instruction.

For each camera view i, the visual encoder outputs a set of tokens $X_v^i \in \mathbb{R}^{N_v \times d}$, which are linearly projected to key vectors $K_{\text{view}_i}^{(l)}$, following the standard attention formulation established in [46]. The resulting cross-attention captures the degree to which each visual patch is relevant to the language instruction:

$$S_t^i = \text{Softmax}\left(\frac{Q_{\text{lang}}^{(l)} K_{\text{view}_i}^{(l)\top}}{\sqrt{d}}\right) \in \mathbb{R}^{1 \times N_v}.$$

This attention distribution quantifies the importance assigned to each visual token when interpreting the task instruction. We extract the attention distribution specifically from the **final vision–language transformer layer**, for two reasons:

1. **Semantic maturity.** Late transformer layers might contain the most semantically integrated features, combining spatial, linguistic, and contextual cues.
2. **Action relevance.** In Pi-0 and other VLA architectures, the action tokens attend to the fused representations produced by the final vision–language layer. Thus, regularizing this layer directly shapes the perceptual information used for motor prediction.

This design parallels observations from prior work such as [35], which shows that late-layer attention better reflects task-relevant perceptual cues. However, unlike prior methods, our approach applies this principle to **robotic control settings**, where attention not only guides prediction but directly influences action generation.

Aligning this spatial attention with human gaze priors yields an inductive bias that is both *compact* and *action-grounded*. This approach mirrors core aspects of human behavior: just as humans internalize a rich understanding of a scene–fusing visual cues with linguistic and contextual knowledge before executing a precise motor action, our method regularizes the model's final representations to

Table 7. Summary of key notations used in gaze-to-attention regularization and VLA token interactions.

Symbol	Description
t	Timestep index of the current observation
i	Camera/view index
N_v	Number of visual tokens (e.g., $16 \times 16 = 256$)
P	Patch grid dimension (e.g., $P = 16$)
$X_l \in \mathbb{R}^{N_l \times d}$	Language token sequence
$X_v^i \in \mathbb{R}^{N_v \times d}$	Visual tokens from camera view i
$Q_{\text{lang}}^{(l)} \in \mathbb{R}^{1 \times d}$	Global query summarizing language semantics
$K_{\text{view}_i}^{(l)} \in \mathbb{R}^{N_v \times d}$	Key vectors for visual tokens from view i
$H_t^i \in \mathbb{R}^{H_g \times W_g}$	Predicted gaze heatmap for view i at time t
$\tilde{H}_t^i$	Temporally aggregated gaze heatmap centered at t
$G_t^i \in \mathbb{R}^{N_v}$	Patch-level gaze distribution for view i
$S_t^i \in \mathbb{R}^{N_v}$	Model's spatial attention over visual tokens
$D_{\text{KL}}(G_{i,t} \parallel S_{i,t})$	KL divergence measuring gaze–attention alignment
I_t^i	RGB frame from view i at time t
ℓ_t	Tokenized language instruction
q_t	Proprioceptive observation at time t
A_t	Predicted short-horizon action sequence
A_t^*	Ground-truth action sequence
λ	Gaze-regularization weighting coefficient
T	Temporal aggregation window size for gaze

guide its decisions. Consequently, the policy is encouraged to mirror the fixation and information-gathering strategies humans employ before and during manipulation.

B.1. Reliability of Predicted Gaze

Because robotic datasets rarely include human eye-tracking labels, we employ *synthetic gaze* generated by pretrained gaze-estimation networks. Among existing models, we adopt the Global–Local Correlation (GLC) network [25] due to a combination of temporal fidelity, robustness, and strong performance on egocentric video tasks.

Temporal Sensitivity. Human gaze during manipulation is inherently dynamic: fixations shift in anticipation of upcoming hand movements. GLC explicitly models these temporal dependencies by processing short clips rather than single frames, producing gaze heatmaps informed by both past and future context. This confers a key advantage over earlier single-frame models such as DeepGaze [24], although DeepGaze and it's new variants show great performance in tasks which require a scanning pattern over a static scene, and in the future, this ability can be leveraged to make our method even better.

Strong Performance in Manipulation-like Settings. GLC achieves high accuracy on egocentric and hand–object interaction datasets, which share structural similarities with robotic manipulation scenes (clutter, hand presence, fine-grained object interactions). These properties make GLC particularly suitable for generating gaze priors for multi-view robotic datasets. In the future, curated teleoperated datasets with ground-truth gaze could further improve interpretability and accuracy by providing real human fixation patterns rather than synthetic estimates.

Ablations on Gaze Quality. To verify that performance improvements stem from meaningful gaze characteristics rather than incidental regularization, we perform additional robustness experiments (see later appendices):

- **DeepGaze comparison:** replacing GLC with DeepGaze reduces performance, indicating that accurate spatial structure of gaze is important.
- **Uniform Gaze:** by equally dividing attention across all the patches, the benefits are not seen anymore, confirming that only *structured* gaze provides useful supervision.

While synthetic gaze is inherently an approximation of true human fixation behavior, our experiments demonstrate that it provides a powerful supervisory signal for shaping transformer attention. We view our results as an initial bound on the benefits achievable with real eye-tracking, and anticipate even greater gains as future teleoperation datasets incorporate true human gaze measurements.

Table 8. Per-task success rates on LIBERO Spatial [29] at 30k training steps. We compare the baseline model, our gaze-regularized model, a DeepGaze-based gaze variant, and a uniform-distribution variant.

Location of Object	w Gaze	DeepGaze	w/o Gaze	Uniform
	30k	30k	30k	30k
Between plate and ramekin	100	85.7	83.3	69.7
Next to ramekin	100	86.7	85.7	59.7
Table center	100	100	100	80.3
On cookie box	91.3	100	100	79.3
In cabinet drawer	73.3	82.0	80	39.3
On ramekin	100	100	100	50.7
Next to cookie box	100	100	100	50.3
On stove	90	91.0	90	10.3
Next to plate	100	55.0	50	70.7
On wooden cabinet	100	73.3	70.3	60.3
Overall Avg.	**95.5**	86.3	85.9	57.1

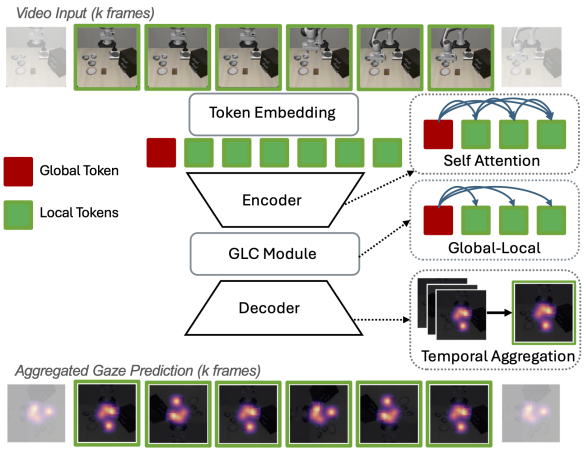

Figure 4. **Closer look at Gaze Prior Generation** A sequence of k video frames is tokenized and processed by the GLC [25] module, where it utilizes global tokens (derived from the sequence) and local tokens, and undergoes self attention as well as Global-Local Correlation to then predict per-frame gaze heatmaps. These heatmaps are temporally aggregated to yield a gaze distribution that captures attention over time and serves as the supervision signal for training- time regularization.

C. Attention–Gaze Alignment Evidence

Beyond task success rates, a core claim of our work is that gaze regularization shapes the model's internal attention to better reflect human fixation patterns. In this section, we first introduce a quantitative Top-k overlap metric to measure alignment between model attention and gaze distributions, and then provide additional qualitative visualizations to illustrate how this alignment manifests across tasks, viewpoints, and time.

C.1. Top-k Attention–Gaze Overlap Metrics

A central question in evaluating our framework is whether gaze regularization meaningfully shifts the model's internal attention toward human fixation patterns. While qualitative visualizations already suggest improved alignment, we seek a more rigorous quantitative measure. To this end, we compute a *Top-k attention–gaze overlap* metric that assesses how frequently the model's most attended patches coincide with regions prioritized by human gaze. For our experiment, we use a value of k=10.

Metric Definition. For each view i at time t, let $S_t^i \in \mathbb{R}^{N_v}$ denote the model's spatial attention distribution and $G_t^i \in \mathbb{R}^{N_v}$ denote the gaze-derived patch-level distribution. We identify the indices of the model's k highest-attended patches:

$$\mathcal{T}_k(S_t^i) = \text{Top-}k(S_t^i).$$

We then compute the total gaze mass contained within these patches:

$$\text{Overlap}_k(t,i) = \sum_{j \in \mathcal{T}_k(S_t^i)} G_{t,j}^i.$$

This yields a score in $[0,1]$, where a value of 1 indicates that all gaze probability lies within the model's top-k attended patches, and 0 indicates complete misalignment.

We observe a substantial improvement in overlap after applying gaze regularization. For example, at k=10, the baseline model achieves an average overlap of 19%, whereas the gaze-regularized model achieves 51%. The relative improvement indicates that the regularized model attends more sharply to the most gaze-salient regions, as shown in Figure 5.

C.2. Attention Map Visualizations

To complement the Top-k quantitative analysis, we include an additional qualitative comparison of spatial attention maps across three settings: the baseline model (no gaze), a model trained with a gaze variant, and our proposed gaze-regularized model. This visualization clearly highlights the characteristic differences produced by each training scheme.

Across all views shown in Figure 5, we observe that the baseline model exhibits diffuse and spatially inconsistent attention, often spreading mass across irrelevant background regions. Using an uniform gaze prior produces diffused attention as well , and still lacks strong task grounding. In contrast, our method produces sharply localized and semantically aligned attention, focusing on regions directly relevant to the instructed manipulation.

These visual patterns are consistent with and supportive of the Top-k overlap results reported earlier: the gaze-regularized model's attention aligns more closely with human fixation structure, reflecting a more task-aware perceptual representation.

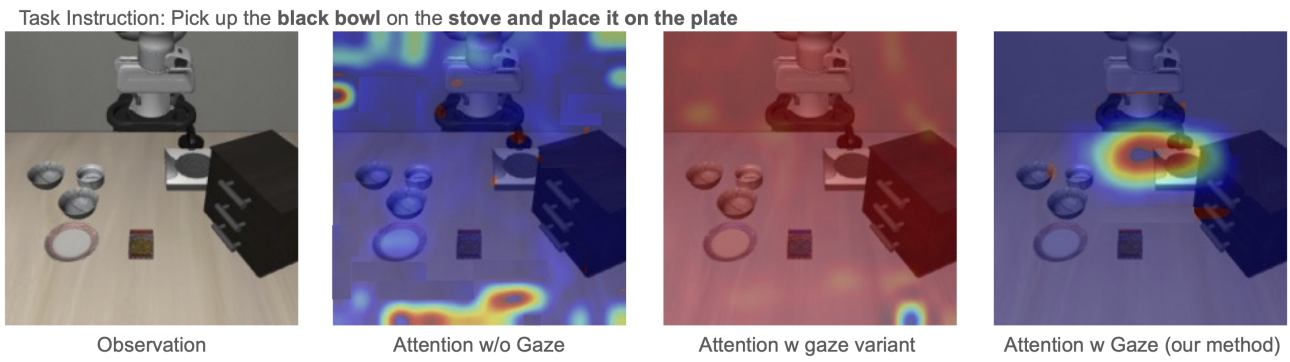

Figure 5. **Additional Visualisations of Attention.** Given the input observation, we show the spatial attention from the baseline model (second), the attention obtained when a perturbed gaze variant is used (third, corresponding to Table 8), and finally the sharper, task-relevant attention produced by our gaze-regularized model (fourth).

C.3. Attention Modulation using average representation of all layers

In the main paper, we regularize the spatial attention extracted from the *final* vision–language transformer layer. This design choice is motivated by the fact that the last layer contains the most semantically integrated features, and its attention maps directly govern the information available to the action tokens. A natural question, however, is whether distributing gaze supervision across *all* layers might further improve performance or stability.

To investigate this, we consider a variant in which we first compute the attention distribution at each transformer layer, then average these distributions across depth, and finally apply the gaze regularization loss to this layer-averaged attention. Intuitively, this variant encourages gaze-aligned information flow throughout the entire network, rather than only at the last layer.

Table 9 reports per-task success rates on LIBERO-Spatial when regularizing this averaged attention across all layers. We observe that this variant achieves competitive performance when compared to the baseline model across most spatial configurations and training checkpoints. At the same time, the results support our design choice in the main paper: concentrating gaze supervision on the final vision–language layer provides a larger increase in accuracy while incurring no additional overhead from multi-layer aggregation.

C.4. Task-Conditioned Gaze and Language-Conditioned VLM Attention

We establish that informative gaze during manipulation is task-dependent and that different language instructions can induce different gaze patterns. In our framework, gaze is predicted from short temporal sequences rather than single images, allowing the gaze model to exploit action progression and implicit task context. Since the model was trained using data from a task-driven setting rather than free viewing, the predicted aggregated gaze yields top-down, task-driven attention rather than bottom-up saliency. From Figure 6, we can also see that the temporal processing of a sequence of frames provides the task context, and hence produces different gaze results for different task instructions, even under similar settings.

Table 9. Per-task success rates on LIBERO Spatial with regularization applied to all layers. The model shows competitive performance with comprehensive regularization.

Location of Object	**w Gaze (All Layers)**		
	10k	20k	30k
Between plate and ramekin	65.0	75.0	90.3
Next to ramekin	55.0	70.0	89.7
Table center	70.0	85.0	100.0
On cookie box	58.3	65.0	70.3
In cabinet drawer	43.3	56.3	60.7
On ramekin	48.3	65.0	99.7
Next to cookie box	65.0	85.0	100.0
On stove	25.0	45.0	79.3
Next to plate	56.7	70.0	99.3
On wooden cabinet	38.3	55.0	80.3
Overall Avg.	52.5	67.1	**87**

Human visual attention in this work refers specifically to egocentric, action-oriented, top-down gaze during object manipulation. Fixations anticipate contact regions, targets, and task-relevant spatial relations rather than free-viewing or social gaze. Temporal aggregation further captures anticipatory fixations that precede motor execution, consistent with findings in the action-perception literature.

Crucially, language is explicitly incorporated through the VLM attention we regularize. The attention map is extracted using a global language token (derived from the in-

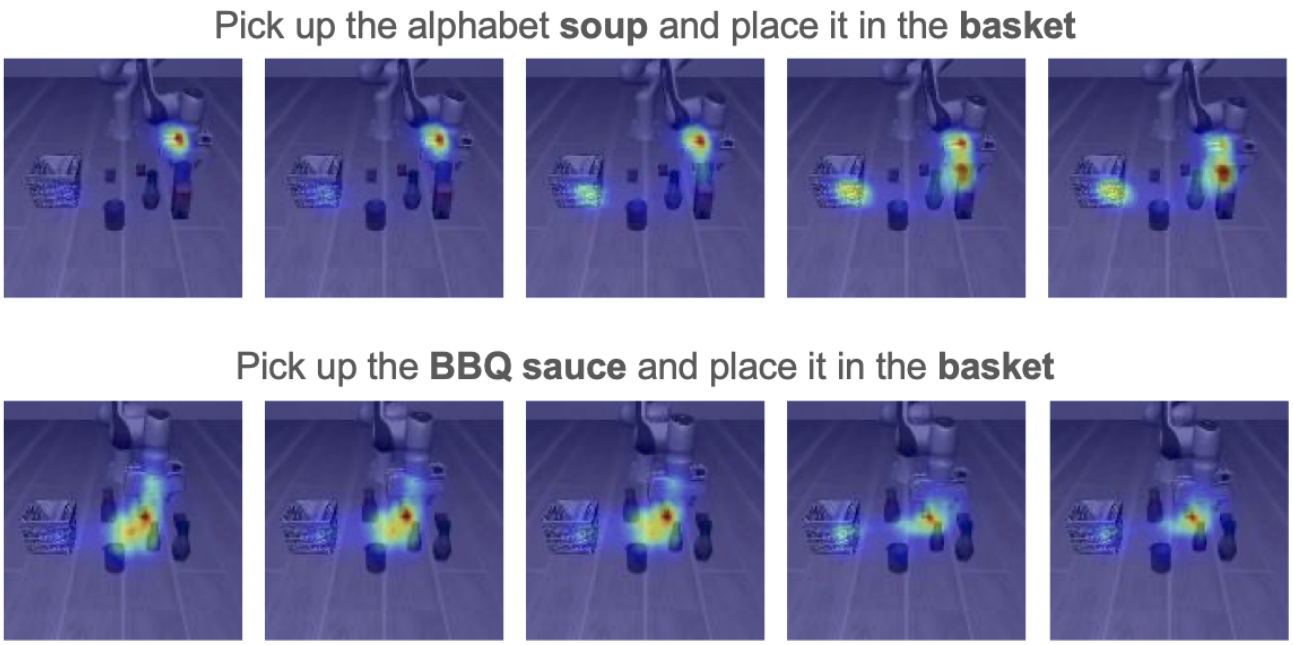

Figure 6. **Reliability of Synthetic Gaze on Simulation Videos** Given the input task, we show the the predicted gaze is accurate and even on similar visual settings, produces different gaze results depending on the language instruction. The model utilizes a temporal sequence of frames, rather than a single frame, and then computes the gaze prediction thus the prediction occurs due to the conditioning through the global context by operating on a sequence of frames

struction) as the query over visual tokens, making it inherently language and task-conditioned. Gaze therefore does not replace task reasoning; it provides a soft spatial prior that biases where a task-aware VLM attends. Regularization is applied as a soft constraint, and gaze–attention overlap is partial (i.e., 51% top-10 overlap in our method vs. 19% in baseline), not enforced to be identical.

If temporally predicted, image-only gaze were incompatible with language conditioned attention, performance would degrade on tasks with similar observations but different instructions (e.g., LIBERO). Instead, we observe consistent improvements across such settings, indicating that temporally predicted gaze complements, rather than conflicts with, task-aware VLM attention.

D. Other Experiments

Beyond the standard evaluation settings presented in the main paper, it is important to understand whether gaze regularization provides benefits under conditions that more closely resemble real-world deployment. Robots operating outside controlled laboratory environments routinely face perturbations in both visual observations and task instructions. In this appendix, we therefore expand our analysis to two additional scenarios: (i) linguistic perturbations that modify the phrasing of task instructions, and (ii) cross-viewpoint degradation where one of the camera inputs becomes unavailable. Together, these experiments shed light on the robustness and generalization properties of gaze-regularized VLA models.

D.1. Perturbations in Language Prompts as Task Distractors

While Section 4.2 introduces visual perturbations, linguistic perturbations can also serve as practical task distractors. Natural language in the real world is rarely fixed: users may rephrase commands, substitute synonyms, or give instructions with subtle differences in wording. To simulate such conditions, we manually replaced verbs in the LIBERO-Spatial instruction set with alternatives such as *grab*, *retrieve*, or *lift* in place of the canonical *pick*. All prompts were kept similar in length to avoid introducing length-based biases.

We then compare model performance under these instruction variations for both the baseline (without gaze regularization) and our gaze-regularized approach in Table 10. The drop in performance is similar across both models, but the gaze-regularized approach still performs better overall, even when the linguistic phrasing deviates from the distribution seen during training.

Table 10. Per-task success rates on LIBERO Spatial under prompt distractors (e.g., replacing "pick" with "grab", "lift", etc.). Both the baseline and gaze-regularized models exhibit performance degradation, but the gaze model remains more robust.

Location of Object	w Gaze (Distractors)	w/o Gaze (Distractors)
	30k	30k
Between plate and ramekin	96.7	78.3
Next to ramekin	95.0	80.0
Table center	97.0	96.7
On cookie box	90.0	96.7
In cabinet drawer	70.0	70.0
On ramekin	96.7	95.0
Next to cookie box	96.7	95.0
On stove	83.3	76.7
Next to plate	85.0	42.0
On wooden cabinet	93.3	60.0
Overall Avg.	**89.9**	79.1

Table 11. Per-task success rates on LIBERO Spatial [29] at 30k training steps. We compare the baseline model, our gaze-regularized model, and a foveated-vision variant.

Location of Object	w/o Gaze	w Gaze	Foveated
	30k	30k	30k
Between plate and ramekin	83.3	100	80.0
Next to ramekin	85.7	100	81.3
Table center	100	100	95.7
On cookie box	100	91.3	90.0
In cabinet drawer	80	73.3	65.3
On ramekin	100	100	90.0
Next to cookie box	100	100	94.0
On stove	90	90	80.7
Next to plate	50	100	44.7
On wooden cabinet	70.3	100	63.3
Overall Avg.	85.9	**95.5**	78.5

D.2. Foveated Vision during Training

Prior work has explored using gaze not only as supervision but also to reshape the visual input via foveated rendering, where regions near the gaze location are preserved at high resolution and the periphery is downsampled or blurred [10, 22, 55]. Following this idea, we implement a simple variant in which, for each timestep and view, we construct a foveated RGB image centered on the peak of the gaze distribution and feed this foveated image directly into the standard visual encoder, without changing any other part of the VLA pipeline.

Under a moderate foveation setting, this variant achieves an overall success rate of **78.5%** on LIBERO-Spatial, which is roughly **8 % lower** than our original non-foveated baseline (85.9%). We hypothesize that, in our multi-view manipulation setting, aggressively reducing peripheral detail removes useful contextual cues (e.g., table geometry, supporting surfaces, or alternative grasps) that the policy relies on for precise spatial reasoning.

Table 12. Per-task success rates on the Missing Views experiment at 30k training steps. The gaze-regularized model consistently outperforms the baseline across all spatial configurations.

Location of Object	w Gaze	w/o Gaze
	30k	30k
Between plate and ramekin	90.3	81.3
Next to ramekin	80.7	71.0
Table center	90.7	81.7
On cookie box	70.7	62.0
In cabinet drawer	69.3	60.7
On ramekin	40.7	34.7
Next to cookie box	69.7	60.7
On stove	21.0	17.7
Next to plate	39.3	32.3
On wooden cabinet	70.3	61.0
Overall Avg.	**64.3**	56.3

D.3. Cross-Viewpoint Robustness

Real-world manipulation often involves partial occlusions or temporary sensor failures. To evaluate robustness under such conditions, we remove one camera view at inference time by replacing its RGB frame with a blank image and measure performance on LIBERO-Spatial. Since models are never trained on missing views, this tests their ability to rely on the remaining cameras and maintain spatial consistency and thus, this scenario evaluates its inherent ability to compensate for missing perceptual input by relying on the remaining views and previously learned cross-view spatial consistency.. Both models experience a performance drop, but the gaze-regularized model consistently retains a higher success rate, indicating that gaze supervision encourages more stable and viewpoint-consistent attention, as shown in Table 12.

D.4. Using Gaze Variants

We further investigate whether different types of gaze supervision influence robustness by evaluating two additional variants: a model trained with DeepGaze [24] (a single-frame gaze predictor) and a Uniform Gaze model where gaze is evenly distributed across all patches. The DeepGaze variant performs moderately well but still falls short of our method, while the Uniform Gaze model exhibits the largest degradation. These trends align with our attention visualizations and Top-k overlap analysis: structured gaze supervision produces sharper, more task-relevant attention, whereas weak or uninformative priors lead to diffuse and unstable attention, reducing performance across tasks. The results are found in Table 8.

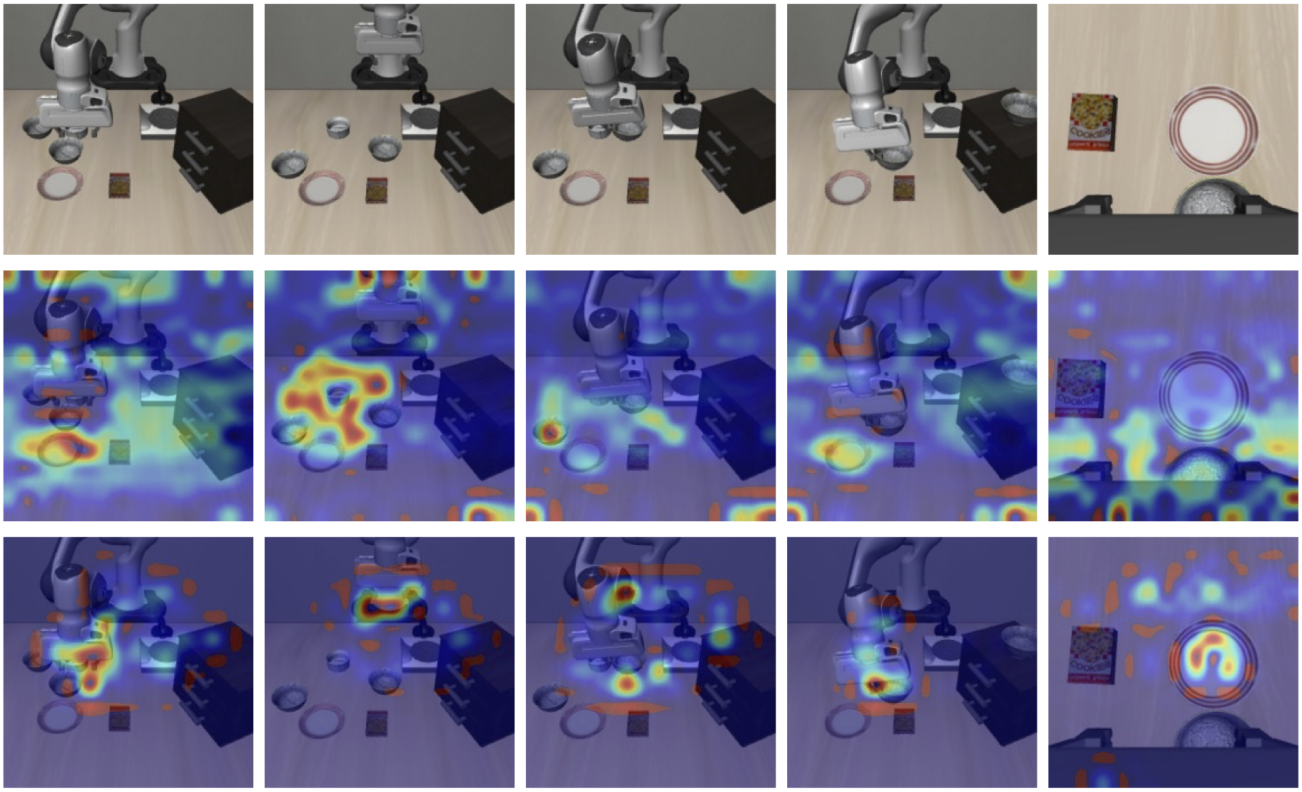

Figure 7. **Additional Visualisations of Attention.** Given the input observation (first), we show the spatial attention from the baseline model (second) and task-relevant attention produced by our gaze-regularized model (third).

D.5. Using Real Human Gaze for Fine-tuning GLC for Gaze Prediction

To enable human-guided gaze prediction for simulation videos, we conducted a data collection study using a screen-based eye tracker which we borrowed briefly for our study. Prior to collection, participants were briefed on each task instruction, ensuring they understood the objective before watching the corresponding simulation video. Their natural eye movements were recorded as they viewed these videos, providing ground truth gaze data for simulation environments. This collected data was then used to fine-tune the GLC model [25], adapting it from its original training on real-world videos to the domain of simulated robotic demonstrations. The resulting model was subsequently used to generate predicted gaze heatmaps for the LIBERO-Spatial benchmark tasks.

To validate the effectiveness of this approach, we compared the performance of our gaze-regularized policy against a baseline trained without gaze supervision. Across the LIBERO-Spatial tasks, the gaze-regularized model consistently outperformed the baseline, demonstrating that even simulation-derived gaze signals provide meaningful guidance for learning visuomotor policies. This performance gap suggests that human attention patterns encode valuable priors about task-relevant visual features that transfer effectively to policy learning.

Importantly, these results were achieved with a relatively modest dataset of human gaze collected specifically for simulation videos. We hypothesize that performance could be further improved by scaling up data collection efforts—incorporating more participants, more diverse tasks, and more finely calibrated eye tracking equipment. Such large-scale, high-quality human gaze data would enable even better adaptation of gaze prediction models to simulation domains, potentially unlocking further gains for gaze-regularized policies. This points to a promising direction for future work: leveraging human attention at scale as a readily accessible form of supervision for robot learning.

E. Pseudocode and Reproducibility

To facilitate reproduction and adaptation of our method, this appendix summarizes the key implementation components of the gaze-regularized training pipeline. We provide pseudocode for the heatmap-to-token projection used to align gaze with visual tokens to obtain the gaze-prior distribution, and for the overall training loop that integrates gaze

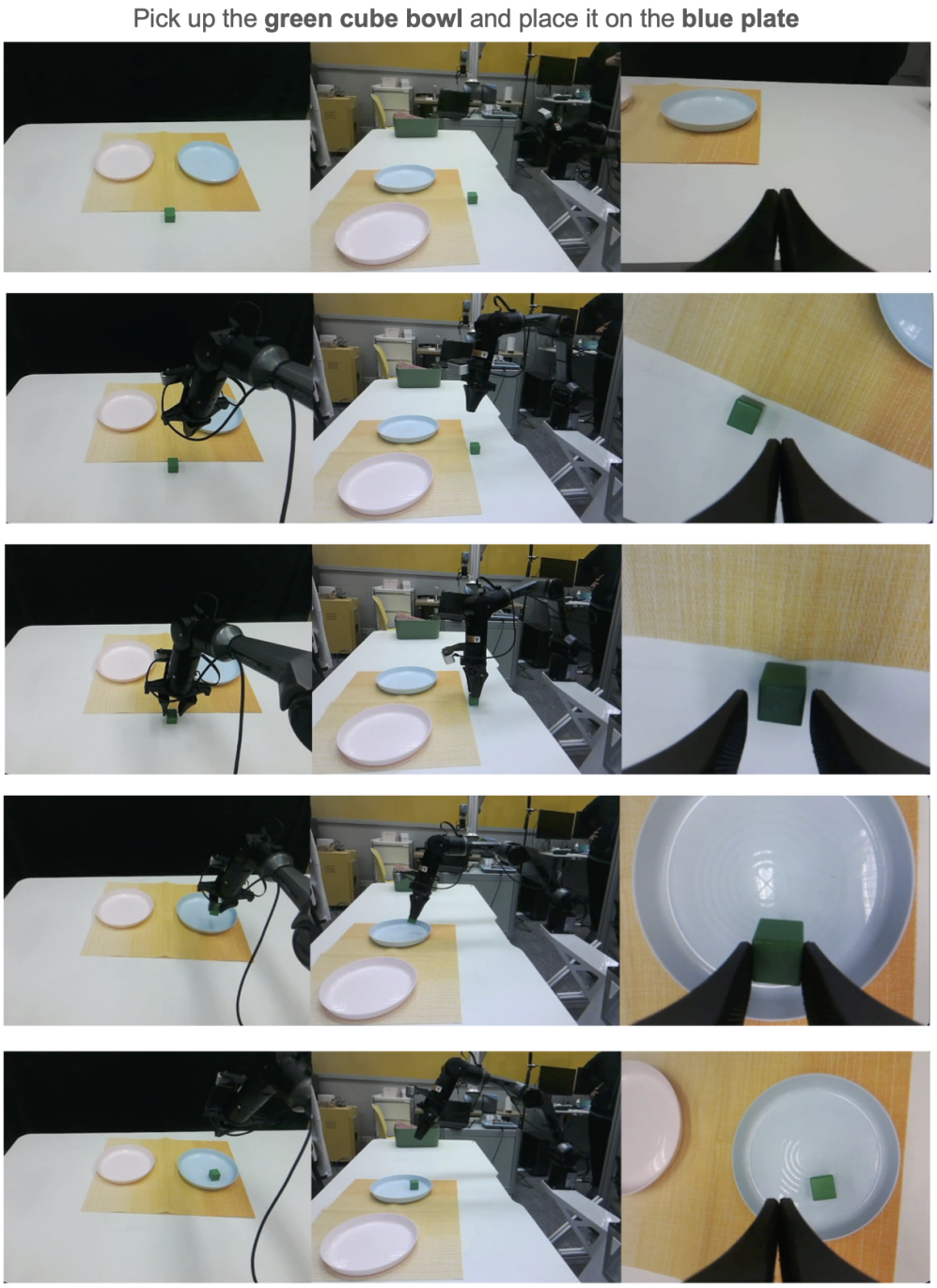

Figure 8. **Visualization of Real-world Task on Aloha Robot** In the figure, we provide some frames from a real world task performed using our gaze-regularized policy to show that our method works outside of simulation as well. Here, the task is to pick up the cube and place it on the correct plate.

regularization into standard VLA optimization.

E.1. Heatmap-to-Token Projection Pseudocode

In this section, we provide pseudocode for converting gaze heatmaps produced by the gaze prediction model into patch-level token distributions that are aligned with the transformer's visual tokens. This procedure is shared across Pi-0 and OpenVLA-based experiments, and can be implemented efficiently using standard tensor operations.

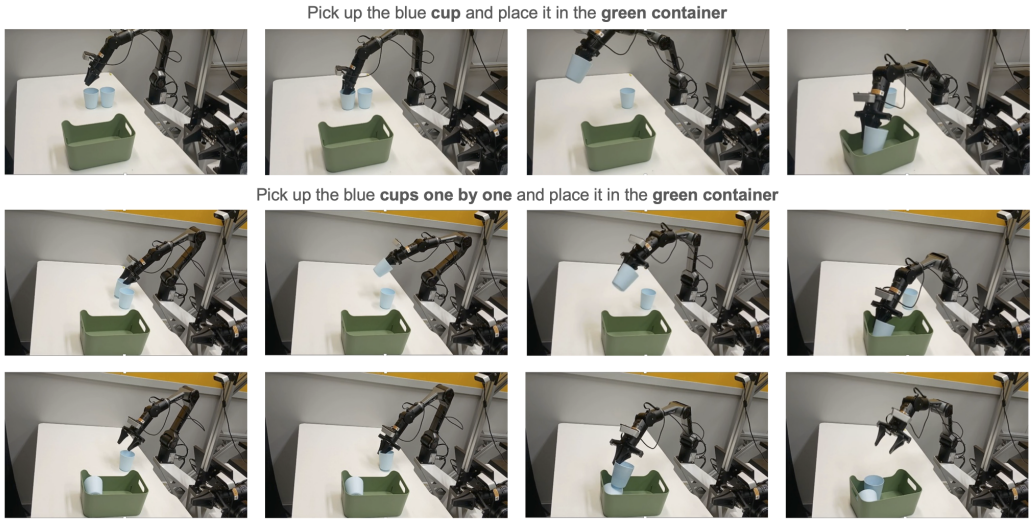

Figure 9. **Visualization of Real-world Task on Aloha Robot** In this figure, we present a short horizon task of picking up a cup and placing it in a container(top) and also another longer horizon task to pick up multiple cups one-by-one, and place them in the container. Both visualisations are obtained using our gaze-regularized policy, highlighting its working functionality even in real-world scenarios

Table 13. Per-task success rates on LIBERO Spatial [29] at 30k training steps. We compare the baseline model, our gaze-regularized model, and the human-gaze-trained variant.

Location of Object	w/o Gaze	w Gaze	Human Gaze
	30k	30k	30k
Between plate and ramekin	83.3	100	100
Next to ramekin	85.7	100	100
Table center	100	100	100
On cookie box	100	91.3	89.3
In cabinet drawer	80	73.3	78.3
On ramekin	100	100	100
Next to cookie box	100	100	100
On stove	90	90	90
Next to plate	50	100	100
On wooden cabinet	70.3	100	90
Overall Avg.	85.9	**95.5**	94.8

E.2. Training Loop with Gaze Regularization

We now provide pseudocode for the full training loop, including: (i) multimodal data loading, (ii) synthetic gaze generation via the GLC network, (iii) heatmap-to-token projection, and (iv) optimization with the combined action and gaze-regularization losses. The procedure is shared across all experiments (Pi-0 and OpenVLA backbones), with minor architecture-specific details encapsulated inside the policy forward pass.

Inference. At test time, we discard the entire gaze branch: no gaze model is invoked and no gaze distributions are computed. The policy operates as:

$$A_t = \pi_{\theta^*}(I_{1:n,t}, \ell_t, q_t),$$

relying only on visual, language, and proprioceptive inputs. The effect of gaze supervision is fully encoded in θ^*, manifesting as gaze-aligned internal attention without any inference-time overhead.

F. Summary of Additions

This supplementary document provides a set of analyses and implementation details that deepen and broaden the claims made in the main paper. We briefly summarize the key additions below and how they support our core hypotheses, and conclude with a discussion of our work.

Clarified notation and methodological details. We introduce a consolidated symbol table (Table 7) and expanded descriptions of how visual tokens, language tokens, and gaze-derived distributions interact within the VLA architecture. In particular, we detail how final-layer vision–language cross-attention is extracted, how it relates to action prediction, and why this layer is the most semantically meaningful target for gaze regularization.

Quantitative and qualitative evidence of attention–gaze alignment. Beyond task success rates, we define a Top-k attention–gaze overlap metric that directly measures how well the model's internal attention aligns with gaze-derived priors. Additional visualization of attention maps further illustrate that gaze regularization produces sharper, more task-relevant, and anticipatory attention patterns which aids the action prediction process.

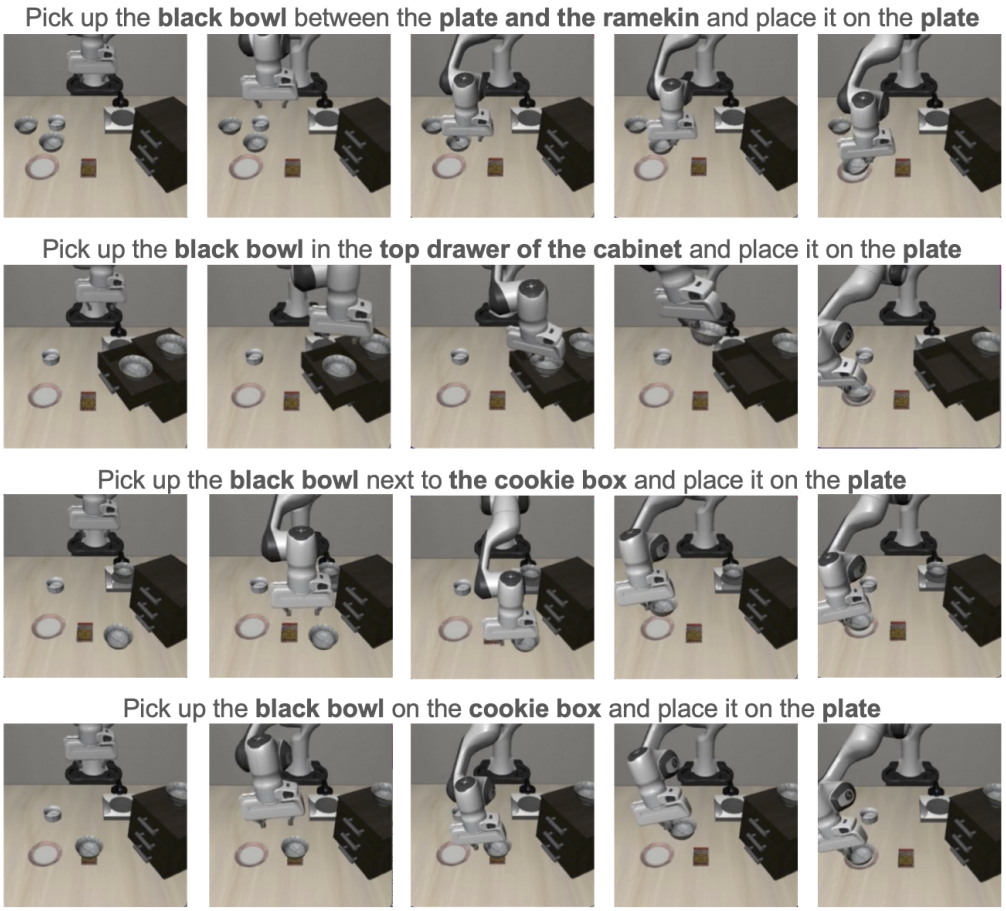

Figure 10. **Visualization Results** In the figure, we provide some visualization results to show how the policy performs on the Libero-Spatial [29] task suites. We provide the task instructions, and some important frames to show the task success. The baseline model performs admirably, but our method enhances the results by using gaze-regularization.

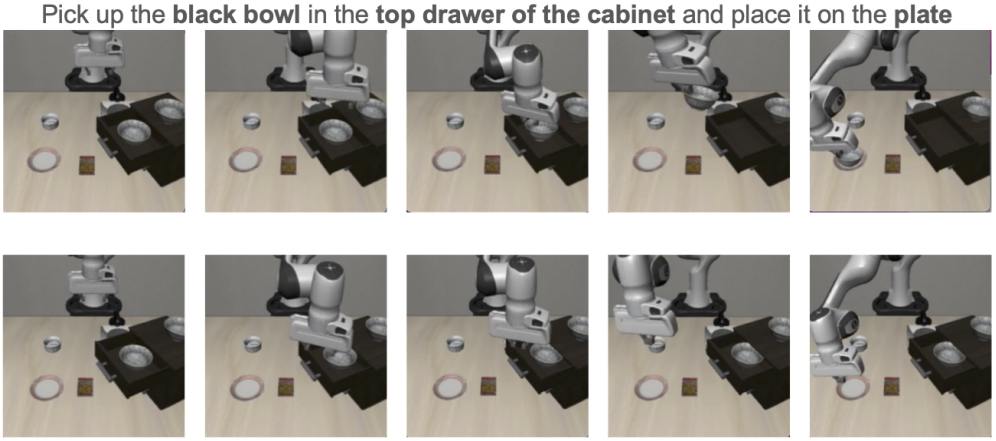

Figure 11. **Failure Case.** We show a failure example from the Libero-Spatial [29] task suite. In this task, the baseline model outperforms the gaze-regularized model, suggesting that stronger or more accurate gaze priors could further improve reliability. The bottom sequence illustrates the failure case where the robot hand fails to grab the bowl in the top drawer and proceeds to carry out the intended action.

Algorithm 1: Heatmap-to-Token Projection

Input: Gaze heatmap $H \in \mathbb{R}^{H_g \times W_g}$, patch grid size P (so $N_v = P^2$).
Output: Patch-level gaze distribution $G \in \mathbb{R}^{N_v}$.

Step 1: Normalize raw heatmap values.
Compute the sum of all heatmap values:

$$Z \leftarrow \sum_{x=1}^{H_g} \sum_{y=1}^{W_g} H(x, y).$$

If $Z = 0$, set $H(x, y) \leftarrow \frac{1}{H_g W_g}$ for all (x, y) (uniform map). Otherwise, normalize:

$$H(x, y) \leftarrow \frac{H(x, y)}{Z} \quad \forall x, y.$$

Step 2: Define patch grid.
Let each patch be of size

$$h_p = \left\lfloor \frac{H_g}{P} \right\rfloor, \quad w_p = \left\lfloor \frac{W_g}{P} \right\rfloor.$$

For patch indices $u, v \in \{0, \ldots, P-1\}$, the spatial region of patch (u, v) is:

$$\mathcal{P}_{u,v} = \{uh_p \le x < (u+1)h_p,\ vh_p \le y < (v+1)w_p\}.$$

Step 3: Aggregate heatmap values per patch.
Initialize $G \in \mathbb{R}^{N_v}$ with zeros.
for $u = 0$ **to** $P - 1$ **do**
  for $v = 0$ **to** $P - 1$ **do**
    $j \leftarrow u \cdot P + v$ `// flattened patch index`

$$G_j \leftarrow \sum_{(x,y) \in \mathcal{P}_{u,v}} H(x, y).$$

Step 4: Re-normalize to ensure a valid distribution.
Compute $Z_G \leftarrow \sum_{j=1}^{N_v} G_j$.
If $Z_G = 0$, set $G_j \leftarrow \frac{1}{N_v}$ for all j. Otherwise:

$$G_j \leftarrow \frac{G_j}{Z_G} \quad \forall j.$$

Return G.

Algorithm 2: Training Loop with Gaze Regularization

Input: Policy π_θ (VLA model),
Gaze prediction model ϕ_{gaze},
Dataset $\mathcal{D}$ of episodes $\{(I_{1:n,t}, \ell_t, q_t, A_t^*)\}$,
Temporal window size T for gaze aggregation,
Regularization scale λ,
Output: Trained parameters θ^*.

Initialize model parameters θ and optimizer state.
Repeat for each training step:

1. Sample a batch of timesteps and episodes from $\mathcal{D}$:

$$\{(I_{1:n,t}, \ell_t, q_t, A_t^*)\}_{b=1}^{B}.$$

2. **Compute synthetic gaze heatmaps.**
For each view $i \in \{1, \ldots, n\}$ and each example in the batch, construct a temporal window of frames:

$$\{I_{i,t-T}, \ldots, I_{i,t}, \ldots, I_{i,t+T}\}.$$

Pass this sequence through the GLC gaze model:

$$[H_{i,t-T}, \ldots, H_{i,t}] \leftarrow \phi_{\text{gaze}}(\{I_{i,t-T}, \ldots, I_{i,t}\}).$$

3. **Temporal aggregation of gaze.**
Aggregate the per-frame heatmaps around time t using a weighted average:

$$\tilde{H}_{i,t} = \sum_{\delta=-T}^{T} w_\delta H_{i,t+\delta}, \quad \sum_{\delta=-T}^{T} w_\delta = 1.$$

This yields a temporally smoothed gaze heatmap per view and frame.
4. Convert the aggregated heatmap $\tilde{H}_{i,t}$ into a patch-level distribution ($G_{i,t}$)
5. Feed the multimodal observation into the VLA model:

$$A_t = \pi_\theta(I_{1:n,t}, \ell_t, q_t),$$

obtaining predicted action sequences A_t.

$$S_t = \{S_{i,t}\}_{i=1}^{n},$$

where $S_{i,t} \in \mathbb{R}^{N_v}$ is the spatial attention over visual tokens for view i.
6. For each batch element and each view, compute the KL divergence between the gaze prior and the model attention.

Until convergence or maximum training steps.
Return θ^*.

Analysis of synthetic gaze quality. We discuss the properties and reliability of the synthetic gaze used in our experiments, motivated by the constraints of existing robotic

datasets. Comparisons against alternative gaze priors (e.g., uniform distributions or weaker gaze models) show that performance gains are tied to the *structure* and *quality* of the gaze signal, rather than to generic regularization alone.

Generalization and robustness experiments. We extend the evaluation to settings that more closely resemble real-world deployment: (i) linguistic perturbations that alter the phrasing of task instructions, and (ii) cross-viewpoint degradation where one camera input is removed. These experiments demonstrate that gaze-regularized models maintain stronger performance under both language and viewpoint perturbations, highlighting improved robustness and cross-view spatial coherence.

Reproducibility and implementation transparency. Finally, we provide pseudocode for the heatmap-to-token projection and for the full training loop with gaze regularization, along with additional implementation notes. These details are intended to make it straightforward to reproduce our results and to adapt the proposed regularization strategy to other VLA architectures and datasets.

Together, these additions reinforce the central message of the our work that incorporating gaze-derived supervisory signals and human priors into VLA training not only improves task performance under standard conditions but also leads to more interpretable, better grounded, and more robust robotic manipulation policies.

Discussion and Limitations Our work presents a simple, modular, and architecture-agnostic strategy for improving action prediction in VLA models by incorporating a human-inspired gaze prior during training. The method requires no modification to the underlying VLA design and can be integrated as a lightweight regularization term, making it immediately applicable to a wide range of existing architectures. By guiding the model's spatial attention toward task-relevant regions-mirroring how humans fixate during manipulation-the policy develops more structured visual grounding, sharper and more discriminative attention maps, and ultimately more reliable action prediction. Across a comprehensive set of experiments, we observe consistent improvements over the baseline model, including enhanced robustness under perturbations, degraded viewpoints, and alternative evaluation protocols. These results highlight that gaze provides a compact yet powerful supervisory signal for spatial reasoning in multimodal transformers. Furthermore, our quantitative and qualitative analyses demonstrate a clear link between sharper attention distributions and improved downstream task success, reinforcing the interpretability of our approach.

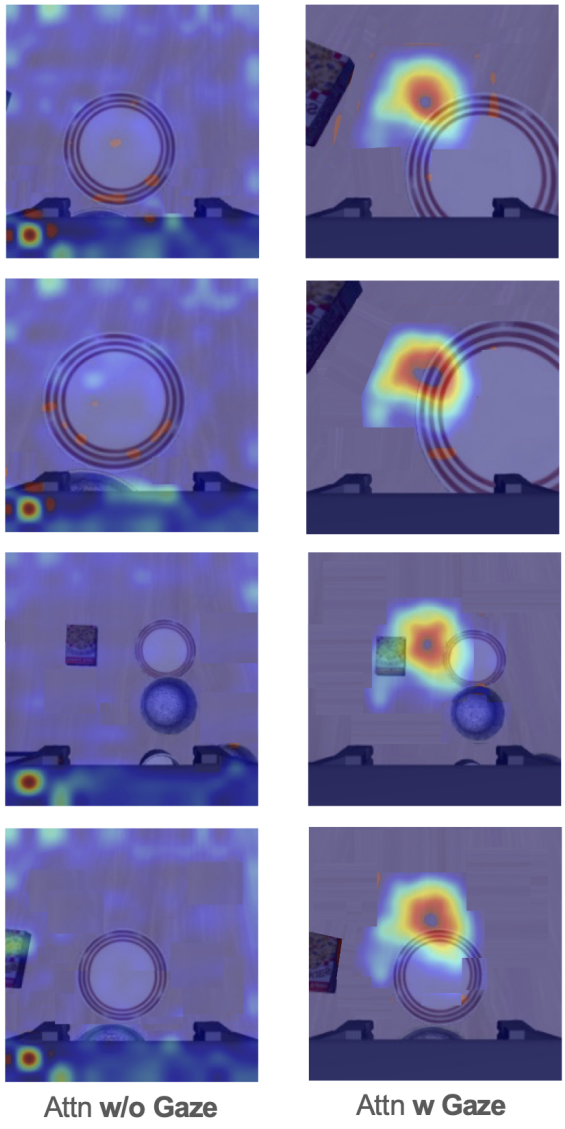

Figure 12. **Attention Comparison.** The baseline model displays diffuse attention spread across the scene, with a single sharp point that is largely task-irrelevant. In contrast, the gaze-regularized model produces noticeably sharper, more concentrated, and consistently task-relevant attention, leading to clearer visual grounding for the instructed action.

While promising, our method also opens several avenues for future refinement. First, the synthetic gaze model used in our experiments-though effective-remains an approximation of real human fixation behavior. A more advanced predictor, or one trained directly on teleoperated demonstrations with ground-truth eye-tracking, could further elevate the quality and temporal precision of gaze heatmaps, strengthening the supervisory signal. Second, our framework currently focuses on RGB-based multi-view

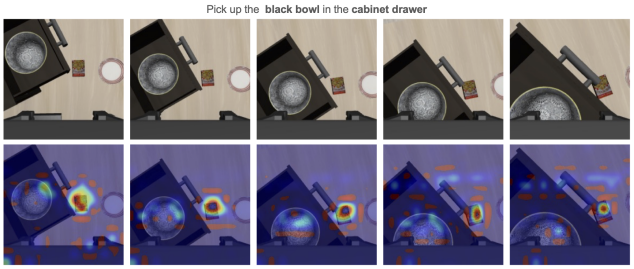

Figure 13. **Visualisation during a failure case.** In this figure, we provide a visualisation of attention during a specific case of failure, where it can be seen that even though the task is to pick up the bowl, attention is not properly distributed on the bowl but rather than on the cabinet handle. Such cases can be mitigated using a better predictor or using a model trained with human supervision on simulated videos

perception; extending gaze regularization to richer modalities such as depth, point clouds, or tactile signals may offer additional benefits, particularly in tasks with complex geometry or occlusions. Third, although our approach is inference-free and directly compatible with real-world deployment, we have not yet evaluated it on a physical robot. A hardware implementation would provide valuable insight into how gaze-aligned attention behaves under real-world variations, including lighting changes, hand occlusions, and workspace clutter. Finally, the interaction between gaze priors and large-scale pretraining remains an open question: future work could explore how gaze can be integrated into foundation-model pretraining pipelines or combined with other forms of human supervision, such as demonstrations or language rationales.

Overall, our findings illustrate that gaze offers a powerful, interpretable, and low-cost source of inductive bias for VLA training. While there is room for further improvement-especially in gaze quality, multimodal integration, and real-world evaluation-our framework represents a meaningful step toward more perceptually grounded, human-aligned, and robust robotic manipulation policies.

G. LLM Usage

We acknowledge the use of LLM in our work for sentence-level re-writing occasionally in our paper to improve the readability, and for suggestions about synonyms, word usage and how to structure and arrange the sections and to check for any spelling/typing mistakes. This was done using ChatGPT and DeepSeek.