OBJECTBINDER-VLA: PERMUTATION-CONSISTENT MULTI-VIEW GROUNDING AND INTENT-CONDITIONED DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language-action (VLA) policies commonly condition control on global visual–language representations that do not explicitly encode task-relevant object support, visibility, multiplicity, or cross-view correspondence. We introduce OBJECTBINDER-VLA, a predictive multi-view object-state interface that preserves global VLA context while exposing a compact, measurable state for control. For each task-conditioned slot and view, a shared Object Binder predicts a categorical distribution over 64 image patches and a NO-OBJECT state, together with independent visibility. A fine-tuned SAM3.1 teacher provides training-only pseudo masks; Hungarian set matching avoids arbitrary identities for exchangeable instances, and teacher-aware duplicate regularization discourages slot collapse. An Intent Query and Object Dynamics Model predict this state at $t + 16$ on LIBERO and $t + 49$ on RoboTwin. Detached future distributions are projected into view–slot tokens for a flow-matching action expert. Stop-gradient boundaries isolate direct updates to the task-specific perception and prediction modules, while the shared Qwen backbone remains jointly optimized. Held-out diagnostics demonstrate spatial grounding and visibility prediction, with no failures across 100 target-order checks of the set objective. The final 50K model achieves 98.10% success (1,962/2,000) on Original LIBERO. These results indicate that explicit object binding and intent-conditioned dynamics can provide auditable intermediate states without replacing global policy context.

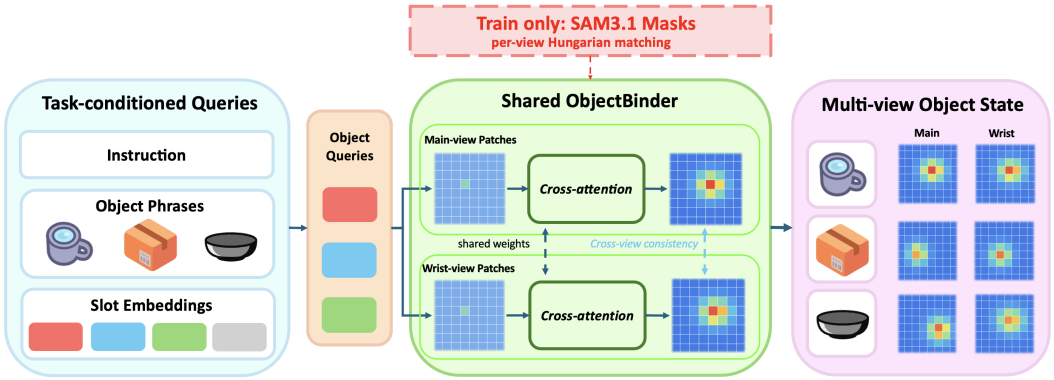


Figure 1: **Task-conditioned multi-view ObjectBinder.** The instruction, object phrases, and learned slot embeddings define a compact set of object queries (left). A shared cross-attention Binder decodes each query independently against main- and wrist-view patch memories, while cross-view consistency relates the resulting view-specific contents (center). Fine-tuned SAM3.1 masks supervise independent per-view Hungarian matching only during training. The output retains, for every valid object and view, explicit spatial support, visibility, and NO-OBJECT probability (right), providing the auditable current representation used by the downstream dynamics model.

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

1 INTRODUCTION

Vision-language-action (VLA) policies adapt pretrained visual–language representations for robot control (Brohan et al., 2023; Kim et al., 2024). Although their dense token sequences provide effective control inputs, they do not explicitly represent the presence, image support, visibility, cross-view correspondence, or expected evolution of task-relevant objects. These variables are particularly important when an object is salient in the main view but small or occluded in the wrist view, temporarily leaves one view, or belongs to a set of visually indistinguishable instances without a canonical ordering. Consequently, policy success does not by itself establish spatial grounding, multiplicity awareness, cross-view consistency, or predictive object dynamics.

Object-centric policies represent proposals, entities, or relations as compact token sets (Zhu et al., 2023b;a; Chapin et al., 2026; Hanyu et al., 2025). Replacing the dense visual sequence with a fixed slot budget can restrict downstream information as scene complexity increases, while latent future features are difficult to assess at the object level. We instead seek a measurable current-and-future object interface that retains the global VLA representation and requires no segmentation at deployment.

We propose OBJECTBINDER-VLA, in which the Object Binder and ODM are jointly designed as a predictive object-state interface. The Binder (Figure 1) exposes the current multi-view object state by predicting, for each slot and view, a categorical distribution over 64 patches and NO-OBJECT, together with independent visibility. The ODM predicts its intent-conditioned evolution at $t + 16$ on LIBERO and $t + 49$ on RoboTwin, the end of its 50-frame action window. Including absence in the normalized selection space precludes an all-zero categorical output and gives absent objects a valid target. Both states augment, rather than replace, the global VLA representation.

Repeated instances have no canonical slot identities. We therefore match predictions to pseudo masks from a fine-tuned SAM3.1 teacher (Meta Superintelligence Labs, 2026) independently in each frame and view using the Hungarian algorithm. The resulting objective is invariant to the ordering of exchangeable targets and does not require a canonical instance-to-slot assignment. A teacher-aware duplicate term penalizes slot collapse only when the teacher multiplicity does not justify the overlap. SAM3.1 supplies training targets but is absent at inference.

The resulting categorical object state is compact and directly inspectable: for LIBERO, $P_t \in \mathbb{R}^{B\times 2\times 4\times 65}$ is normalized along its final dimension, where the first 64 entries form an 8×8 spatial grid and the last denotes NO-OBJECT. The independently supervised visibility variable gates the current slot content but is not part of P_t. Visibility-gated slot features are projected into eight object tokens and appended to the original Qwen hidden states. The policy therefore retains global scene context rather than depending exclusively on the explicit slot set, while the added representation supports direct evaluation of localization, visibility, multiplicity, and duplicate-slot behavior.

The ODM extends static grounding to intent-conditioned object prediction. A state-conditioned Intent Query attends to the current visual–language representation and produces per-view, per-slot intent; the ODM combines it with the current Binder state to predict the categorical distribution at the benchmark-specific horizon, without predicting future visibility. This prediction retains the Binder's patch/NO-OBJECT semantics and is supervised in the same object-level coordinates before projection into future object tokens. Current-frame pseudo-mask losses optimize the Binder, future-state supervision trains the Intent Query and ODM, and detached predictions condition the action expert, preventing the action objective from updating the ODM through its predictions.

The selected Binder achieves 95.63% point recall and 93.64% visibility F1 on a fixed held-out set, with no observed failures across 100 target-order checks or 31 audited duplicate-instance cases. The Binder–ODM combination yields the strongest matched variant at 97.55%, and the final 50K model reaches 98.10% on Original LIBERO.

Our contributions are:

- We jointly design the Object Binder and ODM as a predictive object-state interface that retains global VLA context while exposing current grounding and intent-conditioned evolution in a shared, measurable representation.

- We develop a multi-view Object Binder with patch/NO-OBJECT prediction, independent visibility, Hungarian supervision, and teacher-aware duplicate regularization for exchangeable instances, together with cross-view consistency for identity-valid singleton objects.
- We formulate an ODM in the same categorical state space and condition control on object tokens projected from detached future distributions, enabling dynamics evaluation independently of policy success. The final 50K model obtains 98.10% on Original LIBERO.

2 RELATED WORK

Vision-language-action policies. RT-2 maps pretrained visual–language representations to robot actions (Brohan et al., 2023); OpenVLA introduces an open-source VLA trained on large-scale robot data (Kim et al., 2024); and generalist policy families such as Octo and π_0 target transfer across embodiments and tasks (Ghosh et al., 2024; Black et al., 2024). These methods primarily represent observations through dense visual or visual–language token sequences. OBJECTBINDER-VLA preserves such a global sequence but additionally defines a supervised object state with explicit spatial support, visibility, and absence.

Object-centric representation learning. Slot Attention learns exchangeable object slots through iterative competition (Locatello et al., 2020), while Deep Sets characterizes permutation-invariant set functions (Zaheer et al., 2017). Our setting differs in that the slots are task-conditioned, view-specific, and trained for control using pseudo-mask targets. Because the Binder contains learned slot addresses, we do not claim strict end-to-end permutation equivariance; instead, Hungarian matching makes supervision invariant to the ordering of exchangeable targets.

Object-centric robotic policies. Object structure has been incorporated through detector proposals, explicit 3D representations, and task-conditioned slots. VIOLA conditions control on proposals (Zhu et al., 2023b); GROOT builds object-centric 3D features and segmentation-based correspondences (Zhu et al., 2023a); and STORM learns task-aware slots over frozen foundation-model features (Chapin et al., 2026). Most closely related, SlotVLA supervises propagated slots with boxes, objectness, masks, and Hungarian assignment, then augments selected slots with relation tokens (Hanyu et al., 2025). It prioritizes representation efficiency by replacing dense visual tokens with a fixed-budget object–relation representation, and attributes degradation on complex Spatial and Long tasks partly to that retained-slot budget.

In contrast, OBJECTBINDER-VLA defines each view-specific slot as a categorical distribution over patches and NO-OBJECT with independent visibility, explicitly treats exchangeable multiplicity and duplicate collapse, and appends object tokens to the original global sequence. The explicit object set can therefore be supervised and evaluated without becoming the policy's sole visual input or limiting the global representation to a fixed slot budget. This objective differs from SlotVLA's relation-token compression and GROOT's 3D reconstruction.

World and object dynamics. Object-centric dynamics models discover, track, reconstruct, or predict persistent entities for model-predictive control (Xu et al., 2021). LaWAM predicts a latent action, decodes it into a future feature in a frozen visual space, and conditions the action expert on that subgoal (Chen et al., 2026a). Our ODM instead predicts per-view, per-slot spatial support and NO-OBJECT in the Binder's explicit state space. Localization, absence, and change can therefore be evaluated independently of action success, while the prediction remains directly usable as policy context.

OA-WAM constructs SAM3–DINOv3 object content and pose and maintains temporal identity through an address initialized from language and the first observation; its world and action heads jointly predict the next slot state and action chunk (Liu et al., 2026). Our formulation instead applies per-frame Hungarian supervision, assigns no permanent identity within an exchangeable group, and represents view-dependent absence through NO-OBJECT. OA-WAM consequently emphasizes persistent identity and content-rich states, whereas our ODM models intent-conditioned transitions between permutation-consistent, view-specific states with fixed observable semantics. Stop-gradient boundaries block direct updates to the task-specific Binder and ODM parameters from downstream objectives, while the shared Qwen backbone remains jointly optimized.

Dense visual geometry and multi-view consistency. DINOv3 provides generic, self-supervised dense visual features (Siméoni et al., 2025), while Depth Anything 3 (DA3) is trained to recover spatially consistent geometry from arbitrary views (Lin et al., 2025). We use their frozen intermediate patch features exclusively as matched object-content ablations: identical Binder attention pools either backbone before projection into object tokens. Accordingly, DINOv3 is not treated as an object model, and an intermediate DA3 feature is not interpreted as an explicit depth state.

ConsisVLA-4D selects instruction-relevant patches, fuses multi-view semantic and geometric features, and trains masked future dynamic-object and global-depth decoders (Li et al., 2026). These predictive targets shape the representation during training but are not decoded as an explicit object state for policy conditioning at inference. OBJECTBINDER-VLA instead applies the ODM to the same view-specific state space used by the Binder. ConsisVLA thus emphasizes geometric fusion and predictive representation learning, whereas our formulation emphasizes an explicit, permutation-consistent object-state interface.

Set prediction and segmentation supervision. DETR combines bipartite matching with a null class to train fixed-size unordered predictions (Carion et al., 2020). We adapt this set-prediction formulation to view-conditioned patch supports, while separately supervising visibility and penalizing duplicates in a manner that accounts for the teacher's instance multiplicity. SAM3.1 provides open-vocabulary instance masks that can serve as dense pseudo labels without running the segmentation teacher at policy inference time (Meta Superintelligence Labs, 2026).

3 OBJECTBINDER-VLA

3.1 OVERVIEW AND VISUAL–LANGUAGE INPUTS

Figure 2 summarizes the complete training and inference paths of OBJECTBINDER-VLA, including current object binding, structured intent prediction, object dynamics, and action generation. At time t, the policy observes an instruction x and synchronized RGB images $I_t^{1:V}$. LIBERO uses $V = 2$ cameras (main and wrist), whereas RoboTwin uses $V = 3$ (head, left wrist, and right wrist). Qwen3-VL (Qwen Team, 2025) produces $H_t = \mathrm{Qwen}(x, I_t^{1:V}) \in \mathbb{R}^{B\times L\times 2048}$, which is retained as global policy context. Image-token indices and grid metadata partition the visual states by camera into $X_{t,v} \in \mathbb{R}^{B\times N_p\times 2048}$, with $N_p = 64$ patches per view. The pooled textual context h_t^x averages valid non-image tokens.

We use $K = 4$ object slots. A deterministic, non-neural parser maps the benchmark instruction to at most four object phrases $\phi_{1:K}$ and returns a slot-validity mask $m \in \{0, 1\}^{B\times K}$ together with exchangeability groups. Each phrase representation $e(\phi_k)$ is the mean of its Qwen token embeddings. SAM3.1 masks and teacher instance assignments are used only to define training targets; the inference-time input path contains only the instruction and current observations. Appendix A specifies the teacher data and parsing rules.

3.2 TASK-CONDITIONED MULTI-VIEW OBJECT BINDER

The query for slot k combines a learned slot address with instruction- and phrase-conditioned terms:

$$q_{t,k} = a_k + W_x\,\mathrm{LN}(h_t^x) + W_\phi\,\mathrm{LN}(e(\phi_k)) \in \mathbb{R}^{d_s}, \qquad d_s = 512. \tag{1}$$

Here a_k indexes a representational slot rather than a persistent physical identity. For each view, the patch memory is $M_{t,v,p} = W_m\,\mathrm{LN}(X_{t,v,p}) + e_v$, augmented with a learned null key. The query is decoded separately against the main- and wrist-view memories by a shared two-layer, eight-head cross-attention module, producing view-specific content $C_{t,v,k} \in \mathbb{R}^{d_s}$.

The final-layer attention weights, averaged over the eight attention heads by the multi-head attention operator, parameterize a categorical distribution over the N_p image patches and the null key:

$$P_{t,v,k} = \left[A_{t,v,k,1}, \ldots, A_{t,v,k,N_p}, p^{\emptyset}_{t,v,k}\right], \qquad \sum_{p=1}^{N_p} A_{t,v,k,p} + p^{\emptyset}_{t,v,k} = 1. \tag{2}$$

Here $p^{\emptyset}$ denotes NO-OBJECT. We use "categorical object state" specifically for P_t. Visibility is a separate supervised variable, $s_{t,v,k} = \sigma(f_{\mathrm{vis}}([C_{t,v,k}; q_{t,k}]))$, used to gate current object content. No

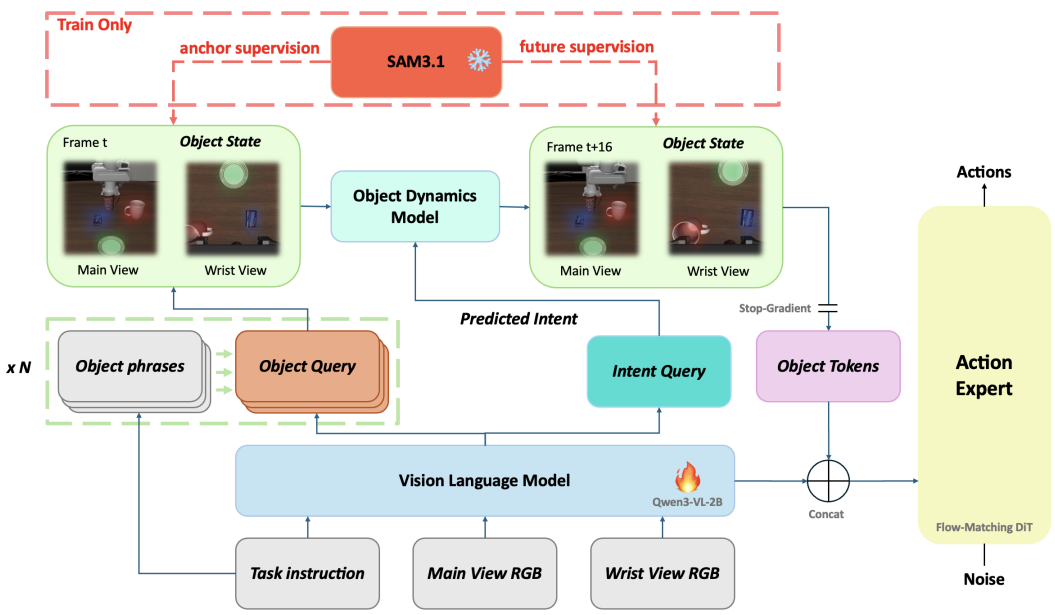


Figure 2: Overview of OBJECTBINDER-VLA with the Object Dynamics Model (ODM). Given the instruction and synchronized main- and wrist-view observations, Qwen3-VL produces global context and task-conditioned queries. The Object Binder represents the current scene, and the Intent Query and ODM predict its categorical state. The illustrated LIBERO path uses horizon $h = 16$; the frame-$t + 16$ panel overlays this prediction on future RGB solely for visualization; future RGB is never used as an inference input. Detached predictions are projected into eight tokens and concatenated, not added, to the original Qwen sequence before the flow-matching action expert. SAM3.1 provides current and future supervision only during training.

canonical fusion is applied across cameras; $P_t \in \mathbb{R}^{B\times V\times K\times(N_p+1)}$ therefore retains view-specific evidence, including cases in which an object is visible in one view but absent from another.

3.3 PERMUTATION-CONSISTENT PSEUDO-MASK SUPERVISION

During training, a fine-tuned SAM3.1 model provides instance masks and per-view visibility labels. A visible mask is resized to the Qwen grid and normalized to unit mass, giving target $Y_{t,v,j} = [\widetilde{M}_{t,v,j}; 0]$; an invisible instance uses $Y_{t,v,j} = [\mathbf{0}; 1]$. The cache also stores valid slots, phrases, multiplicities, and exchangeability groups. Padding slots are excluded from all losses and action tokens, except that visibility BCE supervises them as invisible.

Instance order within an exchangeability group is unspecified. We therefore match predicted slots to teacher instances independently for each frame, view, and group:

$$c_{ij} = \lambda_{\text{sel}} C_{\text{sel}}(P_{t,v,i}, Y_{t,v,j}) + \lambda_{\text{iou}} C_{\text{iou}}(A_{t,v,i}, \widetilde{M}_{t,v,j}) + \lambda_{\text{vis}}\,\text{BCE}(s_{t,v,i}, y^{\text{vis}}_{t,v,j}),$$
$$\pi^*_{t,v,g} = \arg\min_{\pi\in\text{Perm}(g)} \sum_{i\in g} c_{i,\pi(i)}. \tag{3}$$

Here C_{sel} is distributional cross entropy for a visible target and $-\log p^{\emptyset}$ otherwise; the IoU term applies only to visible instances. The three matching terms use equal coefficients. Non-exchangeable objects preserve their phrase-defined slot assignment. Per-view matching for exchangeable groups avoids imposing an arbitrary instance ordering across cameras.

The Binder anchor objective is

$$\mathcal{L}_{\text{bind}} = \lambda_s \mathcal{L}_{\text{select}} + \lambda_v \mathcal{L}_{\text{visibility}} + \lambda_d \mathcal{L}_{\text{duplicate}} + \lambda_c \mathcal{L}_{\text{crossview}}. \tag{4}$$

The selection term is cross entropy over matched (N_p+1)-way distributions, and visibility averages positive and negative BCE separately. The duplicate term combines student-attention overlap with

teacher-exclusive foreign mass. Cross-view InfoNCE uses identity-valid singletons visible in both views and excludes same-phrase cross-episode negatives. Exact coefficients and temperature are reported in Appendix B.

3.4 OBJECT TOKENS AND ACTION CONDITIONING

The base model derives object tokens from the Binder's cross-attended content and appends them to the complete Qwen sequence:

$$T^{\mathrm{obj}}_{t,v,k} = W_o(s_{t,v,k}\,\mathrm{LN}(C_{t,v,k}))\,, \qquad Z_t = [H_t; T^{\mathrm{obj}}_t], \tag{5}$$

Here $W_o : \mathbb{R}^{512} \to \mathbb{R}^{2048}$ is bias-free; invalid-slot tokens are zeroed, and all VK view-specific tokens are retained without set pooling (eight for LIBERO and twelve for RoboTwin). The 16-layer DiT action head is trained by conditional flow matching. For normalized action chunk a, Gaussian noise ϵ, and a Beta-sampled interpolation time τ, it observes $x_\tau = (1-\tau)\epsilon + \tau a$ and regresses the velocity target $a - \epsilon$ with mean-squared error. At inference, the learned velocity field is integrated from Gaussian noise using four uniform explicit-Euler steps. The head predicts an eight-step, 7-DoF action chunk on LIBERO and a 50-step, 14-DoF bimanual chunk on RoboTwin. The token sequence follows the input slot order and remains differentiable, so the base Binder receives both anchor and action gradients. SAM3.1 is not required at inference.

3.5 OBJECT DYNAMICS MODEL

Let $h_{\mathcal{D}} = 16$ for LIBERO and 49 for RoboTwin, where the latter indexes the last frame of the 50-frame window $t{:}t+49$, and define $t_b^+ = \min(t + h_{\mathcal{D}}, T_b - 1)$. The current-conditioned ODM predicts $P_{t_b^+}$ from time-t information only, without future images, actions, or inverse-dynamics latents; it therefore models the policy-implied future state rather than action-conditioned physical dynamics. With sg denoting stop-gradient, the Intent Query uses the detached current Binder state to form per-view, per-slot queries over the current Qwen sequence:

$$\begin{aligned} r_{t,v,k} &= E_p(\mathrm{sg}(P_{t,v,k})) + q^L + e_v, \\ L_{t,v,k} &= f_L(r_{t,v,k} + \mathrm{MHA}(r_{t,v,k}, W_H H_t, W_H H_t)) \in \mathbb{R}^{32}. \end{aligned} \tag{6}$$

The shared query q^L introduces no slot identity; consequently, the ODM is permutation-equivariant with respect to permutations of its slot inputs. This property is local to the ODM and is not a claim of strict equivariance for the complete Binder pipeline. A four-layer, eight-head Transformer then processes the VK masked view–slot inputs $U_{t,v,k} = E_p(\mathrm{sg}(P_{t,v,k})) + W_L L_{t,v,k} + e_v$ and predicts

$$\widehat{P}_{t^+} = \mathrm{softmax}(W_F\,\mathrm{ODM}(U_t; m)) \in \mathbb{R}^{B\times V\times K\times(N_p+1)}. \tag{7}$$

Future SAM3.1 targets use the same patch/NO-OBJECT space. In the evaluated LIBERO schema, exchangeable pairs are matched independently per view, and $\mathcal{L}_{\mathrm{future}}$ is the mean categorical cross entropy over valid matched slots. The ODM predicts P_{t^+} only; it does not predict the auxiliary visibility variable s_{t^+}.

In the evaluated ODM variant, each predicted distribution is detached and projected to a 2048-dimensional token:

$$T^{\mathrm{future}}_{t,v,k} = m_k f_{\mathrm{proj}}(\mathrm{sg}(\widehat{P}_{t^+,v,k})), \qquad Z^{\mathrm{ODM}}_t = [H_t; T^{\mathrm{future}}_t]. \tag{8}$$

The resulting policy input retains the global Qwen sequence and replaces the base model's current object tokens with VK predicted-future tokens (eight in LIBERO). The projection head is optimized by $\mathcal{L}_{\mathrm{act}}$; the stop-gradient prevents this loss from updating the latent predictor or ODM. The training objective is

$$\mathcal{L}_{\mathrm{ODM}} = \mathcal{L}_{\mathrm{act}} + \lambda_f \mathcal{L}_{\mathrm{future}} + \lambda_b \mathcal{L}_{\mathrm{bind}}. \tag{9}$$

For the evaluated ODM, $\lambda_f = 0.015$ and $\lambda_b = 0.32$; the relative weights inside $\mathcal{L}_{\mathrm{bind}}$ are unchanged.

The three objectives have disjoint direct update paths to task-specific modules. In the ODM variant, the Binder is optimized only by $\mathcal{L}_{\mathrm{bind}}$, because P_t is detached and current Binder tokens are not supplied to the action expert. The latent predictor and ODM are optimized by $\mathcal{L}_{\mathrm{future}}$, whereas the future-token projector and action expert are optimized by $\mathcal{L}_{\mathrm{act}}$. Qwen is shared across the three paths and therefore receives gradients from all three objectives. Thus, the stop-gradient boundaries isolate direct updates to task-specific modules but do not isolate the shared visual–language representation.

4 EXPERIMENTAL SETUP

We evaluate on Original LIBERO (Liu et al., 2023), LIBERO-Plus (Fei et al., 2026), and six RoboTwin 2.0 tasks (Mu et al., 2025; Chen et al., 2025a), fully fine-tuning the official Qwen3-VL-2B-Instruct weights. Primary runs use $50{,}000 \times 64 = 3.2$M samples (four GPUs, batch 16/GPU). The matched 30K ablations hold the data, initialization, action expert, batch size, and 2,000-episode evaluation protocol fixed. A separate 3,200-step Binder pretraining stage is shared by Binder-bearing variants. Table 5 in Appendix A gives the complete settings.

5 RESULTS

5.1 PRIMARY RESULTS ON ORIGINAL LIBERO

Original LIBERO is our primary comparison because it provides complete four-suite task coverage. Table 1 compares the final 50K checkpoint with representative VLA policies and world-model-augmented policies. The two model families are separated to make their different prediction objectives explicit. All results for our model are obtained from the complete 2,000-episode evaluation with 500 episodes per suite.

Table 1: Original LIBERO success rate (%). Representative published methods are grouped into conventional VLA policies and world-model-augmented policies/WAMs. OBJECTBINDER-VLA is evaluated for 500 episodes per suite. Where the numerical source differs from the method paper, Octo follows OpenVLA (Kim et al., 2024); π_0 and π_0-FAST follow OpenVLA-OFT (Kim et al., 2025); and GR00T-N1 and $\pi_{0.5}$ follow VLA-JEPA (Sun et al., 2026). † denotes the VLA-JEPA ablation without human-video pretraining, rather than the full model. Bold denotes the best result among the methods shown in each column.

Model	Spatial	Object	Goal	LIBERO-10	Overall
Vision–language–action policies					
SlotVLA (ORC) (Hanyu et al., 2025)	60.0	91.0	86.0	31.0	67.0
TraceVLA (Zheng et al., 2024)	84.6	85.2	75.1	54.1	74.8
Octo (Ghosh et al., 2024)	78.9	85.7	84.6	51.1	75.1
OpenVLA (Kim et al., 2024)	84.7	88.4	79.2	53.7	76.5
SpatialVLA (Qu et al., 2025)	88.2	89.9	78.6	55.5	78.1
π_0-FAST (Pertsch et al., 2025)	96.4	96.8	88.6	60.2	85.5
4D-VLA (Zhang et al., 2025a)	88.9	95.2	90.9	79.1	88.6
GR00T-N1 (NVIDIA et al., 2025)	94.4	97.6	93.0	90.6	93.9
π_0 (Black et al., 2024)	96.8	98.8	95.8	85.2	94.2
UniVLA (Bu et al., 2025)	96.5	96.8	95.6	92.0	95.2
PointACT (Chen et al., 2026b)	97.4	99.6	96.2	90.6	96.0
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	98.8	98.2	**98.0**	92.4	96.9
OpenVLA-OFT (Kim et al., 2025)	97.6	98.4	97.9	94.5	97.1
World-model-augmented policies and world-action models					
WorldVLA (512×512) (Cen et al., 2025)	87.6	96.2	83.4	60.0	81.8
DreamVLA (Zhang et al., 2025b)	97.5	94.0	89.5	89.5	92.6
VLA-JEPA† (Sun et al., 2026)	94.8	99.6	95.8	94.0	96.1
Fast-WAM (Yuan et al., 2026)	98.2	**100.0**	97.0	95.2	97.6
OA-WAM (Liu et al., 2026)	**98.9**	99.0	97.4	95.9	97.8
Ours					
OBJECTBINDER-VLA + ODM (50K)	98.6	99.4	97.4	**97.0**	**98.1**

The 50K model succeeds in 1,962/2,000 episodes (98.10%), the highest overall rate among the representative methods in Table 1. It obtains 98.6%, 99.4%, 97.4%, and 97.0% on Spatial, Object, Goal, and LIBERO-10, respectively, and 90.0% on the exchangeable-instance two-moka task. The matched ablation supports compatibility between predicted-state conditioning and strong control, but does not isolate the effect of 50K training.

5.2 ROBUSTNESS ON LIBERO-PLUS

We evaluate the same 50K checkpoint, without LIBERO-Plus fine-tuning, on all 10,030 episodes spanning seven controlled distribution shifts. Following the aggregation convention used for the compared methods, Table 2 reports the unweighted mean across the seven perturbation axes.

Table 2: LIBERO-Plus success rate (%) across seven perturbation axes. Axis Mean is the unweighted arithmetic mean of the seven displayed axis scores. Published values are transcribed from the comparison in SG-WAM (Zhao et al., 2026a) and the official reports of OpenWAM-α (Wang et al., 2026), 4D-WAM (Yang et al., 2026), and Faster-WAM (Zhao et al., 2026b); Spatial Forcing (Li et al., 2025) and Mantis (Yang et al., 2025) identify the corresponding original methods. † denotes the VLA-JEPA ablation without human-video pretraining. Bold denotes the best value among the representative methods included in the table.

Model	Camera	Robot init.	Lang.	Light	Backgr.	Noise	Layout	Axis Mean
VLA policies								
Spatial Forcing	20.1	13.4	40.9	29.1	33.4	25.7	39.3	28.84
Mantis	15.7	41.8	45.9	45.1	28.9	39.2	62.5	39.87
UniVLA	4.3	50.3	71.8	59.1	80.0	25.3	34.3	46.44
π_0	13.8	6.0	58.8	85.0	81.4	79.0	68.9	56.13
OpenVLA-OFT	56.4	31.9	79.5	88.7	93.3	75.8	74.2	71.40
World-model-augmented policies and world-action models								
WorldVLA	0.1	27.9	41.6	43.7	17.1	10.9	38.0	25.61
Fast-WAM	16.4	44.5	68.9	78.2	53.7	37.7	60.7	51.44
VLA-JEPA†	40.3	55.7	72.9	88.2	70.5	38.2	74.6	62.91
OpenWAM-α	33.8	**76.1**	88.0	**97.0**	87.1	39.8	77.5	71.33
4D-WAM	45.15	64.26	90.63	94.29	57.71	69.08	**79.21**	71.48
SG-WAM	**58.6**	48.9	81.4	89.8	86.1	**80.7**	74.2	74.24
Faster-WAM	53.75	71.57	**94.67**	96.27	61.29	63.57	79.09	74.32
Ours								
OBJECTBINDER-VLA + ODM (50K)	50.84	57.03	89.33	95.01	**94.24**	72.83	75.28	**76.37**

OBJECTBINDER-VLA + ODM obtains an Axis Mean of 76.37%, the highest among the representative methods in Table 2. Across all 10,030 evaluation episodes, the corresponding episode-level micro-average is 74.61% (7,483 successes). Performance is strongest for lighting, background, and language perturbations, whereas camera viewpoint and robot initialization remain the principal failure modes. Suite-level scores are 82.43% Spatial, 76.73% Object, 72.79% Goal, and 66.89% LIBERO-10.

5.3 CROSS-BENCHMARK EVALUATION ON ROBOTWIN

To test transfer beyond LIBERO's single-arm embodiment, we train one policy on six bimanual RoboTwin 2.0 tasks using Clean and Randomized data (Chen et al., 2025a;b). From head and two wrist views, it predicts P_{t+49} and a 50-step, 14-DoF action chunk over $t{:}t+49$. Inference uses four explicit-Euler flow-integration steps; episode- and chunk-derived seeds fix the initial Gaussian noise for deterministic evaluation. Under the same 3.2M-sample-equivalent budget, our policy achieves 81.2% success over 1,200 simulator-oracle episodes, comprising 83.5% on Clean and 78.8% on Randomized data; Table 6 reports per-task results. Table 3 additionally summarizes published RoboTwin 2.0 averages.

Table 3: Average success rate (%) on RoboTwin 2.0. ACT, DP, RDT, π_0, $\pi_{0.5}$, VPP, VPDD, and Video2Act are transcribed from Video2Act (Jia et al., 2025); OpenVLA-OFT and JEPA-VLA are transcribed from JEPA-VLA (Miao et al., 2026). Published results follow the RoboTwin 2.0 evaluation settings reported in the respective papers.

Method	Avg.
ACT	21.9
DP (Chi et al., 2025)	31.2
OpenVLA-OFT	32.1
VPDD	38.5
$\pi_{0.5}$	42.8
π_0	43.2
VPP	44.7
JEPA-VLA	45.6
RDT (Liu et al., 2024)	46.0
Video2Act	54.1
OBJECTBINDER-VLA + ODM	**81.2**

5.4 ABLATION STUDY

Table 4: Ablation study on Original LIBERO (%; 2,000 episodes). All variants are trained for 30K steps under the matched protocol. Differences below one percentage point from a single evaluation should not be interpreted as statistically conclusive.

Model	Spatial	Object	Goal	LIBERO-10	Overall
VLA baseline	98.6	98.6	97.4	94.4	97.25
ObjectBinder-VLA	97.4	99.0	98.4	94.6	97.35
ObjectBinder-VLA + ODM	99.0	99.4	97.4	94.4	**97.55**
Attention-DINO	97.4	99.4	97.0	94.0	96.95
Attention-DA3	98.4	98.6	96.8	94.0	96.95

ObjectBinder-VLA differs from the baseline by 0.10 percentage points, whereas the ODM variant attains the highest value in this matched ablation. Attention-DINO and Attention-DA3 pool frozen semantic or geometry-aware features with current Binder attention and project eight tokens into the Qwen sequence (Siméoni et al., 2025; Lin et al., 2025). Both attain 96.95%. Under this interface, the results favor predicted dynamics tokens over the evaluated frozen-feature tokens; they do not establish that either backbone is intrinsically unsuitable.

5.5 OBJECT-STATE QUALITY

On the fixed held-out set, the selected Binder obtains 95.63% point recall, 93.64% visibility F1, and 88.76% NO-OBJECT accuracy. The audit records no failures across 31 duplicate-instance cases or 100 target-order checks of the set-matched objective. Small wrist-view objects remain the primary limitation (56.25% point recall); complete metric definitions and qualitative examples appear in Appendix C.

6 DISCUSSION AND CONCLUSION

OBJECTBINDER-VLA provides an explicit multi-view object interface and intent-conditioned dynamics while retaining global VLA features. Its normalized patch/NO-OBJECT state supports auditable localization and avoids an all-zero categorical output; the final policy reaches 98.10% on Original LIBERO. Neither frozen DINOv3 nor DA3 content improves the matched 30K result, although this comparison does not identify which representational or optimization factor accounts for the outcome. Limitations include the small wrist-view audit subset, inherited SAM3.1 errors, and the absence of persistent 3D identity; relation-aware fusion remains future work.

Reproducibility Statement

The model architecture, optimization objectives, gradient boundaries, and training and evaluation protocols are documented in the main text and appendices. Appendix A reports the complete implementation and training configuration, while Appendix B specifies the future-state objective and gradient isolation. Code, experiment configurations, and evaluation scripts will be included in the supplementary material.

AI Assistance Disclosure

Generative AI tools were used to aid or polish writing. All technical content, experimental results, analyses, and claims were reviewed and verified by the authors.

References

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, 2023.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. UniVLA: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 2020.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. WorldVLA: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

Alexandre Chapin, Emmanuel Dellandréa, and Liming Chen. STORM: Slot-based task-aware object-centric representation for robotic manipulation. *arXiv preprint arXiv:2601.20381*, 2026.

Jialei Chen, Kai Wang, Kang Chen, Shuaihang Chen, Feng Gao, Wenhao Tang, Zhiyuan Li, Weilin Liu, Zhuyu Yao, Boxun Li, Yuanbo Xu, and Chao Yu. LaWAM: Latent world action models for efficient dynamics-aware robot policies. *arXiv preprint arXiv:2606.15768*, 2026a.

Shizhe Chen, Paul Pacaud, and Cordelia Schmid. PointACT: Vision-language-action models with multi-scale point-action interaction. *arXiv preprint arXiv:2605.21414*, 2026b.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025a.

Tianxing Chen, Kaixuan Wang, Zhaohui Yang, Yuhao Zhang, Zanxin Chen, Baijun Chen, Wanxi Dong, Ziyuan Liu, et al. Benchmarking generalizable bimanual manipulation: RoboTwin dual-arm collaboration challenge at CVPR 2025 MEIS workshop. *arXiv preprint arXiv:2506.23351*, 2025b.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion Policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10–11), 2025. doi: 10.1177/02783649241273668.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: A progressive robustness benchmark for visual-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38574–38583, June 2026.

Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Ria Doshi, Charles Xu, Jianlan Luo, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Taisei Hanyu, Nhat Chung, Huy Le, Toan Nguyen, Yuki Ikebe, Anthony Gunderman, Duy Nguyen Ho Minh, Khoa Vo, Tung Kieu, Kashu Yamazaki, Chase Rainwater, Anh Nguyen, and Ngan Le. SlotVLA: Towards modeling of object-relation representations in robotic manipulation. *arXiv preprint arXiv:2511.06754*, 2025.

Yueru Jia, Jiaming Liu, Shengbang Liu, Rui Zhou, Wanhe Yu, Yuyang Yan, Xiaowei Chi, Yandong Guo, Boxin Shi, and Shanghang Zhang. Video2Act: A dual-system video diffusion policy with robotic spatio-motional modeling. *arXiv preprint arXiv:2512.03044*, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, et al. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Robotics: Science and Systems*, 2025.

Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. *arXiv preprint arXiv:2510.12276*, 2025.

Wei Li, Jizhihui Liu, Yixing Li, Junwen Tong, Rui Shao, and Liqiang Nie. ConsisVLA-4D: Advancing spatiotemporal consistency in efficient 3d-perception and 4d-reasoning for robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. RDT-1B: A diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

Yushan Liu, Peibo Sun, Shoujie Li, Yifan Xie, Lingfeng Zhang, Xintao Chao, Shiyuan Dong, Fang Chen, Xiao-Ping Zhang, and Wenbo Ding. OA-WAM: Object-addressable world action model for robust robot manipulation. *arXiv preprint arXiv:2605.06481*, 2026.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, 2020.

Meta Superintelligence Labs. SAM 3.1: Object multiplex model checkpoints. `https://huggingface.co/facebook/sam3.1`, 2026. Official model card and release documentation.

Shangchen Miao, Ningya Feng, Jialong Wu, Ye Lin, Xu He, Dong Li, and Mingsheng Long. JEPA-VLA: Video predictive embedding is needed for VLA models. *arXiv preprint arXiv:2602.11832*, 2026.

Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, Lunkai Lin, Zhiqiang Xie, Mingyu Ding, and Ping Luo. RoboTwin: Dual-arm robot benchmark with generative digital twins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27649–27660, 2025.

NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, Jiayuan Gu, Bin Zhao, Dong Wang, and Xuelong Li. SpatialVLA: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. VLA-JEPA: Enhancing vision-language-action model with latent world model. *arXiv preprint arXiv:2602.10098*, 2026.

Yuran Wang, Siqiao Huang, Mingleyang Li, Chenhao Zhang, Jiaqi Liang, Weiyang Jin, Yue Chen, Xuemin Chi, Donghao Zhou, Qize Yu, Yu-Kai Wang, Yuhan Rui, Shenzhe Yao, Zhen Yuan, Zhenhao Shen, Kefei Zhu, Zijie Zhu, Ning Gao, Xiaowei Chi, Guanqi He, Shanghang Zhang, Hao Dong, Lin Shao, and Hang Zhao. OpenWAM: An open, modular exploration towards systematic world-action model pretraining. *arXiv preprint arXiv:2609.07398*, 2026.

Zhenjia Xu, Zhanpeng He, Jiajun Wu, and Shuran Song. Learning 3d dynamic scene representations for robot manipulation. In *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pp. 126–142, 2021.

Lishan Yang, Wenxuan Song, Xi Wang, Pingyue Sheng, Zheng Fang, Ziyang Zhou, Junjie He, Haodong Yan, Jiayi Chen, Nan Sun, Qiao Sun, Pengwei Wang, Lingqiao Liu, Yan Wang, Yuxiang Gao, Feras Dayoub, and Haoang Li. 4D-WAM: Infusing spatiotemporal awareness into world action models through trajectory fields. *arXiv preprint arXiv:2608.08023*, 2026.

Yi Yang, Xueqi Li, Yiyang Chen, Jin Song, Yihan Wang, Zipeng Xiao, Jiadi Su, You Qiaoben, Pengfei Liu, and Zhijie Deng. Mantis: A versatile vision-language-action model with disentangled visual foresight. *arXiv preprint arXiv:2511.16175*, 2025.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. 4D-VLA: Spatiotemporal vision-language-action pretraining with cross-scene calibration. *arXiv preprint arXiv:2506.22242*, 2025a.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, Fan Lu, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. DreamVLA: A vision-language-action model dreamed with comprehensive world knowledge. *arXiv preprint arXiv:2507.04447*, 2025b.

Ruiteng Zhao, Zhengshen Zhang, Yue Su, Wenshuo Wang, Jiahui Li, Zhiyuan Yang, Francis E. H. Tay, Marcelo H. Ang, Jr., and Haiyue Zhu. SG-WAM: Self-guided world modeling in geometry-aware policy space. *arXiv preprint arXiv:2608.01397*, 2026a.

Weiheng Zhao, Haoyi Jiang, Xin Shi, Liu Liu, Fan Huang, Zhizhong Su, Wei Sui, and Xinggang Wang. Faster-WAM: Efficient inference-time future conditioning for robust world action models. *arXiv preprint arXiv:2608.04404*, 2026b.

Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. TraceVLA: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*, 2024.

Yifeng Zhu, Zhenyu Jiang, Peter Stone, and Yuke Zhu. Learning generalizable manipulation policies with object-centric 3d representations. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 3418–3433, 2023a.

Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. VIOLA: Imitation learning for vision-based manipulation with object proposal priors. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 1199–1210, 2023b.

A Implementation Details

Table 5 summarizes the evaluated matched configuration. The Binder checkpoint was selected using a fixed held-out protocol in which duplicate and set-objective target-order failures served as exclusion criteria, followed by point recall and prespecified tie-breakers. Policy success was not used for checkpoint selection.

SAM3.1 teacher. For the LIBERO experiments, the official SAM3.1 multiplex checkpoint is fine-tuned for four epochs on 10,720 simulator-rendered training images with pixel-accurate masks spanning the four Original LIBERO suites and 39 object categories. The checkpoint is selected by segmentation AP on a disjoint 2,824-image validation split. Neither policy-evaluation rollouts nor LIBERO-Plus observations are used to fit the teacher. The selected teacher is used offline to build current- and future-frame caches; it is not loaded by the policy at inference.

Deterministic object parser. The parser is rule-based and contains no learned language model. For the fixed LIBERO vocabulary, an explicit instruction table identifies two phrase groups; each group expands to at most two slots. Plural constructions such as "both moka pots" produce repeated phrases in one exchangeability group, whereas relation-qualified objects such as left and right plates retain distinct groups. Unknown instructions use a conservative noun-phrase template fallback, and unused slots are treated as padding. RoboTwin uses an analogous task-specific rule set for the six evaluated tasks.

B Additional Reproducibility Notes

Padding slots are masked from all losses except visibility BCE, which labels them invisible. Exchangeable instances are aligned per frame and view before selection and future-state losses are evaluated. For matched target Y^* and valid set Ω, the prediction loss is

$$\mathcal{L}_{\text{future}} = -\frac{1}{|\Omega|} \sum_{(b,v,k)\in\Omega} \sum_{c=1}^{65} Y^*_{t_b^+,b,v,k,c} \log \widehat{P}_{t_b^+,b,v,k,c}. \tag{10}$$

Current Binder state is detached at the ODM input, and predicted future distributions are detached before projection into action-conditioning tokens. The DINO attention study uses frozen official

Table 5: Core implementation settings.

Component	Setting
VLM	Qwen3-VL-2B-Instruct; official pretrained weights; full fine-tuning
VLM hidden width	2048
LIBERO views / slots	2 (main, wrist) / 4
RoboTwin views / slots	3 (head, left wrist, right wrist) / 4
Binder	width 512; 2 layers; 8 heads; head-averaged attention
Selection space	8×8 patches + NO-OBJECT
LIBERO action expert	DiT-B; 16 layers; horizon 8; action dim. 7
RoboTwin action expert	DiT-B; 16 layers; horizon 50; action dim. 14
ODM prediction offset	LIBERO: 16 ($t + 16$); RoboTwin: 49 ($t + 49$)
Action inference	4-step explicit-Euler flow integration
Primary training budget	$50{,}000 \times 64 = 3.2$M samples
LIBERO parallelism	4 GPUs; batch 16/GPU; global batch 64
RoboTwin data / budget	3,300 Clean+Randomized trajectories; 3.2M samples
Ablation training	30K steps; 4 GPUs; global batch 64
Optimizer / peak LR	AdamW; Qwen 10^{-5}; action side 10^{-4}
Primary / ablation schedule	5K / 3K warmup; cosine to 50K / 30K
Binder initialization	Separate selected 3,200-step checkpoint
Binder loss weights	select 0.05; visibility 0.02; duplicate 0.05; cross-view 0.05
Cross-view objective	symmetric InfoNCE; temperature 0.1
ODM loss weights	future 0.015; Binder anchor 0.32
Original LIBERO evaluation	500 episodes/suite; 2,000 total
LIBERO-Plus evaluation	10,030 episodes across seven shifts
RoboTwin inference	Deterministic 4-step flow integration; seeded initial noise; action window $t{:}t + 49$
RoboTwin evaluation	100 episodes/task/domain; 1,200 total; simulator-oracle success

Table 6: Per-task success rate (%) for OBJECTBINDER-VLA + ODM on the six evaluated RoboTwin 2.0 tasks. Each task–domain pair contains 100 episodes, with success determined by the simulator task oracle. BR: blocks ranking; DS: place dual shoes; OS: place object on scale; PS: place phone on stand; BD: put bottles in dustbin; SB: stack three blocks.

Domain	BR	DS	OS	PS	BD	SB	Avg.
Clean	90.0	84.0	92.0	93.0	77.0	65.0	83.5
Randomized	84.0	75.0	90.0	97.0	58.0	69.0	78.8
Combined	87.0	79.5	91.0	95.0	67.5	67.0	**81.2**

DINOv3-L/16 features with shape $[B, 2, 256, 1024]$; Binder attention is bilinearly upsampled from 8×8 to 16×16, renormalized, and used to pool each slot before a $1024 \rightarrow 2048$ projection. Training-cache and online inference features were verified to be elementwise identical after FP16 conversion.

C OBJECT-STATE DIAGNOSTICS AND QUALITATIVE RESULTS

The step-3200 Binder was selected by the fixed audit in Table 7, without policy-success feedback; step 4000 reduced overall and small-wrist point recall to 95.13% and 43.75%, respectively. Point recall tests whether the highest-probability spatial patch lies inside the matched visible teacher support. Top-k support IoU selects k patches, where k equals the number of occupied teacher cells. Visibility F1 thresholds $s_{t,v,k}$ at 0.35, while NO-OBJECT accuracy classifies a slot as visible when $p^{\emptyset}_{t,v,k} < 0.5$. Multiplicity point recall restricts point recall to exchangeable slots; the small-wrist subset contains visible wrist-view objects whose cached support occupies at most 2% of the grid. A duplicate failure occurs when two simultaneously visible members of one exchangeable group share

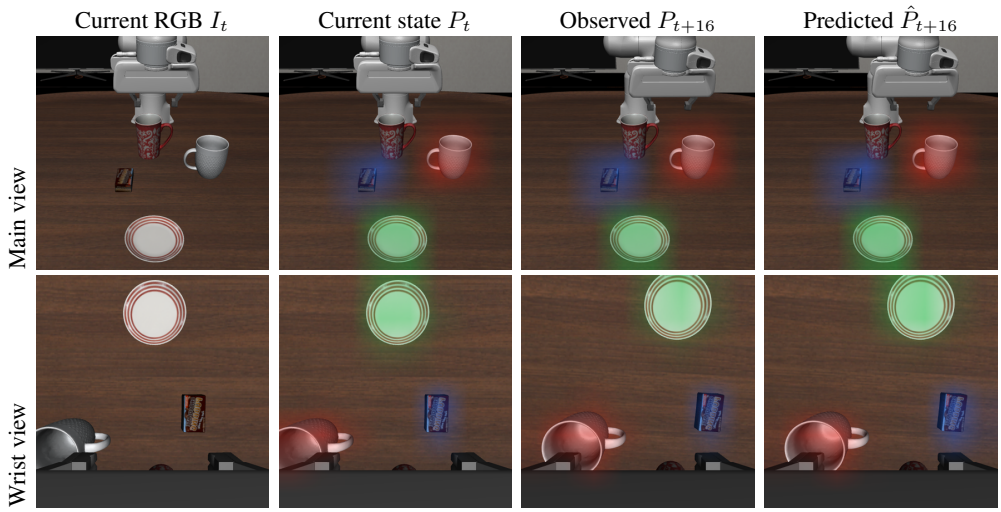


Figure 3: Qualitative binding and prediction on LIBERO-10 task 6, episode 5: "put the white mug on the plate and put the chocolate pudding to the right of the plate." Rows show the main and wrist cameras; columns show the current RGB, current Binder state, SAM3.1-supervised state at horizon 16, and ODM prediction. The colored valid slots for the mug, pudding, and plate remain spatially consistent across both views, and the prediction closely follows the observed future state.

the same spatial argmax. A target-order failure is recorded when reversing an exchangeable teacher group changes the summed set-matching objective by more than 10^{-5}.

Table 7: Fixed held-out Object Binder diagnostics.

Metric	Result
Overall point recall	95.63%
Main-view point recall	96.52%
Wrist-view point recall	94.53%
Top-k support IoU	74.16%
Visibility F1	93.64%
NO-OBJECT accuracy	88.76%
Multiplicity point recall	89.06%
Small-wrist point recall	56.25%
Duplicate failures	0 / 31
Set-objective target-order failures	0 / 100
Trainable Binder parameters	9.74M



Figure 4: Qualitative binding and prediction on the RoboTwin 2.0 Clean `put_bottles_dustbin` task. Rows show the head, left-wrist, and right-wrist cameras; columns show the current RGB observation, current Binder state P_t, and ODM prediction $\widehat{P}_{t+49}$ at the final frame of the 50-frame action window. Colors identify the four task-conditioned slots for the two foreground bottles, the soda bottle, and the dustbin. The displayed effective state combines spatial attention with predicted visibility, thereby suppressing invalid or poorly visible slots. Unlike Figure 3, this visualization does not include the observed future teacher state.