UAM: A DUAL-STREAM PERSPECTIVE ON FORGETTING IN VLA TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language–action (VLA) models typically adapt a pretrained vision–language model (VLM) to robot control to transfer broad multimodal knowledge. Yet action fine-tuning can erode the competence that motivates this initialization, a loss we call the *embodiment tax*. Can a VLA reduce this loss while allowing its semantic backbone to adapt, without auxiliary vision–language co-training? Inspired by the recognition and visuomotor pathways of biological vision, we investigate a representational-bottleneck hypothesis: a semantic pathway that must also supply control-specific visual features may face competing adaptation demands. We study this hypothesis through the Unified Action Model (UAM), which couples a semantic expert, a parallel visual Dorsal Expert, and an action expert using mixture-of-transformers. Starting from aligned understanding–generation pretraining, we compare Dorsal initialization, visual versus query inputs, and visual-dynamics supervision. The strongest tested design pairs generative initialization with future-observation prediction from robot trajectories, providing a complementary learning target while all three experts remain trainable. Under matched initialization and action-training settings, UAM retains 96.7% of pretrained scores across four reported multimodal metrics, versus 54.8% without the Dorsal pathway, while also improving OOD action performance by 35.0% with no auxiliary vision–language replay, VLM freezing, or gradient blocking is used during adaptation. After training on only 3,000 real-robot trajectories, UAM achieves the highest average success among the compared variants on semantic generalization tasks involving unseen objects, novel target compositions, and instruction variation. Together, these results support aligned generative pretraining and visual-dynamics supervision as a means to improve the retention–control trade-off in unified multimodal models. Demos can be found at our anonymous website.

1 INTRODUCTION

Vision–language–action (VLA) models adapt pretrained vision–language models (VLMs) to language-conditioned robot control (Brohan et al., 2023; Kim et al., 2024; Team et al., 2024; Black et al., 2024; Intelligence et al., 2025). The motivation extends beyond learning a mapping from images to low-level actions. Broad vision–language pretraining develops capabilities such as open-vocabulary recognition, attribute and spatial understanding, text recognition, multilingual grounding, and instruction following. Robot demonstrations cover a much narrower range of scenes and language, making it difficult to acquire all of these capabilities from action data alone. A VLA's appeal therefore lies partly in what it inherits: knowledge that may help a policy interpret unfamiliar objects and goals beyond its demonstrated behaviors.

This inheritance can be fragile. Our coupling diagnostics show that action fine-tuning can improve control while degrading pretrained multimodal scores, across the tested backbones and action-head designs (Section 3.1). We call this measured loss the *embodiment tax*: part of the competence for which the VLM was selected is sacrificed as it becomes an action model. We assess retention relative to the same model's pretrained scores, alongside the resulting policy's control performance.

The concern is practical as well as evaluative. A policy may imitate familiar trajectories while becoming less able to recognize a new object, resolve an unfamiliar object–target pairing, or understand a rephrased instruction. Robot trajectories provide narrow supervision for recovering broad knowledge

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

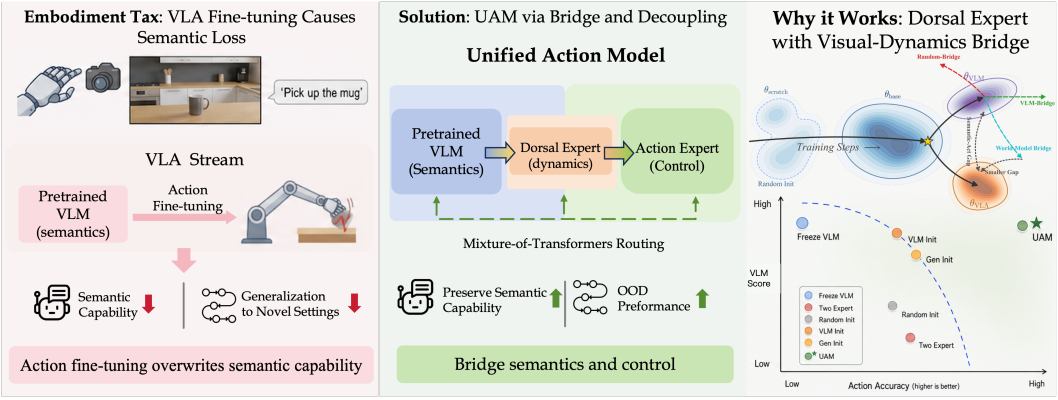


Figure 1: **A dual-stream view of the embodiment tax.** Action adaptation can degrade pretrained multimodal competence. UAM adds a control-oriented visual pathway alongside the semantic VLM. Generative initialization and visual-dynamics supervision equip this pathway with predictive features, improving the retention–control trade-off.

once it is lost. We therefore evaluate both what the semantic model continues to know and whether the action policy can use relevant knowledge under semantic distribution shifts.

Freezing the VLM preserves its outputs by construction, but leaves control dependent on static features that may not fully support the spatial and dynamic demands of manipulation (Zhang et al., 2026). Auxiliary vision–language (VL) co-training supplies an explicit retention signal (Zhou et al., 2025; Intelligence et al., 2025). Knowledge insulation separates gradient paths from newly initialized action modules while allowing other supervision to update the backbone (Driess et al., 2025). These approaches motivate a complementary question: *can the organization of visual representations improve retention while the semantic expert remains trainable, without replaying general-purpose VL data during action adaptation?*

The ventral–dorsal organization of biological vision distinguishes recognition-oriented and visuomotor processing (Ungerleider, 1982; Goodale & Milner, 1992). We use this distinction as a design analogy for a *representational-bottleneck hypothesis*. A semantic pathway must identify the instructed object and also supply features for locating the arm, approaching a grasp, and tracking scene evolution. Separate action-expert parameters do not remove this shared visual dependence: action gradients can still change semantic features. A parallel, visually grounded pathway could supply complementary control-oriented representations and reduce the need to obtain every action-relevant feature by adapting the semantic encoder.

We investigate this hypothesis using the **Unified Action Model (UAM)**, which couples a semantic expert, a visual Dorsal Expert, and an action expert through mixture-of-transformers (MoT). UAM initializes its semantic and generative experts from BAGEL (Deng et al., 2025). The action expert attends to both visual pathways for language-grounded information and interaction features. We treat the design of the Dorsal Expert as an empirical question: what visual prior and learning signal allow it to support control while preserving the semantic pathway's competence? We vary initialization, visual versus query inputs, supervision, and semantic-expert trainability to examine when a second pathway helps (Section 3.3).

With generative initialization and visual-dynamics supervision, UAM retains 95.7% of pretrained scores across six multimodal metrics while improving action performance under end-to-end robot adaptation, without parameter freezing, gradient blocking, or auxiliary VL replay (Section 4.1, Table 3). After training on 3,000 real-robot trajectories, it achieves the highest average success among the compared variants on semantic OOD tasks involving unseen objects, novel object–target compositions, and instruction variation (Section 4.2, Fig. 5). Visual-routing analysis shows complementary attention patterns: action queries attend to task objects through semantic tokens and interaction regions through Dorsal tokens (Section 4.3, Fig. 6). These patterns are consistent with the two pathways providing complementary semantic and control-oriented information.

Our contribution is an empirical account of forgetting and retention within unified predictive VLAs. First, the coupling study shows that separating action-expert parameters leaves substantial semantic

degradation. Second, the design study identifies generative initialization with visual-dynamics supervision as the most effective tested pathway for control and OOD generalization. Third, matched ablations demonstrate improved retention and control with this pathway while the semantic expert remains trainable. These findings establish semantic preservation as a property to investigate in unified multimodal adaptation and provide concrete design variables for future studies.

2 Related Work

VLA adaptation and knowledge preservation. RT-2 and OpenVLA predict tokenized actions (Brohan et al., 2023; Kim et al., 2024); Octo and the π_0 family study generalist policies and continuous action prediction (Team et al., 2024; Black et al., 2024; Intelligence et al., 2025). These models use pretrained visual and linguistic knowledge to condition action generation, but robot adaptation can interfere with that knowledge. Co-training supplies additional multimodal supervision, while knowledge-insulation methods regulate gradient flow between semantic and action modules (Zhou et al., 2025; Driess et al., 2025). UAM examines a complementary intervention: a separately supervised visual pathway that supplies control-oriented features while the semantic expert remains trainable. Its evaluation therefore considers both action performance and the multimodal competence retained after adaptation.

Predictive models and unified embodied architectures. Visual prediction supports robot learning through video generation, predictive representations, and understanding–generation–action models (Du et al., 2024; Hu et al., 2024; Zhang et al., 2025; Lv et al., 2025; Li et al., 2025). Unified multimodal models supply aligned understanding and generation representations (Xie et al., 2024; Deng et al., 2025). BagelVLA (Hu et al., 2026) combines these representations with action generation, visual prediction, and single-step extraction of predictive features. UAM adopts this predictive expert formulation as a testbed for action-induced forgetting. Our contribution is a controlled study of how the second pathway's initialization, input, supervision, and semantic-expert trainability affect control and retention. The resulting evidence connects a supervised generative pathway to preserved multimodal competence during robot adaptation, complementing prior work on predictive action generation.

3 From Forgetting to a Dorsal Expert

We first measure forgetting under different semantic–action couplings, then use the remaining degradation to motivate a second visual pathway. The design study asks which initialization, inputs, and supervision make that pathway useful for control. The matched retention experiments in Section 4.1 then test whether the selected design also preserves pretrained multimodal scores.

3.1 Measuring Forgetting Across VLA Couplings

A policy π_θ maps images I_t, an instruction L, and robot state s_t to an action chunk. The adaptation dataset $\mathcal{D}_{\text{act}} = \{\tau_i\}_{i=1}^{N}$ consists of trajectories containing time-indexed observations, instructions, states, and actions. Each training example pairs the current context with demonstrated actions; UAM additionally uses a future observation from the same trajectory as a visual prediction target. We evaluate the semantic expert through its multimodal question-answering interface independently of the action policy, separating retained competence from successful execution.

Retention metric. For an evaluation set $\mathcal{B}$, let $s_b(f)$ denote a higher-is-better score on metric b and f_0 the corresponding pretrained semantic model. To avoid averaging incompatible score scales, we normalize each metric by its own nonzero reference score:

$$R_{\mathcal{B}}(f) = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \frac{s_b(f)}{s_b(f_0)}, \qquad \Delta_{\mathcal{B}}(f) = 1 - R_{\mathcal{B}}(f). \tag{1}$$

We report $100R_{\mathcal{B}}$ as retention (%); values above 100% are possible if scores improve. The four-metric matched control uses MMMU, MME-P, MME-S, and MMBench. The broader comparison additionally includes MM-Vet and MathVista where pretrained reference scores are available. These averages count both reported MME columns. The earlier coupling diagnostic uses its original normalized three-benchmark aggregate S, with retention $S(f)/S(f_0)$; we keep this diagnostic separate from the matched comparisons under Eq. 1.

Coupling diagnostic. We compare a frozen semantic backbone, a trainable backbone with a parallel MoT action expert, and a trainable backbone with a sequential MLP action head. The diagnostic covers a Qwen-based 7B VLM and PaliGemma-1; checkpoint sources and training details are given in Appendix B. Each model is adapted on CALVIN action data without auxiliary VL co-training, and evaluated on both control performance and retained multimodal scores.

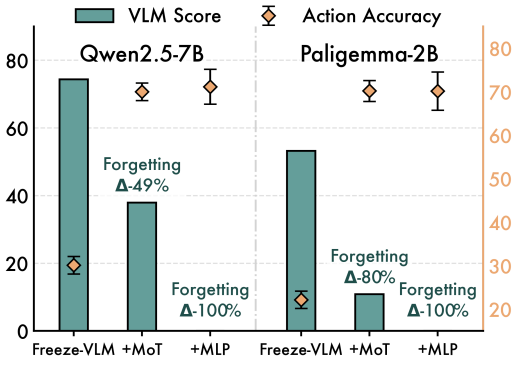


Figure 2: **Coupling affects forgetting.** VLM and action scores under frozen, MoT, and MLP couplings. Protocols are in Appendix B.

Freezing preserves the semantic score but reduces control performance. Both trainable couplings improve action performance while degrading multimodal scores, with MoT retaining more than MLP (Fig. 2). The 7B MoT diagnostic retains $37.94/74.32 = 51.0\%$ of its original aggregate score. The pattern also occurs with PaliGemma, but the comparison changes both model family and scale; it does not isolate a causal effect of parameter count.

3.2 A Bottleneck Hypothesis

A separate action expert does not necessarily create a separate visual representation. In a two-expert MoT VLA, the action expert still consumes features supplied by the semantic pathway. Action gradients can therefore change representations that also support general multimodal understanding. We hypothesize that an additional, visually grounded pathway can reduce this competing demand.

Why parameter separation may be insufficient. The semantic expert must identify the instructed object and also expose spatial features needed to reach, grasp, or move it. MoT gives the action expert its own parameters, but control-specific adaptation can still modify the semantic features on which that expert depends. A Dorsal pathway offers an alternative source of visual information about the arm, interaction geometry, and scene evolution. The goal is not to prevent communication between recognition and control: the action expert still needs language-grounded semantic information. Instead, the hypothesis is that providing complementary visual features reduces the need to obtain every control-relevant representation by changing the semantic pathway.

UAM couples a semantic expert E_{sem}, a Dorsal Expert E_{dor}, and an action expert E_{act}. Schematically,

$$Z_{\text{sem}} = E_{\text{sem}}(I_t, L), \quad Z_{\text{dor}} = E_{\text{dor}}(X_{\text{dor}}), \quad a_t = E_{\text{act}}(Z_{\text{sem}}, Z_{\text{dor}}). \tag{2}$$

Here $X_{\text{dor}} \in \{I_t, q\}$, where q denotes learnable queries; expert parameters are omitted for brevity. These expressions summarize token sources; the implementation exchanges information across parallel MoT layers according to the attention masks in Appendix C. UAM routes ViT-encoded observations to the semantic expert and VAE-encoded observations to the generative Dorsal Expert. Proprioceptive state is supplied to the action expert. The two visual pathways provide complementary token sources for action prediction.

3.3 Visual Priors and Matched Supervision

We study the six design points in Table 1. Within the sweep, variants share the action expert and MoT coupling and differ in the Dorsal pathway. Each frozen or trainable semantic-expert setting is trained separately. The sweep compares control, OOD behavior, and retention after real-robot adaptation; Section 4.1 provides additional matched controls.

We assess a useful Dorsal design against two requirements: it should support in-distribution control and semantic OOD generalization, and the resulting model should retain its pretrained multimodal competence. Reading frozen and trainable settings together helps determine whether a pathway can support control when adaptation of the semantic expert is restricted. In the frozen setting, the Dorsal and action experts must learn to use the available visual and semantic features without changing

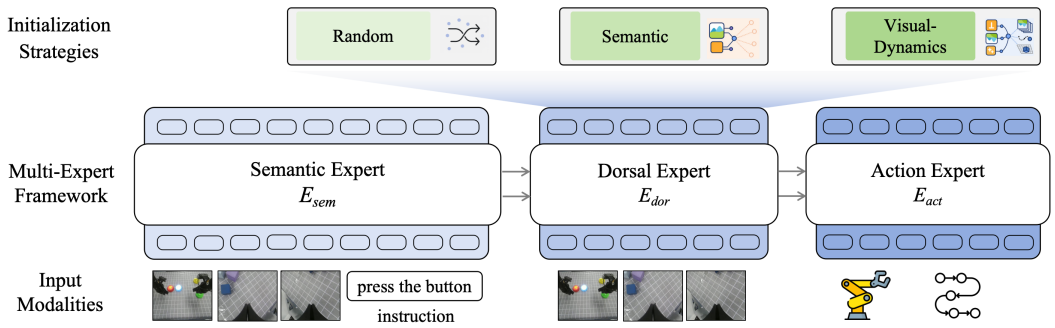


Figure 3: **UAM and the Dorsal design space.** Three experts exchange information through MoT. Comparisons vary the second pathway's initialization, input, and visual-dynamics supervision.

Table 1: **Dorsal Expert designs and retention.** Variants share the action expert and MoT coupling. Retention is the four-metric average after real-robot fine-tuning with a trainable semantic expert (Appendix F).

Variant	Initialization	Input	Auxiliary loss	Retention (%)
No Dorsal	–	–	None	54.8
Random-init	Random	Vision	None	73.2
VLM-init (2a)	Semantic VLM	Vision	None	92.4
VLM-init (2b)	Semantic VLM	Queries	None	–
Gen-init (3a)	Generative UMM	Vision	None	94.2
UAM (3b)	Generative UMM	Vision	Visual dynamics	**96.7**

the semantic expert's parameters. The matched adaptation controls then directly assess retention alongside control performance.

Variant 1: random initialization. Adding a randomly initialized Dorsal Expert raises retention from 54.8% to 73.2% (Table 1), showing that an additional visual pathway already alleviates semantic degradation. Its retention nevertheless remains below the pretrained variants. Control performance is comparable to the two-expert baseline in simulation but weaker on real-world OOD tasks under both frozen and trainable semantic settings (Fig. 4). Thus, the extra pathway improves semantic preservation, while its initialization remains important for learning representations that generalize beyond the demonstrations.

Variant 2: semantic initialization and input modality. Semantic initialization with visual input (2a) further improves retention to 92.4%, compared with 73.2% for random initialization. This supports using pretrained visual representations to preserve the semantic expert during adaptation. Despite this high retention, 2a remains weaker than UAM on real-world OOD tasks, indicating that preserving VLM competence alone does not ensure effective action generalization. Replacing visual input with learnable queries (2b) reduces control performance, especially under frozen-semantic training. Visual tokens provide direct observation features, whereas queries must acquire scene information through interactions with the other experts.

Variant 3: generative initialization. Generative initialization without visual-dynamics supervision (3a) retains 94.2% of pretrained scores, but its control performance remains similar to 2a and its OOD generalization falls below UAM. Adding future-observation supervision in UAM (3b) raises retention to 96.7% and yields the strongest combined control performance. In several settings, it remains competitive with the trainable two-expert reference even when its semantic expert is frozen (Fig. 4). Together, the variants show that an additional pretrained pathway supports semantic preservation, while visual future supervision is a key complement for action generalization. The high retention of 2a and 3a highlights the distinction: retaining semantic knowledge provides a foundation, and predictive supervision helps the Dorsal pathway supply the interaction-oriented features needed to use that knowledge in control.

Visual-dynamics supervision. UAM adopts BagelVLA's predictive training formulation (Hu et al., 2026), with $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{act}} + \mathcal{L}_{\text{wm}}$. Here $\mathcal{L}_{\text{act}}$ is action flow matching, and $\mathcal{L}_{\text{wm}}$ is visual latent flow

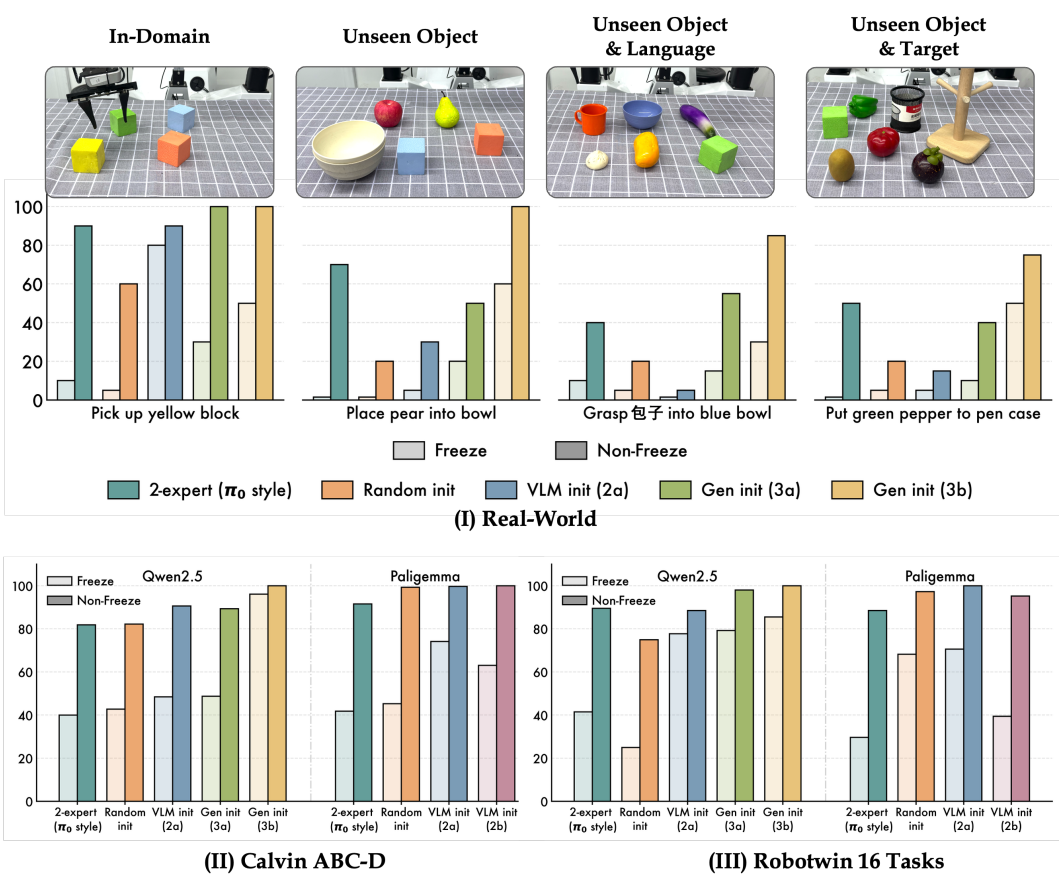


Figure 4: **Control performance across Dorsal designs.** Simulation and real-world results under frozen and trainable semantic-expert settings. Generative initialization with visual-dynamics supervision gives the strongest combined performance.

matching for a future observation from the same robot trajectory. Robot trajectories supply both action targets and auxiliary visual supervision, without external VL replay. All three expert parameter sets remain trainable without gradient blocking. At inference, the policy uses intermediate generative representations from a single denoising step without reconstructing a future image.

The three-expert coupling, visual prediction objective, and single-step feature extraction are inherited components; the Dorsal Expert denotes the generative expert's role in this study. We simplify image inputs and omit large-scale embodied pretraining to study adaptation directly on robot trajectories (Appendix C.3). The new finding is that this predictive formulation supports strong semantic retention when paired with the selected prior and supervision, beyond its established use for action generation. The design ablations characterize retention and control; the matched comparisons in Section 4.1 quantify retention and the benefit of semantic adaptation.

4 Experiments and Analysis

We examine multimodal retention, its relationship to robot control, and qualitative visual routing. The real-robot models are adapted for 30,000 steps on 3,000 ALOHA trajectories. Simulation models use their respective benchmark data and protocols (Appendix F). UAM contains approximately 7B semantic, 7B generative, and 2B action-expert parameters, or 16B in total. Training and inference costs are reported in Appendix H.

Training and control evaluation. The reported Qwen-based training uses a learning rate of 10^{-5}. CALVIN ABC–D models are trained for 30,000 steps with an effective batch size of 192 and action chunks of length 10; evaluation averages the number of completed tasks over 1,000 five-task

sequences in environment D. The real-robot setting uses chunks of length 24 and three camera views: the primary view and two wrist views.

4.1 Multimodal Retention under Matched Initialization

Table 2 reports multimodal benchmark scores, and Table 3 compares adaptation settings with the same semantic initialization. The ablations examine the Dorsal pathway, semantic-expert trainability, and knowledge insulation (KI). The No-Dorsal model retains the action expert, action data, optimization settings, and training steps while removing the Dorsal pathway and its visual loss. Pretrained references and evaluation details are provided in Appendix G.

Table 2: **Multimodal understanding benchmarks.** Cross-model scores provide context rather than matched retention comparisons. Asterisks denote VL co-training; dashes indicate unavailable scores.

Model	Params	MMMU	MME-P	MME-S	MMBench	MM-Vet	MathVista	MMStar	TextVQA
Vision Language Models									
Qwen2-VL Bai et al. (2025)	1.5B	41.1	—	1872	74.9	49.5	43.0	48.0	79.7
Qwen2.5-VL Yang et al. (2024)	3B	53.1	—	2157	79.1	61.8	62.3	56.3	79.3
LLava-OV Li et al. (2024)	7B	48.8	1580	—	80.8	57.5	63.2	61.7	—
Qwen2-VL Bai et al. (2025)	7B	54.1	—	2327	83.0	62.0	58.2	60.7	84.3
Qwen2.5-VL Yang et al. (2024)	7B	58.6	—	2347	83.5	67.1	68.2	63.9	84.9
Kimi-VL Team et al. (2025)	3B/16B	57.0	—	—	83.1	66.7	68.7	61.3	—
DeepSeek-VL2 Lu et al. (2024)	4B/27B	51.1	—	2253	83.1	60.0	62.8	61.3	84.2
Unified Multimodal Models									
Show-o_{512} Xie et al. (2024)	1.3B	26.7	1097	—	—	—	—	—	—
Janus Wu et al. (2025)	1.5B	30.5	1338	—	69.4	34.3	—	37.6	—
Janus-Pro Chen et al. (2025b)	7B	41.0	1444	—	79.2	50.0	—	—	—
LlamaFusion Shi et al. (2024)	8B	41.7	1604	—	72.1	—	—	—	—
SEED-X Ge et al. (2024)	13B	35.6	1457	—	70.1	43.0	—	—	—
BAGEL Deng et al. (2025)	7B MoT	55.3	1687	2388	85.0	67.2	73.1	—	—
Vision–Language–Action Models									
OpenVLA Kim et al. (2024)	7B	0	0	—	0	0	0	0	0
ECoT Zawalski et al. (2024)*	7B	5.4	0	—	—	—	—	0	0
DiVLA Wen et al. (2025)*	2B	17.2	187	—	—	—	—	21.1	7.5
ChatVLA Zhou et al. (2025)*	2B	37.4	1435	—	69.0	—	—	47.2	71.2
$\pi_{0.5}$-base Intelligence et al. (2025)*	2B MoT	18.7	1032	1241	7.3	—	—	—	—
UAM	7B MoT	**53.7**	**1607**	**2289**	**83.7**	**63.4**	**68.2**	**61.3**	**84.2**

Effect of the Dorsal pathway. UAM retains 96.7% of pretrained scores across the four matched metrics, compared with 54.8% for No-Dorsal, an improvement of 41.9 percentage points. This higher retention accompanies stronger control performance: CALVIN average completed sequence length increases from 3.30 to 4.13, and success on the additional baozi OOD task rises from 35% to 70% (Table 3). Thus, adding the supervised Dorsal pathway improves both semantic preservation and action performance under the same adaptation settings.

The retention gains extend across the evaluated capabilities. UAM retains 97.1% on MMMU and 98.5% on MMBench, versus 59.0% and 29.5% for No-Dorsal, while its two MME retention ratios reach 95.3% and 95.9%. Including MM-Vet and MathVista gives 95.7% average retention across all six metrics with pretrained references. Table 2 additionally reports scores of 61.3 on MMStar and 84.2 on TextVQA. The benchmark suite covers multidisciplinary understanding, perception, reasoning, and text recognition (Yue et al., 2024; Fu et al., 2023; Yu et al., 2023; Lu et al., 2023; Chen et al., 2024; Singh et al., 2019), showing that UAM maintains broad multimodal competence after robot adaptation.

Effect of semantic-expert adaptation. Freezing UAM's semantic expert preserves 100% of its pretrained scores, with a CALVIN score of 3.35 and baozi OOD success of 30%. Allowing this expert to adapt raises these action scores to 4.13 and 70%, respectively, while retaining 96.7% of its multimodal scores (Table 3). The trainable setting therefore combines substantial semantic preservation with improved control, indicating that the Dorsal pathway supports effective action learning while leaving the semantic expert available for task adaptation.

Table 3: **Retention and control ablations.** Retention and baozi OOD success (%) use real-robot-trained checkpoints; CALVIN sequence length uses separately trained CALVIN checkpoints. Training and evaluation protocols are in Appendix F; ablation details are in Appendix G.

Model	Retention (%)	CALVIN	Baozi OOD (%)
BAGEL + action (No Dorsal)	54.8	3.30	35
UAM, frozen semantic expert	100.0	3.35	30
BAGEL + action + KI, no QA replay	2.3	2.89	25
BAGEL + action + KI, LLaVA-665K	83.2	2.72	45
UAM	96.7	4.13	70

Comparison with knowledge insulation. The KI variant blocks continuous action-expert gradients into the semantic backbone and trains it with discrete action-token supervision. Adding LLaVA-665K QA co-training raises retention from 2.3% to 83.2% and baozi OOD success from 25% to 45%, with CALVIN scores of 2.89 and 2.72, respectively. In this ablation setting, UAM achieves higher retention (96.7%) and stronger action performance (4.13 on CALVIN and 70% baozi OOD success) without auxiliary QA replay (Table 3). These comparisons support the effectiveness of a supervised visual pathway for jointly preserving multimodal competence and improving control during robot adaptation. Implementation details are given in Appendix G.

4.2 Semantic Generalization in Robot Manipulation

We assess UAM's semantic generalization in real-robot manipulation; task details and demonstrations are provided in Appendix D.

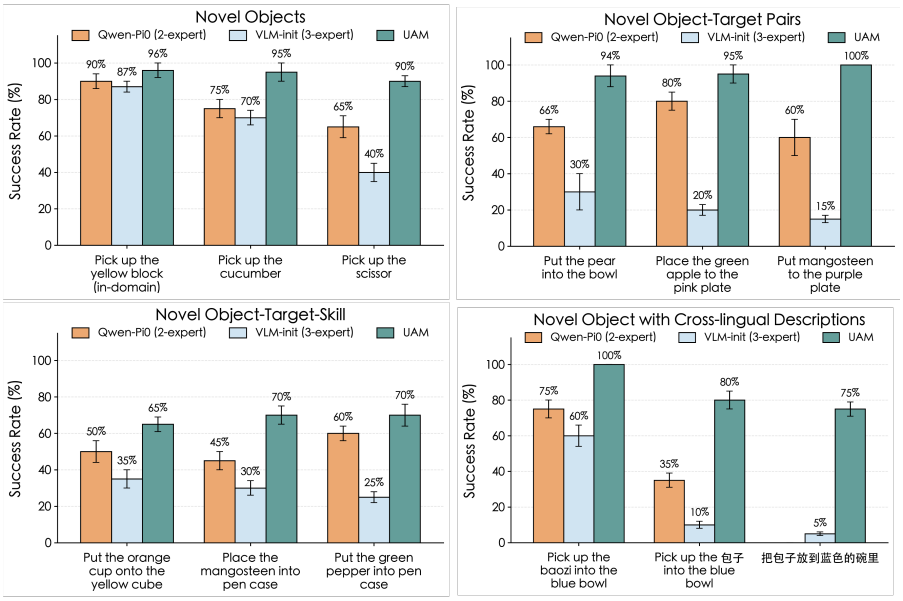


Figure 5: **Real-world semantic generalization.** Success rates (%) for the two-expert baseline, VLM-initialized three-expert variant, and UAM after training on 3,000 robot trajectories. Panels compare novel objects, unseen object–target pairs, new placement configurations, and cross-lingual instructions; the yellow-block task provides an in-domain reference. The final panel uses pinyin, Chinese–English code-mixing, and fully Chinese instructions for the same manipulation goal. Task demonstrations appear in Appendix D, Figs. 8 and 9.

As shown in Fig. 5, UAM achieves the highest average success among the compared variants. It reaches 95% and 90% success on the unseen cucumber and scissor tasks, and 94–100% on novel object–target pairs. Its advantage also extends to new placement configurations, including placing objects onto supports or into unfamiliar containers. Under instruction variation, UAM achieves 100%, 80%, and 75% success with pinyin, mixed-language, and fully Chinese instructions, respectively. These results show that UAM supports object recognition, compositional grounding, and multilingual instruction following within a learned action policy.

The VLM-initialized three-expert variant performs comparably to the two-expert baseline on the in-domain task (87% versus 90%), but achieves lower success on the OOD tasks. UAM improves across these settings by combining generative initialization with visual-dynamics supervision. The comparison highlights the importance of the second pathway's representation and learning objective: a predictive visual pathway provides interaction-oriented features alongside the semantic expert's object and instruction representations. Together with the retention results in Section 4.1, these evaluations demonstrate that UAM combines broad multimodal competence with effective semantic generalization after robot adaptation.

4.3 Qualitative Analysis of Visual Routing

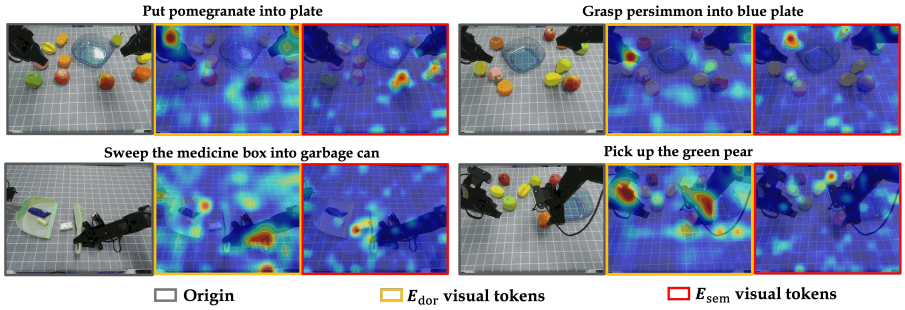


Figure 6: **Complementary visual routing during action generation.** Each task shows the input camera image, action-query attention over Dorsal tokens (yellow border), and attention over semantic tokens (red border). Semantic-token attention emphasizes instructed objects and goal regions, while Dorsal-token attention covers the robot arm, tools, and surrounding interaction context. The visualization procedure is detailed in Appendix E.

We analyze visual attention during action generation to examine the complementary roles of the semantic and Dorsal pathways. Figure 6 shows distinct, complementary patterns across the two token sources. In the pomegranate-placement and green-pear examples, attention over semantic tokens concentrates on the instructed fruit and, for placement, the receiving region. Attention over Dorsal tokens extends to the gripper and surrounding workspace. The sweeping example exhibits a similar division: semantic-token attention highlights the object to be moved, while Dorsal-token attention covers the tool and its interaction context. Across these tasks, the semantic pathway emphasizes what the policy should act on, whereas the Dorsal pathway supplies spatial context for execution.

This visual organization is consistent with the roles of UAM's two pathways. The semantic expert provides object and goal grounding, while the generatively initialized Dorsal Expert learns predictive features through future-observation supervision. The action expert accesses both token sources within the same prediction, combining instruction-relevant entities with information about the robot and its surroundings. These complementary attention patterns give a qualitative account of how UAM's dual-stream design supports manipulation, alongside the improvements in retention and semantic generalization reported in Sections 4.1 and 4.2.

5 Discussion and Conclusion

We studied the embodiment tax in VLA training and introduced UAM to investigate semantic preservation through complementary visual pathways. Built from a unified multimodal model (UMM), UAM combines a generative Dorsal Expert with visual-dynamics supervision, retaining over 95% of pretrained multimodal scores while improving action performance and semantic generalization under end-to-end robot adaptation. These results identify strong semantic retention as a distinctive property of this UMM-based design. The current method relies on representation alignment from joint understanding–generation pretraining, as provided by BAGEL; extending it to non-unified backbones therefore presents a potential alignment challenge. Our findings motivate further study of unified multimodal pretraining and predictive visual pathways as foundations for robot policies that acquire action skills while preserving broad semantic competence.

AI USE STATEMENT

AI tools assisted with manuscript editing, including language polishing and grammar correction.

REPRODUCIBILITY STATEMENT

To support reproduction, Appendices C and F document the architecture, training settings, and evaluation protocols; Appendix G describes the additional controls and their reporting scope.

ETHICS STATEMENT

Preserved multimodal benchmark performance does not guarantee safe robot behavior. Manipulation policies can misinterpret instructions, fail on unfamiliar objects, or produce unsafe motions. Deployment requires task-specific validation, supervision, and appropriate physical safeguards. Our experiments do not establish reliability for unsupervised operation around people.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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.

Kevin Black, Manuel Y. Galliker, and Sergey Levine. Real-time execution of action chunking flow policies. *arXiv preprint arXiv:2506.07339*, 2025. URL https://arxiv.org/abs/2506.07339.

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. *arXiv preprint arXiv:2307.15818*, 2023.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

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.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Danny Driess, Jost Tobias Springenberg, Brian Ichter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Z Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, et al. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. *arXiv preprint arXiv:2505.23705*, 2025.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.

Yucheng Hu, Jianke Zhang, Yuanfei Luo, Yanjiang Guo, Xiaoyu Chen, Xinshu Sun, Kun Feng, Qingzhou Lu, Sheng Chen, Yangang Zhang, Wei Li, and Jianyu Chen. BagelVLA: Enhancing long-horizon manipulation via interleaved vision-language-action generation. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.083.

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, 2025. *URL https://arxiv. org/abs/2504.16054*, 1(2):3, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Qi Lv, Weijie Kong, Hao Li, Jia Zeng, Zherui Qiu, Delin Qu, Haoming Song, Qizhi Chen, Xiang Deng, and Jiangmiao Pang. F1: A vision-language-action model bridging understanding and generation to actions. *arXiv preprint arXiv:2509.06951*, 2025.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):7327–7334, 2022.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili YU. LMFusion: Adapting pretrained language models for multimodal generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Kc1WTxZbrP.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy, 2024. URL https://arxiv.org/abs/2405.12213.

Leslie G. Ungerleider. Two cortical visual systems. In *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 1982. URL https://api.semanticscholar.org/CorpusID:142774685.

Junjie Wen, Yichen Zhu, Minjie Zhu, Zhibin Tang, Jinming Li, Zhongyi Zhou, Xiaoyu Liu, Chaomin Shen, Yaxin Peng, and Feifei Feng. Diffusionvla: Scaling robot foundation models via unified diffusion and autoregression. In *Forty-second International Conference on Machine Learning*, 2025.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12966–12977, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

Jianke Zhang, Xiaoyu Chen, Qiuyue Wang, Mingsheng Li, Yanjiang Guo, Yucheng Hu, Jiajun Zhang, Shuai Bai, Junyang Lin, and Jianyu Chen. Vlm4vla: Revisiting vision-language-models in vision-language-action models. *arXiv preprint arXiv:2601.03309*, 2026.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, XinQiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. DreamVLA: A vision-language-action model dreamed with comprehensive world knowledge. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=PK07eretkF.

Zhongyi Zhou, Yichen Zhu, Minjie Zhu, Junjie Wen, Ning Liu, Zhiyuan Xu, Weibin Meng, Yaxin Peng, Chaomin Shen, Feifei Feng, et al. Chatvla: Unified multimodal understanding and robot control with vision-language-action model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5377–5395, 2025.

A Evaluation Details in Methodology

Evaluation Protocol To systematically investigate the forgetting phenomenon and the effectiveness of our proposed solution, we decouple the evaluation of VLA models into two distinct dimensions:

- *Semantic Retention (Forgetting Metric):* We evaluate the fine-tuned VLA models on standard multimodal understanding benchmarks (MMMU, MME, and MMBench) to quantify the preservation of the VLM's pre-trained general knowledge and reasoning abilities.
- *Action Performance (Control Metric):* To ensure that semantic retention does not compromise control accuracy, we rigorously test the action success rate in both simulation and the real world.
  - For simulation, we utilize the Calvin (ABC-D split) (Mees et al., 2022) and RoboTwin (Chen et al., 2025a) for the in-domain manipulation test.
  - For RoboTwin, the reported design-sweep training uses 800 demonstrations (16 tasks with 50 demonstrations each). Specifically, from the original 50 tasks, we selected 1 to 2 tasks per skill category that exhibited the lowest success rates, thereby representing the most challenging scenarios. The models were trained and subsequently evaluated on this suite (utilizing 50 expert demonstrations per task), which includes the following tasks: Beat Block Hammer, Blocks Ranking Size, Click Alarmclock, Dump Bin Bigbin, Handover Block, Hanging Mug, Lift Pot, Move Can Pot, Pick Diverse Bottles, Place Can Basket, Place Dual Shoes, Put Bottles Dustbin, Scan Object, Stack Blocks Three, Stamp Seal, and Turn Switch.
  - For real-world deployment, we evaluate the models on a dual-arm ALOHA system, assessing out-of-distribution (OOD) semantic generalization capabilities (unseen objects during action-only finetuning). Following training on the 3,000 collected demonstration trajectories (without an additional large-scale robot pretraining stage; the foundation experts are pretrained), we evaluate the model across various OOD tasks, which target objects that are unseen in the training set, alongside the deliberate introduction of additional distractor objects. Detailed tasks are shown in Fig 4 and Appendix D.

This dual-metric evaluation framework is designed to simultaneously verify two crucial aspects: first, whether the VLA method retains its semantic understanding alongside strong manipulation proficiency after training exclusively on action data; and second, whether it can effectively transfer its inherent vision-language generalization capabilities to the manipulation of out-of-distribution (OOD) objects in real-world robotic deployments.

B Empirical Observations on the Forgetting Dilemma

We initialize the Qwen2.5-7B model using VLM parameters from Bagel (Deng et al., 2025) (which empirically outperforms the original Qwen2.5VL-7B) and the PaliGemma model using PaliGemma-1. Building upon these two VLMs, we construct VLAs by integrating action experts in two distinct ways: a parallel Mixture-of-Transformers (MoT) architecture following π_0 Black et al. (2024), and a sequential MLP action head following VLM4VLA Zhang et al. (2026). We perform full-parameter fine-tuning on these models using the Calvin ABC dataset for 10k steps. Specifically, we employ a learning rate of 1×10^{-5} for the Qwen-7B series and 5×10^{-5} for the PaliGemma series. Subsequently, we evaluate the understanding capabilities of the VLM component and the overall action performance of the VLA. Note that the simulation evaluation here primarily assesses in-domain action accuracy rather than generalization to novel objects.

The results, presented in Tab 4 and Fig 2, reveal that even when the action policies have largely converged, the degree of retained VLM performance varies significantly across different architectures. For the baseline where the VLM is frozen, although it entirely preserves the model's intrinsic capabilities and yields the highest understanding scores, it suffers from severely degraded action performance. Conversely, all approaches that unfreeze the VLM incur a degradation in its original capabilities. Specifically, the VLA employing the 7B-LLM and the MoT architecture exhibits the best retention of VLM capabilities. In contrast, while the sequential MLP head achieves comparable action accuracy to the MoT architecture, its inherent VLM capabilities are catastrophically destroyed

(yielding a VQA score of 0). MoT preserves more multimodal score than MLP in these diagnostic runs.

Table 4: Detailed results of Fig 2. For the "Action" metric, we report the normalized average completion rate on the test simulation environment as the indicator of action performance. For the "VLM" capability, we use the normalized average score across three standard VQA benchmarks as the evaluation metric.

VLA Arch (VLM+Action Head)	Action	VQA-AVG	MME	MMMU	MMBench
Qwen2.5-Freeze + MLP (7B+10M)	30.12	74.32	2354	53.7	85.2
Qwen2.5 + MoT (7B+1B)	65.98	37.94	1675	30.46	23.54
Qwen2.5 + MLP (7B+10M)	71.14	0	0	0	0
Paligemma-Freeze + MLP (2.3B+10M)	22.18	53.18	1670	34.66	65.23
Paligemma + MoT (2.3B+0.3B)	70.18	10.83	400	6.2	12.00
Paligemma + MLP (2.3B+10M)	70.12	0	0	0	0

C Dorsal Expert Design Implementation Details

In this section, we provide detailed implementations of the ablation variants for the Dorsal Expert discussed in Section 3.2. Specifically, we integrate two or three experts using the Mixture-of-Transformers (MoT) mechanism, enabling information exchange among the multiple experts through the design of distinct attention masks.

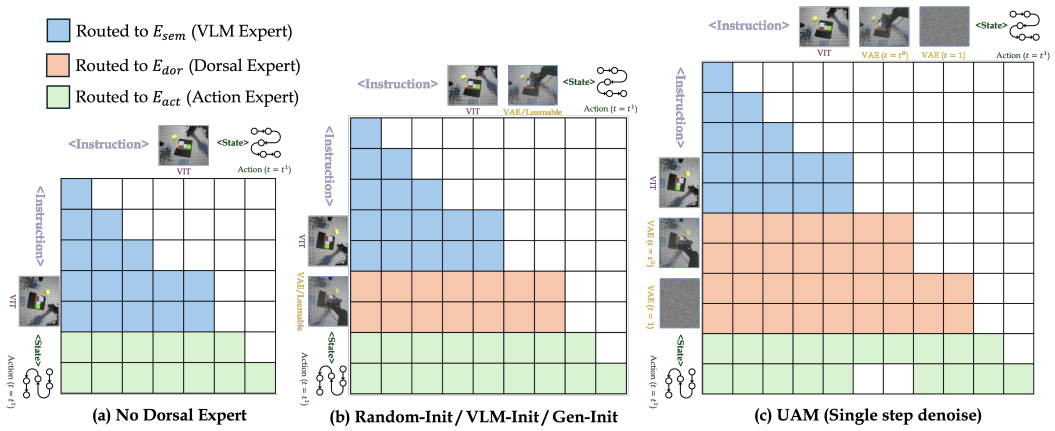


Figure 7: Attention Mask used for Different Dorsal Expert Design. We visualize the input sequence formats and their corresponding attention masks across different architectural designs (where white regions indicate masked attention). Additionally, distinct colors are employed to denote how different types of tokens are routed to their respective experts by the MoT mechanism.

C.1 No Dorsal Expert (2-Expert)

We establish the architecture without the Dorsal expert as our baseline. Specifically, following the approach of π_0 Black et al. (2024), we couple the VLM expert and the action expert via the MoT, while directly feeding the proprioceptive state into the action expert. The detailed attention implementation is illustrated in Fig 7(a).

C.2 Random-init, VLM-init and Gen-init

The implementations of the Random-init, VLM-init, and Gen-init variants share a similar architectural design, differing solely in the initialization weights of the intermediate Dorsal Expert. Specifically,

for the Random-init variant, the Dorsal Expert is initialized using standard random weights. For VLM-init, we directly duplicate the weights of E_{sem} (the VLM expert) to initialize the Dorsal Expert. For Gen-init, it is initialized using the pre-trained weights of the Gen expert from the pretrained UMM model (Bagel Deng et al. (2025)). All three variants employ a similar attention masking strategy, as depicted in Fig 7(b).

C.3 UAM

The implementation of UAM builds upon the Gen-init variant by incorporating an additional world modeling supervision signal. We observe that the prior alignment between the dynamics expert and the VLM is crucial. Consequently, we implement this approach exclusively on the Qwen model, utilizing the pre-trained Bagel Deng et al. (2025) weights (as applying it to the PaliGemma series would necessitate extensive additional alignment between multimodal understanding and generation). Specifically, following the implementation of BagelVLA Hu et al. (2026), we simplify its original image input format and discard the large-scale embodied pre-training phase, opting instead to train the model directly on action data. Fig 7(c) illustrates the attention masking strategy and the input sequence concatenation scheme for UAM. Notably, tokens encoded by the ViT are routed to E_{sem} (the VLM expert), whereas tokens encoded by the VAE are directed to E_{dor} (the Dorsal expert). Finally, we adopt the single-step denoising mechanism from BagelVLA to facilitate the interaction within the dual flow-matching framework.

D Details and Demos of Real World OOD Tasks

On the real-world bimanual robotic platform, we train the UAM using 3,000 collected demonstration trajectories. Specifically, the tasks within this training dataset can be broadly categorized into the following types:

- **Pick & Place:** Grasping and placing objects. The training set includes toy fruits, a computer mouse, colorful blocks, toy phones, and so on. The placed targets include colorful plates, baskets, boxes, and so on.
- **Water Flower:** grasping the handle of a toy watering can to simulate the pouring action of watering a plant.
- **Stack Cubes:** Stack blocks of four different colors.
- **Stack Bowls:** Stack bowls of three different colors according to a specified color sequence.
- **Pour Fries:** Open the lid of a carton and pour the toy fries from inside it onto a plate.
- **Sweep Rubbish:** Grasp a toy broom, sweep the randomly placed tissue paper trash on the table into a dustpan, and then put down the broom.
- **Press Button:** Press different buttons in a specified color sequence.
- **Drawer Operation:** Opening and closing a drawer.
- **Other Complex Motion Tasks:** Additionally, the training dataset further includes tasks that require complex manipulations, which encompass: wiping a table with a cloth, inserting flowers into a vase, hanging a cloth, and plugging in a power cable, and so on. These tasks primarily demand high manipulation precision, placing minimal requirements on the model's semantic generalization capabilities.

During evaluation, our primary objective is to verify whether the model possesses sufficient semantic generalization capabilities. Specifically, we utilize pick-and-place, placement, and recognition tasks to assess the model's ability to manipulate elements unseen during training, including novel objects, novel targets, novel distractor objects, and novel language instructions (or linguistic variations). We evaluate the model across a diverse range of such out-of-distribution (OOD) tasks, including the real-world evaluation scenarios depicted in Fig 4 and 5. Demos of evaluation demonstrations are also provided in Fig 8 and Fig 9.

What the OOD settings test. Unseen-object tasks ask the policy to ground an object name without having seen that object in action demonstrations. Composition shifts test whether known or novel objects can be placed at the instructed target instead of following a familiar pairing. Language variation changes how the goal is expressed, including pinyin and code-mixed instructions, while retaining the underlying manipulation. Together, these settings test several routes by which pretrained semantics may benefit control. Their common pick-and-place structure helps focus the evaluation on semantic variation.

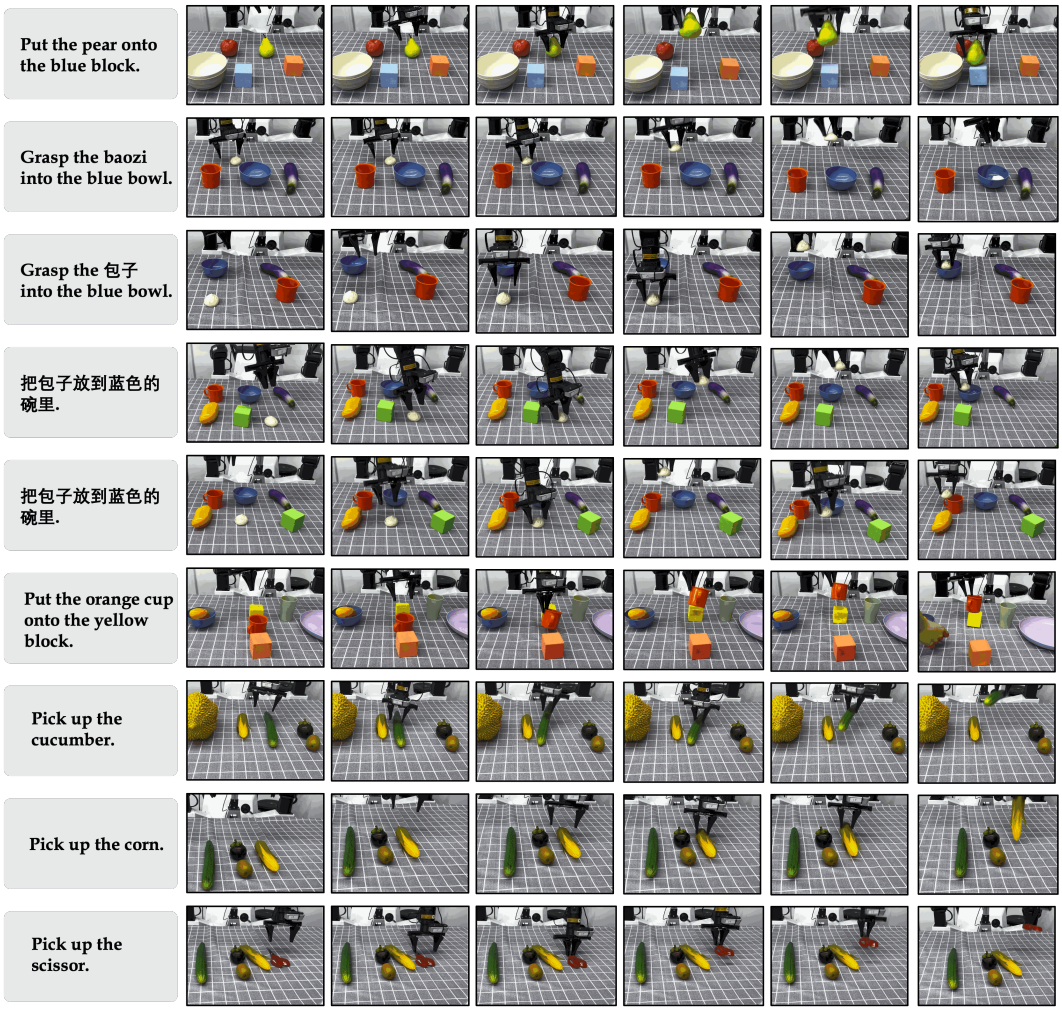


Figure 8: **Evaluation videos of UAM on semantic out-of-distribution tasks**. The text on the left denotes the input language instructions. All demonstrated scenarios are unseen during training.

E Implementation Details of Attention Visualization

In this section, we provide the core implementation details for the attention heatmaps presented in Fig. 6 and further analyze the emergent attention decoupling phenomenon observed in our dual-stream architecture.

E.1 Visualization Pipeline

To visualize the attention during action generation, we extract the MoT self-attention weights from the action prediction tokens (Queries) to the visual tokens (Keys) in the middle of self-attention layers (14/28). The weights are averaged across all attention heads and action steps to obtain a stable,



Figure 9: **Evaluation videos of UAM on semantic out-of-distribution tasks**. The text on the left denotes the input language instructions. All demonstrated scenarios are unseen during training except ”in-domain".

global distribution. The core challenge of visualization lies in accurately reconstructing the 2D spatial grids from the flattened 1D visual tokens, as the two visual encoders employ different downsampling strategies. To preserve the original 640×480 (4:3) aspect ratio and avoid spatial misalignment caused by forced square cropping, the preprocessing module dynamically adjusts the input resolution:

- **Semantic Visual Tokens** (E_{sem}): Based on a ViT architecture with a patch size of 14, it generates 252 visual tokens, which are reshaped into a 14×18 (Height $\times$ Width) 2D grid.
- **Dorsal Visual Tokens** (E_{dor}): Based on a VAE architecture with a downsampling stride of 16, it generates 192 visual tokens, which are reshaped into a 12×16 2D grid.

Finally, these low-resolution 2D grids are upsampled back to the original 640×480 resolution via bicubic interpolation and overlaid onto the original RGB images using a JET colormap.

E.2 Analysis: Emergent Attention Decoupling

As illustrated in Fig 6, the visualization reveals a compelling phenomenon: without any explicit pixel-level or grounding supervision, the model naturally learns to decouple semantic understanding from spatial-motor control, consistent with the ventral–dorsal design analogy, without establishing biological equivalence. The Semantic Stream (E_{sem}) acts as the "What" pathway. The heatmaps (red boxes) show that E_{sem} tokens are highly activated on task-relevant semantic entities specified in the language instructions. For instance, in the "Put pomegranate into plate" and "Pick up the green pear" tasks, the attention of E_{sem} sharply focuses on the target fruits and the goal receptacles, while almost completely ignoring the robot's own embodiment. This indicates that E_{sem} is primarily responsible for visual grounding and identifying "what" to manipulate. The Dorsal Stream (E_{dor}) acts as the "How/Where" pathway. Conversely, the heatmaps (yellow boxes) demonstrate that E_{dor} tokens consistently attend to the robot's end-effector (the gripper), the immediate interaction boundaries, and the spatial relationship between the arm and the objects. In the "Sweep the medicine box into garbage can" task, E_{dor} heavily focuses on the broom and the spatial gap between the broom and the box.

F Training and Evaluation Details

For experiments on Qwen, we used a learning rate of 1e-5 and employed packed datasets within the FSDP framework to maximize resource utilization. For experiments on Paligemma, we take use of the official implementation of π_0, using a learning rate of 5e-5. For action fine-tuning and evaluation, we adopted different settings for various downstream scenarios:

- Calvin ABC-D Simulation Environment: We trained on 8 A800 GPUs (effective batch size 192) for 30k steps. We used an action chunk size of 10. For evaluation, we tested on 1,000 tasks of length 5 from the D-split and reported the mean task completion length.
- Robotwin Simulation Environment: We trained on 8 A800 GPUs using 800 clean demonstrations (16 tasks $\times$ 50) for 30k steps. We used an action chunk size of 16, sampling one action every 3 steps (effective action horizon of 48).
- Real-Robot Tasks: We trained on 8 A800 GPUs for 30k steps. We used an action chunk size of 24, inputting three views (primary, left wrist, right wrist). For evaluation, we tested each task type 20 times for each checkpoint with 3 different seed values, with randomized initial positions. We analyzed the success rate error across multiple seed tests, as shown in the five error bars in Figure 5.

G Additional Ablation Protocols and Results

The ablations in Table 3 share UAM's BAGEL semantic initialization, action data, optimizer, and training protocol. No-Dorsal removes the generative expert and its visual-dynamics loss, while frozen UAM keeps the semantic expert fixed throughout adaptation. This comparison evaluates the contribution of the supervised Dorsal pathway and semantic-expert adaptation. The earlier coupling diagnostic in Appendix B uses its own pretrained references and training duration.

Knowledge-insulation protocol. The KI variant blocks continuous action-expert gradients into the semantic backbone while training it with discrete FAST action tokens. We compare this setting with and without LLaVA-665K QA co-training. In both settings, discrete action supervision updates the semantic parameters; the QA variant additionally supplies multimodal supervision. Table 3 lists aggregate retention of 2.3% and 83.2%, respectively. UAM reaches 96.7% retention without QA replay, illustrating the effectiveness of its supervised visual pathway in this comparison.

Additional OOD control. The single-task baozi comparison yields success rates of 35%, 30%, 25%, 45%, and 70% for No-Dorsal, frozen UAM, KI without QA replay, KI with QA replay, and UAM, respectively. UAM achieves the highest success among these adaptation settings, consistent with the control improvements in Section 4.1. This control is evaluated separately from the multilingual suite in Fig. 5, which includes multiple placement and language variants.

H Training and Inference Cost

Inference latency. We compare action-chunk latency on an NVIDIA A100 using ten groups of 100 calls per model (Table 5). These measurements use an unoptimized implementation, without deployment-specific sequence parallelism, CUDA optimizations, or real-time chunking (RTC). The two-expert baseline contains 7B semantic and 2B action-expert parameters; UAM adds a 7B Dorsal Expert. Mean latency increases from 1329 ms to 1485 ms, an overhead of 156 ms (11.7%). UAM uses intermediate generative features from a single denoising step, without reconstructing a future image during action inference.

Table 5: Unoptimized action-chunk latency on an NVIDIA A100 over ten groups of 100 calls.

Model	Expert parameters	Calls	Latency (ms)
Qwen-based two-expert	$7\text{B} + 2\text{B}$	10×100	1329 ± 36
UAM	$7\text{B} + 7\text{B} + 2\text{B}$	10×100	1485 ± 43

Latency–performance trade-off. Adding the Dorsal Expert to the two-expert policy incurs a modest inference overhead while substantially improving semantic retention and generalization. In the matched ablation, retention increases from 54.8% to 96.7%, CALVIN performance rises from 3.30 to 4.13, and baozi OOD success doubles from 35% to 70% (Table 3); Fig. 5 further shows gains across object, composition, and language shifts. These improvements make the three-expert MoT design an attractive trade-off for practical manipulation, where reliable semantic grounding and task completion outweigh the modest additional computation.

Deployment optimization. For deployment, we target approximately 300 ms per action chunk through sequence parallelism and CUDA-level optimization. RTC (Black et al., 2025) can additionally overlap the next chunk's computation with execution of the current chunk, reducing pauses between chunks. Together, these optimizations aim to minimize the added expert's effect on task execution while retaining its substantial performance benefits. The three-expert design thus offers a promising deployment option when semantic robustness and generalization are central to successful manipulation.