FlowVLA: Visual Chain of Thought-based Motion Reasoning for Vision-Language-Action Models (Supplementary Document)

***Abstract*—In this supplementary document, we provide additional implementation details, extended empirical results, and in-depth methodological analyses to complement our main paper. Specifically, this document covers:**

- **Implementation and Training Details: Comprehensive configurations including dataset formulations, training hyperparameters, and an analysis of computational efficiency during pre-training and fine-tuning.**
- **Additional Experiments and Ablations: Extended evaluations such as multi-view (wrist camera) experiments, CALVIN long-horizon generalization, statistical significance analysis, reduced-data policy evaluation, robust real-world evaluations with error bars, and ablation studies on intra-frame attention masking.**
- **In-depth Methodological Analysis: Deeper theoretical and architectural discussions, including clarifications on causal motion reasoning, the robustness of optical flow on textureless/specular objects, and the rationale behind our decoupled training paradigm and shared VQ-GAN tokenization.**
- **Discussions on Scaling and Baselines: Broader insights regarding model and data scaling, alongside extended comparisons with related flow-based baselines to further highlight our core contributions.**

I. Implementation and Training Details

A. Details of Training Data

In this section, we provide a comprehensive breakdown of the specific video datasets utilized for pre-training and fine-tuning our model across various experimental settings:

- **Experiments on LIBERO Benchmark:** The model was trained using the standard demonstration videos provided within the LIBERO dataset suites [1] (e.g., LIBERO-Spatial, LIBERO-Object), following the official benchmark protocols.
- **Experiments on SimplerEnv Benchmark:** For evaluations in the SimplerEnv simulation, we trained the model on the BridgeData V2 dataset [2], utilizing its large-scale real-world demonstration data to learn generalizable manipulation policies.
- **Real-World Experiments:** We initialized the model with the checkpoint trained on BridgeData V2 to acquire broad manipulation priors. We then fine-tuned this checkpoint using our custom-collected real-robot dataset to adapt to the specific physical setup and tasks.

B. Analysis of Training Efficiency

In this section, we discuss the impact of our method on training time and overall computational efficiency. A potential concern is that interleaving optical flow tokens effectively doubles the length of the visual segment in the input sequence during the pre-training phase. However, this increased sequence length does not result in a proportional increase in training time, primarily due to the following factors:

- **Parallel Processing via Teacher Forcing:** During pre-training, the model is trained using teacher forcing. Consequently, the entire sequence (including both RGB and flow tokens) is processed in parallel within a single forward pass, rather than being generated sequentially step-by-step.
- **Hardware and Algorithmic Efficiency:** Thanks to the parallelization capabilities of modern GPUs and memory-efficient attention implementations such as FlashAttention [3], the computational overhead of processing longer sequences is significantly mitigated.
- **Marginal Wall-Clock Overhead:** In our empirical evaluations, the wall-clock training time per iteration increased by only approximately **15%** compared to the baseline without flow tokens. We consider this a highly manageable and worthwhile computational cost, given the significant downstream performance improvements achieved by incorporating explicit motion representations.

C. Pre-training Efficiency Analysis

In this section, we analyze the pre-training efficiency of our proposed method. We demonstrate that our FlowVLA exhibits superior training efficiency compared to the baseline UniVLA [4], achieving better generation quality with fewer training iterations.

To substantiate this, we quantitatively evaluated the generation quality of the World Model (measured by **LPIPS** ↓ via closed-loop autoregressive rollouts) at different pre-training checkpoints (4k, 8k, and 12k steps) on the Bridge-V2 dataset [2].

TABLE I: Pre-training Efficiency Analysis on the Bridge-V2 dataset [2]. **LPIPS** (↓) is measured using closed-loop autoregressive rollouts at different pre-training steps. Lower is better.

Model	4k Steps	8k Steps	12k Steps
UniVLA	0.3052	0.2338	0.1944
FlowVLA (Ours)	**0.2614**	**0.1647**	**0.1333**

As shown in Table I, the comparison reveals a stark difference in convergence behavior:

- **Faster Convergence:** Our FlowVLA demonstrates highly efficient learning. Notably, FlowVLA trained for only 8k steps achieves a significantly better LPIPS score (0.1647)

than the baseline UniVLA fully trained for 12k steps (0.1944).

- **Better Final Quality:** Under the same 12k training steps, FlowVLA achieves a much lower LPIPS (0.1333) compared to UniVLA (0.1944), proving that our pre-training is not only faster but also yields a more accurate physical world model.

We attribute this remarkable efficiency to the explicit modeling of optical flow:

- **Structural Efficiency (Decoupling):** Direct video generation (as in UniVLA) burdens the model with simultaneously learning complex texture reconstruction and motion dynamics. By decoupling motion (flow) from appearance, our FlowVLA allows the pre-training to focus purely on learning physical dynamics first. This separation simplifies the optimization space, leading to much faster convergence.
- **Denser Supervision Signal:** The flow tokens provide explicit, pixel-level motion guidance. Compared to the implicit supervision in next-frame prediction (where motion must be inferred from subtle pixel changes), flow provides stronger and more direct gradient signals for learning world dynamics, explaining why our FlowVLA outperforms the baseline with significantly fewer training steps.

II. Additional Experiments and Ablation Studies

A. Additional Experiments with Wrist-View Cameras

In the main manuscript (e.g., Table I), we evaluated our proposed FlowVLA and the baseline UniVLA [4] using only a single global (third-person) camera view. This single-view setting was deliberately chosen to rigorously evaluate the models' spatial and motion reasoning capabilities under constrained visual inputs. Consequently, the reported average success rate of our reproduced UniVLA without wrist-view (84.0%) is lower than the 95.5% reported in the original UniVLA paper, which relies on multi-view observations (including a wrist-mounted camera) for high-precision manipulation.

To provide a comprehensive evaluation and a direct comparison with the original UniVLA multi-view setup, we conducted an additional experiment where wrist-view camera images are incorporated into both the pre-training and post-training phases of our FlowVLA framework.

The quantitative results on the LIBERO benchmark are presented in Table II. As shown, adding wrist-view observations makes the LIBERO tasks overall easier, leading both methods to near-saturated performance. FlowVLA (w/ Wrist) achieves a slightly higher average success rate than UniVLA (96.2% vs. 95.5%), but we do not treat this near-saturated setting as the primary evidence for a large performance gap.

B. CALVIN ABC→D Evaluation

For further analysis, we additionally report results on the CALVIN benchmark [5]. We follow the UniVLA evaluation protocol [4] and evaluate the ABC→D split, where the policy is trained on environments A, B, and C and evaluated on the held-out environment D. Following the standard CALVIN protocol, we evaluate 1000 five-instruction sequences and report SR@1–SR@5, where SR@k denotes the fraction of sequences in which the policy completes at least k consecutive subtasks. As shown in Table III, FlowVLA achieves a higher average sequence length, with the gain most visible at longer horizons: SR@5 improves from 0.751 to 0.815.

C. Statistical Significance Analysis

To assess whether the observed gains are statistically reliable, we conduct a fixed-sample statistical analysis on the two simulation benchmarks used in the paper. Each rollout is treated as a Bernoulli success/failure trial. We report success counts, success rates, Wilson 95% confidence intervals (CI), and a one-sided Fisher's exact test for the pairwise comparison against UniVLA, with $H_0 : p_{\mathrm{FlowVLA}} \leq p_{\mathrm{UniVLA}}$ and $H_1 : p_{\mathrm{FlowVLA}} > p_{\mathrm{UniVLA}}$. A smaller p-value means that the observed improvement is less likely to be explained by random rollout variation alone. As shown in Table IV, FlowVLA achieves statistically significant pooled improvements on both LIBERO and SimplerEnv.

For completeness, Table V provides the per-suite and per-task success counts used in the pooled analysis. The LIBERO evaluation uses 500 rollouts per suite, while the expanded SimplerEnv evaluation uses 96 rollouts per task under the same task protocol as the main SimplerEnv benchmark.

D. Reduced-Data Policy Evaluation

To directly evaluate policy performance under reduced data, we report task success rates on the SimplerEnv benchmark. We evaluate the 100%, 50%, and 25% BridgeData V2 training-data fractions and use the original 24-rollout-per-task SimplerEnv evaluation protocol for each fraction. For each data fraction, we fine-tune FlowVLA and UniVLA with the same training protocol while varying only the amount of expert demonstration data. As shown in Table VI, FlowVLA achieves higher average success rates than UniVLA in the reduced-data regimes, providing policy-level evidence that the benefit of motion-reasoning pre-training persists when downstream training data is reduced.

E. Extended Real-World Evaluations with Error Bars

In our core experiments, the reported success rates were initially derived from a single evaluation session consisting of 25 trials per task, following the standard evaluation protocols widely adopted in previous works [6], [7], [8].

To provide robust error bars and thoroughly evaluate the stability of our proposed method under real-world physical variations, we conducted 4 additional independent evaluation sessions for our approach. This brings the total to 5 independent evaluations (totaling 125 trials per task).

The extended results are presented in Table VII. For the baseline methods, we report the results from the standard single evaluation (25 trials). For our method, we report the **Mean** $\pm$ **Standard Deviation** over the 5 evaluations. Note that the mean values for our method reflect the average across

TABLE II: Performance comparison on the LIBERO benchmark with wrist-view cameras enabled. Both models utilize multi-view observations (global + wrist).

Method	Spatial	Object	Goal	Long	Avg
UniVLA [4] (w/ Wrist)	95.4	98.8	93.6	94.0	95.5
FlowVLA (Ours, w/ Wrist)	**96.6**	**99.0**	**94.4**	**94.6**	**96.2**

TABLE III: CALVIN ABC→D evaluation following the UniVLA protocol. We report success rates of completing at least 1–5 consecutive subtasks and the average completed sequence length.

Method	SR@1	SR@2	SR@3	SR@4	SR@5	Avg. Len
UniVLA [4]	0.989	0.948	0.890	0.828	0.751	4.41
FlowVLA (Ours)	**0.991**	**0.952**	**0.923**	**0.864**	**0.815**	**4.55**

TABLE IV: Fixed-sample statistical significance analysis on simulation policy evaluation. Each rollout is treated as a Bernoulli success/failure trial, and results are pooled over the four LIBERO suites or four tasks in the SimplerEnv benchmark. We report success counts, success rates (SR), Wilson 95% confidence intervals (CI), and one-sided Fisher exact-test p-values under $H_0 : p_{\mathrm{FlowVLA}} \leq p_{\mathrm{UniVLA}}$ and $H_1 : p_{\mathrm{FlowVLA}} > p_{\mathrm{UniVLA}}$. The p-value is a pairwise benchmark-level statistic comparing FlowVLA with UniVLA, not a per-method quantity.

Benchmark	UniVLA			FlowVLA (Ours)			p-value
	Count	SR	95% CI	Count	SR	95% CI	
LIBERO	1680/2000	84.0%	[82.3, 85.5]%	**1762/2000**	**88.1%**	**[86.6, 89.4]%**	$< 10^{-3}$
SimplerEnv	249/384	64.8%	[59.9, 69.5]%	**279/384**	**72.7%**	**[68.0, 76.9]%**	0.012

TABLE V: Per-suite and per-task success counts used for the fixed-sample statistical analysis.

Benchmark	Split / Task	UniVLA	FlowVLA (Ours)
LIBERO	Spatial	463/500 (92.6%)	**466/500 (93.2%)**
LIBERO	Object	469/500 (93.8%)	**475/500 (95.0%)**
LIBERO	Goal	433/500 (86.6%)	**458/500 (91.6%)**
LIBERO	Long	315/500 (63.0%)	**363/500 (72.6%)**
SimplerEnv	Put Spoon	60/96 (62.5%)	**67/96 (69.8%)**
SimplerEnv	Put Carrot	59/96 (61.5%)	**60/96 (62.5%)**
SimplerEnv	Stack Block	41/96 (42.7%)	**59/96 (61.5%)**
SimplerEnv	Put Eggplant	89/96 (92.7%)	**93/96 (96.9%)**

TABLE VI: Reduced-data end-to-end policy evaluation on the SimplerEnv benchmark. We report task success rates (%) under different fractions of BridgeData V2 fine-tuning data.

Data	Method	Put Spoon	Put Carrot	Stack Block	Put Eggplant	Avg.
100%	UniVLA	62.5	62.5	41.6	95.8	65.6
100%	**FlowVLA (Ours)**	**70.8**	62.5	**62.5**	**100.0**	**74.0**
50%	UniVLA	54.2	50.0	33.3	83.3	55.2
50%	**FlowVLA (Ours)**	**62.5**	**54.2**	**45.8**	83.3	**61.5**
25%	UniVLA	41.7	37.5	20.8	66.7	41.7
25%	**FlowVLA (Ours)**	**45.8**	**41.7**	**29.2**	**75.0**	**47.9**

all 125 trials. As demonstrated in the results, even when accounting for environmental variance and physical randomness, our approach consistently and significantly outperforms the baselines.

F. Robustness in Realistic Settings and the Choice of Motion Representation

In this section, we provide additional experiments and analyses to evaluate our model's performance in environments with complex backgrounds and to further justify our architectural choice of utilizing 2D optical flow rather than 3D scene flow.

1) Robustness to Realistic Backgrounds: To verify our model's performance in more realistic settings beyond simple clean backgrounds, we conducted additional real-world experiments using patterned tablecloths (e.g., pink zigzag and grey grid patterns) to simulate complex visual textures (see Figure 1).

- **Results:** The RAFT-based optical flow estimation remained robust despite the background variations. The model successfully completed manipulation tasks with reliable performance comparable to the white-background setting, demonstrating that background texture changes do not negatively affect the flow estimation or policy execution.

2) Scene Flow vs. Optical Flow: While Scene Flow (3D motion) theoretically offers better geometric generalization, in-

TABLE VII: Results on the Real-world AgileX Cobot platform. We report the task success rates (%). For our method, we conducted 5 independent evaluations (25 trials each) to report **Mean** $\pm$ **Standard Deviation**, demonstrating stability. Baseline results are from a single evaluation (25 trials). Best results are in **bold**.

Model	Stack Bowls (single-arm)	Place Vegetable (single-arm)	Place Bottles (bimanual)	Lift Pot (bimanual)	Avg.
ACT [9]	32.0	24.0	12.0	8.0	19.0
OpenVLA [6]	28.0	20.0	20.0	12.0	20.0
UniVLA [4]	48.0	40.0	16.0	20.0	31.0
FlowVLA (Ours)	$\mathbf{53.6 \pm 7.7}$	$\mathbf{57.6 \pm 6.6}$	$\mathbf{32.0 \pm 6.3}$	$\mathbf{25.6 \pm 6.1}$	**42.2**

corporating full 3D scene flow is architecturally incompatible with our current VLA backbone, primarily due to the absence of a mature discrete tokenizer for scene flow. To approximate scene flow, we experimented with an interleaved sequence of RGB, Depth, and Optical Flow tokens.

This approach led to a performance drop rather than an improvement. On the long-horizon LIBERO-10 benchmark [1], the success rate decreased from **72.6%** (pretrained with Optical Flow only) to **70.4%** (pretrained with Optical Flow + Depth).

Our investigation revealed that this performance degradation stems primarily from a critical **context length bottleneck**. Crucially, adding depth tokens significantly increases the sequence length. Given the memory constraints of high-performance GPUs (e.g., 80GB A100), we were forced to reduce the temporal context window to fit the model in memory. This reduction severely impaired the World Model's ability to model **long-horizon dependencies**, which is critical for complex tasks like those in LIBERO-10.

Consequently, we conclude that maximizing temporal context with 2D flow is more beneficial for overall policy performance than explicitly modeling 3D geometry that sacrifices the history length.

G. Quantitative Evaluation of World Model Generation

In Section IV.C of the main paper, we present qualitative visualizations demonstrating the predictive capabilities of our world model. To provide a more comprehensive and objective assessment, we supplement these with quantitative evaluations in this section.

Closed-loop Autoregressive Rollout. We first clarify that both the visualizations in the main text and the quantitative metrics reported here are based on **closed-loop autoregressive rollouts**. In this recursive setting, the model's own predictions are iteratively fed back as inputs for subsequent steps. This evaluation protocol rigorously tests the model's stability and robustness against compounding errors over long time horizons.

Quantitative Results. To measure the perceptual quality of the generated future frames, we employ the Learned Perceptual Image Patch Similarity (LPIPS) metric. The evaluation is conducted on 100 randomly selected samples from the validation set under the aforementioned closed-loop rollout setting.

As shown in Table VIII, our FlowVLA achieves a significantly lower LPIPS score (0.1333) compared to the UniVLA baseline (0.1944). This substantial improvement indicates that our explicit flow-guided generation mechanism effectively mitigates blurring and artifacts, producing future frames with higher perceptual fidelity even during extended closed-loop generation.

TABLE VIII: Quantitative evaluation of World Model generation quality on 100 randomly selected samples from the validation set. **LPIPS** ($\downarrow$) is measured using closed-loop autoregressive rollouts.

Model	LPIPS $\downarrow$
UniVLA (Baseline)	0.1944
FlowVLA (Ours)	**0.1333**

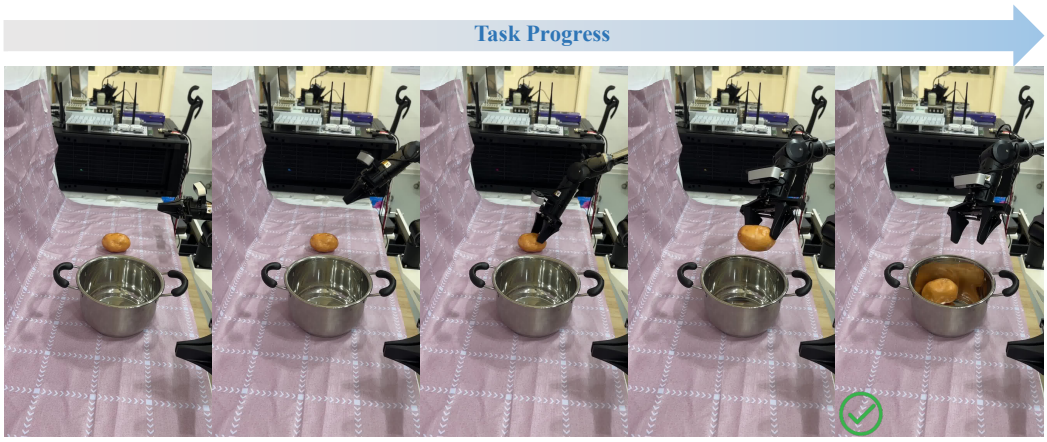
H. Ablation Study on Intra-Frame Attention Masking

In the design of our visual Chain of Thought (CoT), we employ a strict causal (lower-triangular) attention mask across the entire interleaved sequence of tokens. An alternative, theoretically sound approach for modeling simultaneous spatial data is to utilize a non-step-wise (bidirectional) mask within each individual image or flow block. The rationale behind this is that within a single spatial frame, there is no inherent temporal causality assumption among the patches (as explored in recent tokenization literature).

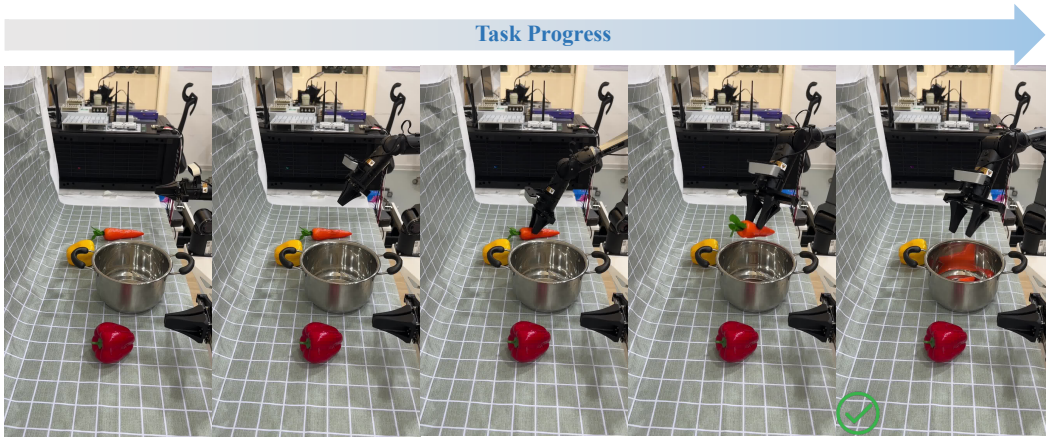
To evaluate the impact of this architectural choice, we conducted an ablation study implementing a bidirectional intra-frame mask throughout both the World Model pre-training and the downstream policy fine-tuning stages. The results on the SimplerEnv benchmark [10] are presented in Table IX.

Contrary to the theoretical intuition, replacing the strict causal mask with a bidirectional intra-frame mask leads to a noticeable drop in downstream task execution success rates (from 74.0% to 55.2% on average). We attribute this performance degradation primarily to a **mismatch with the foundation attention priors**.

Our backbone model, Emu3 [11] (8B parameters), is natively pre-trained under a strict Next-Token Prediction (NTP) paradigm using a standard causal mask. The model's attention matrices are heavily optimized for this unidirectional structure over massive-scale foundation pre-training data. Introducing bidirectional intra-frame attention creates a structural distribution shift in the attention mechanism. While the model can partially adapt to this new mask layout, our training scale is relatively small compared to Emu3's original web-scale pre-training (which encompasses the massive Aquila language corpus and millions of filtered image/video pairs). Consequently, the model struggles to fully realign its deeply entrenched structural priors, leading to suboptimal policy performance.



(b) Task: “Picking up a potato and placing it into a pot."



(b) Task: "Picking up a carrot and placing it into a pot."

Fig. 1: **Robustness Validation of our FlowVLA in Realistic Settings.** To address concerns about simple white backgrounds, we tested our model in complex real-world environments. **(a)** Shows successful execution on a pink patterned tablecloth. **(b)** Shows successful execution on a grey grid tablecloth. These results demonstrate that our flow-based policy generalizes well beyond simple clean backgrounds.

Thus, we maintain the global causal masking mechanism in our final FlowVLA architecture.

III. In-depth Methodological Analysis

A. Clarification on the Terminology: Causal Motion Reasoning

While there is an ongoing debate in the LLM community regarding whether Chain-of-Thought (CoT) constitutes true reasoning or merely mimics logical patterns, the context of our FlowVLA framework differs fundamentally. Unlike the abstract and often ambiguous nature of linguistic reasoning in LLMs, our intermediate flow prediction is grounded in **physical causality**. We frame this process (Equation 5 in the main text) as “Reasoning” specifically from the perspective of **Causal Motion Reasoning**:

- **Explicit Physical Causality:** Standard video generation models often act as a “black box”, directly hallucinating future pixels ($v_t \rightarrow v_{t+1}$) without necessarily understanding the underlying motion dynamics. In contrast, FlowVLA is forced to decouple **dynamics** (how things move) from **appearance** (what things look like). We define this intermediate flow prediction ($v_t \rightarrow f_t$) as “Reasoning” because the model explicitly infers the **causal mechanism** of motion—specifically the vector field that physically transports pixels—before generating the future state ($f_t \rightarrow v_{t+1}$). This structural constraint ensures the model derives the effect (next frame) from its cause (motion), rather than merely fitting statistical correlations.
- **Evidence of Causal Dependency (Counterfactual Intervention Experiment):** To validate that the flow serves as a necessary reasoning step rather than a passive feature, we conducted a **counterfactual intervention experiment**, which is visualized in the main manuscript.

TABLE IX: Comparison of intra-frame masking strategies on the SimplerEnv benchmark [10]. Replacing the strict causal mask with a bidirectional mask (applied to both World Model pre-training and fine-tuning) leads to a noticeable decrease in task success rates (%), primarily due to the structural mismatch with the foundation model's pre-trained attention priors.

Masking Strategy	Put Spoon	Put Carrot	Stack Block	Put Eggplant	Avg.
Bidirectional Intra-frame	54.2	50.0	29.2	87.5	55.2
Strict Causal (Ours)	**70.8**	**62.5**	**62.5**	**100.0**	**74.0**

We artificially inverted the ground-truth optical flow (GT flow) tokens before feeding them into the video generation head. Crucially, we observed that the model's predicted next-frame RGB **strictly followed the inverted flow**, generating a future state where the robot arm and object moved in the opposite direction to the ground truth image (GT image). This confirms that the model is not retrieving memorized video statistics but is actively **deriving the result based on the intermediate flow**, effectively treating the flow as a critical reasoning bridge in the causal chain.

B. Clarification on Optical Flow and Motion Dynamics

It is important to distinguish between kinematic representations and full dynamic descriptions. We acknowledge that optical flow itself is a kinematic representation (velocity) rather than a complete dynamic description involving mass, forces, and contact.

Therefore, we clarify that **we do not equate optical flow with motion dynamics.** Instead, our core philosophy is to utilize optical flow as a **"guidance medium"** or a **"strong supervision signal"** to facilitate the World Model in capturing underlying motion dynamics.

Since physical forces and contacts are invisible in standard video data, they cannot be modeled directly without specialized sensors. However, these dynamic interactions can be visually observed as pixel-level motion, which is captured by optical flow. By explicitly tokenizing and predicting optical flow, we force the model to focus on the **consequences of dynamics**. This acts as a proxy task, enabling the model to implicitly learn the latent physical rules governing interactions. This capability is validated by the counterfactual intervention experiment presented in Section III-A and visualized in the main manuscript.

To further illustrate the importance of this distinction, we can contrast our approach with recent methods that *do* treat flow directly as dynamics, such as *3DFlowAction* [12] discussed in Section IV-C. By feeding predicted flow as a hard constraint into an optimization solver, *3DFlowAction* implicitly assumes the generated flow is geometrically perfect. Consequently, as shown in our failure analysis (**Figure 4(c)**), even minor prediction noise or geometric inconsistencies (which are inevitable in generative models) can cause their solver to calculate infeasible actions, leading to misalignment issues (e.g., failing to stack the green bowl precisely onto the yellow bowl). In contrast, by using flow merely as a supervision signal for latent learning, our model avoids this fragility, learning to extract robust control intent even from noisy visual predictions.

C. Detailed Analysis of the Decoupled Training Paradigm

In this section, we provide a detailed discussion on our choice to completely decouple flow pre-training from action post-training, rather than integrating visual Chain-of-Thought (VCoT) into a joint training pipeline with reasoning dropout (similar to the ECoT-Lite [13] strategy). We detail the rationale behind this design choice from two perspectives:

- **Why post-train exclusively on action tokens?** Our decision to post-train exclusively on action tokens is driven by the principle of **representation internalization**. During the flow pre-training stage, the backbone has already internalized robust physical priors and motion dynamics into its hidden representations. By explicitly leaving out the visual CoT generation during post-training, we enable the policy to better focus its representational capacity on precisely fitting the action distribution. This allows the model to deeply benefit from the learned dynamics implicitly, without being bottlenecked by explicit pixel-level generation.
- **Why avoid full CoT with reasoning dropout?** While techniques like loss weighting could partially address multi-task conflicts, we explicitly avoided the "Reasoning Dropout" strategy (jointly training on $v \to f \to a$ with random dropout) due to the severe **token imbalance and attention distraction** inherent to high-dimensional visual generation.
  Specifically, ECoT-Lite was originally designed for textual reasoning, where intermediate tokens are sparse (typically a few dozen tokens). In contrast, generating a visual CoT (optical flow) requires hundreds of discrete visual tokens per step. If we adopted reasoning dropout, during the non-dropped steps, inserting this massive block of flow tokens between the observation v and action a would cause significant *attention distraction*—the action tokens would have to attend over a huge intermediate visual context, diluting their direct grounding on the original observation. Furthermore, the sheer volume of flow tokens would overwhelmingly dominate the backbone's gradient updates, destabilizing the learning of the sparse action signals.
  Empirical Validation. To validate this architectural bottleneck, we implemented the "Reasoning Dropout" strategy (jointly training $v \to f \to a$ with 50% flow dropout) using our backbone and evaluated it on the LIBERO benchmark. As shown in Table X, this joint training strategy suffers a significant performance drop compared to our decoupled approach (**80.9% vs. 88.1%** average success rate). This degradation directly validates our theoretical analysis: forcing the model to balance massive

visual generation with sparse action prediction within the same training phase leads to suboptimal policy learning. Therefore, completely decoupling flow pre-training from action post-training serves as a fundamentally more stable paradigm.

D. Analysis of Shared VQ-GAN for Optical Flow Tokenization

In our FlowVLA framework, we utilize a frozen VQ-GAN [14], originally pre-trained on regular RGB images, to tokenize both RGB observations and optical flow maps. This design choice is driven by both theoretical representation benefits and reconstruction performance.

Unified Token Space. Our primary motivation for sharing a frozen VQ-GAN is to maintain a unified token space. Because our Transformer backbone models the joint distribution of RGB and flow tokens ($v \rightarrow f \rightarrow v$), sharing the same codebook ensures that semantically similar structural features—such as an object edge in an RGB image and a corresponding motion boundary in an optical flow map—are mapped to compatible latent representations. Fine-tuning a separate VQ-GAN exclusively for optical flow would create a disjoint latent space, potentially making the cross-modal prediction task ($v_t \rightarrow f_t$) significantly harder for the Transformer to learn. Thus, we prioritize a unified representation over marginal domain-specific reconstruction gains.

Reconstruction Capability. To process the flow, we map the 2-channel flow vectors into a 3-channel image-like format. While these maps differ statistically from natural photographs, they retain strong spatial correlation and smoothness—core features that the VQ-GAN is well-equipped to encode. The encoder effectively treats them as "texture-less" structural maps. To validate whether the RGB-pretrained VQ-GAN is suitable for this domain shift, we evaluated its reconstruction error on our processed flow maps (averaged over 50,000 samples). As shown in Table XI, the Mean Squared Error (MSE) is 5.27×10^{-4}. According to recent studies on image tokenization [15], an error of this magnitude is considered extremely low, indicating that the tokenized representation preserves the structural fidelity of the optical flow with negligible artifacts for downstream tasks.

Marginal Gains from Fine-Tuning. We also investigated the potential benefits of explicitly fine-tuning the VQ-GAN on flow data. Our observations indicate that the improvement is marginal. Specifically, fine-tuning the VQ-GAN exclusively on flow maps yields only an approximate **5% relative reduction** in reconstruction error, lowering the MSE slightly from 5.27×10^{-4} to 5.01×10^{-4}. Given this minimal gain, the architectural simplicity and cross-modal alignment provided by the shared, frozen VQ-GAN prove to be the more effective strategy.

E. Robustness of Optical Flow on Textureless and Specular Objects

A potential concern with explicitly relying on optical flow is the performance degradation of flow estimators in challenging visual conditions, such as environments containing textureless, transparent, or highly specular (reflective) objects. While textureless surfaces are traditionally difficult for local pixel matching algorithms, modern flow estimators like RAFT [16] mitigate this issue by leveraging global context and iterative refinement to propagate motion from structural boundaries to the interior regions.

To empirically demonstrate this robustness within our pipeline, we evaluated our model on manipulating a stainless steel pot—an object that is not only textureless but also highly specular, representing a severe edge case for optical flow estimation (see Figure 2).

As shown in the flow visualization, despite the pot's smooth metallic surface lacking internal texture, the model accurately predicts a dense and complete flow mask across the entire object. This confirms that our flow-based policy does not fail on textureless objects; instead, it effectively infers motion from the object's visible boundaries and the gripper's movement, providing reliable and continuous motion representations for the downstream policy.

IV. Discussions on Scaling and Baselines

A. Analysis on Model and Data Scaling

A critical question regarding our proposed method is whether simply scaling up the pre-training data or model size could naturally resolve the limitations of direct next-frame prediction, thereby rendering the explicit inductive bias of optical flow unnecessary. To comprehensively address this, we structure our analysis in two parts: evaluating a state-of-the-art large-scale video model and conducting a systematic data scaling law experiment.

1) Qualitative Evaluation on Large-Scale Pre-trained Models: First, to verify whether scaling up the pre-training alone can solve the limitations of next-frame prediction, we tested NVIDIA Cosmos-Predict2.5 [17], a state-of-the-art large-scale video generation model. As shown in Figure 3, Cosmos still exhibits the specific failure modes we identified:

- **Physical Inconsistency:** We observe noticeable morphological deformation of the robot arm and gripper across the sequence, struggling to maintain rigid-body constraints.
- **Semantic/Dynamics Misalignment:** The model fails to precisely follow fine-grained spatial instructions (e.g., placing a block *next to* instead of "on top of").

This suggests that while massive pre-training improves visual fidelity, it does not automatically yield the **fine-grained physical reasoning** required for robotics.

2) Data Scaling Law: UniVLA vs. FlowVLA: Ideally, to isolate the effect of our flow inductive bias at a massive scale, we would integrate our Visual CoT directly into Cosmos. However, Cosmos employs a bidirectional attention DiT architecture, which fundamentally differs from our causal autoregressive framework, making direct integration and fair ablation infeasible.

Therefore, as a controlled alternative to investigate whether scaling up data diminishes the utility of our flow inductive bias, we resorted to our autoregressive baseline, UniVLA [4]. We conducted a systematic **Data Scaling Law** experiment on the Bridge-V2 dataset [2] by varying the amount of training data. Specifically, we evaluated the generation quality

TABLE X: Comparison of training strategies on the LIBERO benchmark. Our decoupled two-stage paradigm (FlowVLA) significantly outperforms the joint training with Reasoning Dropout strategy (ECoT-Lite style) across all task suites.

Training Strategy	Spatial	Object	Goal	Long	Avg.
Reasoning Dropout (ECoT-Lite style)	86.4	89.2	83.6	64.2	80.9
FlowVLA (Ours, Decoupled)	**93.2**	**95.0**	**91.6**	**72.6**	**88.1**

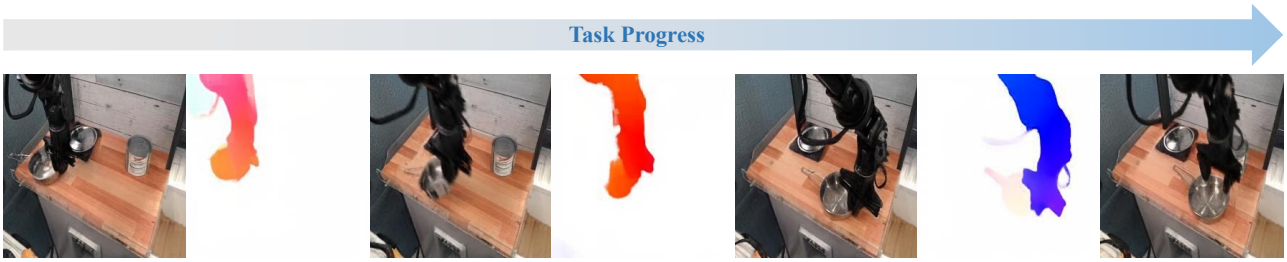
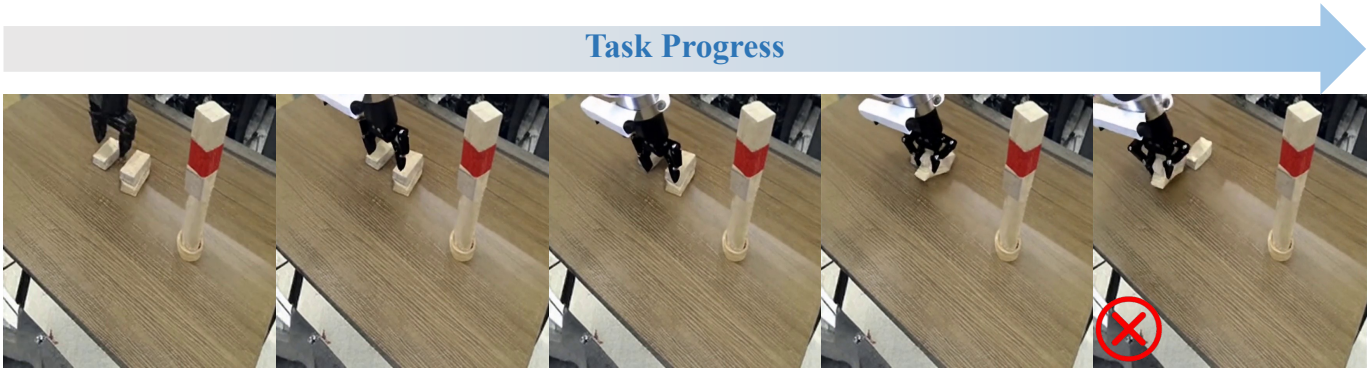


Fig. 2: **Robustness on Textureless/Reflective Objects.** We visualize the optical flow predictions while manipulating a stainless steel pot. Despite the object being textureless and metallic, RAFT accurately predicts dense flow across the entire object surface, demonstrating that the model effectively propagates motion cues from boundaries to textureless regions.



(a) Task: "Put the rectangular on top of the rectangular block next to it."

(b) Task: "Move toy diagonally little bit top on the right side."

Fig. 3: **Failure Cases of Large-Scale Video Generation Model (NVIDIA Cosmos-Predict2.5).** Despite being trained on massive datasets, the model still exhibits significant issues in physical reasoning for robotic manipulation. (a) The model fails to follow the instruction "on top of", placing the block *next to* the target instead. (b) The generated motion does not align with the "top right" directive. Crucially, in both cases, we observe morphological deformation of the robot arm (e.g., the gripper shape distorts over time), indicating a lack of rigid-body dynamics understanding. This suggests that simply scaling up visual generation does not automatically solve the need for explicit dynamics modeling (e.g., optical flow) in robotics. For visualizations of correct predictions generated by our proposed method, please refer to Figure 8 and Figure 9 in the main manuscript.

TABLE XI: Reconstruction Loss (MSE) of the frozen RGB-pretrained VQ-GAN on Flow maps.

Modality	Reconstruction MSE (normalized) ↓
Flow Maps	5.27×10^{-4}

of the World Model (measured by **LPIPS** ↓ via closed-loop autoregressive rollouts) using 10%, 50%, and 100% of the training data for both UniVLA and our FlowVLA.

TABLE XII: Data Scaling Analysis of World Model generation quality on the Bridge-V2 dataset. **LPIPS** (↓) is measured using closed-loop autoregressive rollouts. Lower is better.

Model	10% Data	50% Data	100% Data
UniVLA	0.2840	0.2250	0.1944
FlowVLA (Ours)	**0.2316**	**0.1825**	**0.1333**

As shown in Table XII, our FlowVLA demonstrates **significant sample efficiency** and highly efficient data utilization. Notably, **FlowVLA trained on 50% of the data achieves a better LPIPS score (0.1825) than the UniVLA baseline trained on the full 100% dataset (0.1944)**. Furthermore, our FlowVLA's performance at 10% (0.2316) is comparable to UniVLA at 50% (0.2250). This confirms that explicitly modeling optical flow acts as a powerful structural prior, dramatically reducing the amount of data required to learn robust world dynamics.

In conclusion, our experiments confirm that explicit modeling of dynamics (optical flow) remains a necessary and highly efficient inductive bias that cannot be easily bypassed by simply scaling up pre-training parameters or downstream data.

B. Clarification on the Architectural Baseline and Core Contributions

In Section IV.E of the main text, we presented an ablation study demonstrating the impact of our proposed components. We would like to clarify that, functionally, removing the entire Visual Chain-of-Thought (V-CoT) component aligns our ablated model with the general architecture of standard world model-based VLAs, of which UniVLA [4] is a representative example (excluding the use of wrist cameras).

Despite this structural similarity when our proposed module is ablated, we emphasize that building upon strong, established base models is a **common and effective practice** in the VLA community. This approach allows researchers to isolate and rigorously evaluate the specific impact of their proposed innovations. For example, recent prominent works follow a similar trajectory:

- *OpenVLA-OFT [8] (Robotics: Science and Systems 2025)* builds upon OpenVLA [6].
- *FAST: Efficient Action Tokenization for Vision-Language-Action Models [18] (Robotics: Science and Systems 2025)* builds upon π_0 [19].
- *3D Diffusion Policy [20] (Robotics: Science and Systems 2024)* builds upon Diffusion Policy [7].

Our core contribution does not lie in proposing an entirely new backbone from scratch, but rather in introducing the **Visual Chain-of-Thought (V-CoT) paradigm**. We propose a novel method to transform a generic World Model into a **motion-causality reasoning engine** by explicitly modeling optical flow as an intermediate reasoning step. This paradigm shift makes generic World Models significantly more effective when fine-tuned as a policy model. Crucially, this methodology is **model-agnostic** and can be applied to other World Model architectures to enhance their understanding of physical dynamics. Our extensive experiments demonstrate the effectiveness of this approach in both simulation and real-world environments.

C. Extended Comparison with Related Flow-based Methods

While optical flow or scene flow is a well-established tool in computer vision for velocity estimation and prediction, we emphasize that the core contribution of our FlowVLA is not merely the usage of optical flow. Rather, it is the proposal of a "Causal Reasoning Paradigm" for training World Models, where optical flow serves as a specific vehicle to realize this causality.

Crucially, we posit that possessing such causal reasoning capability is a prerequisite for a World Model to serve as a reliable backbone for Vision-Language-Action (VLA) models. Unlike standard video generators that may hallucinate physically implausible motions based on statistical correlations, a causal-aware backbone ensures that the VLA model grounds its planning in realistic dynamics expressed by the optical flow, which is fundamental for safe and effective robot interaction. Below, we provide a detailed comparison to demonstrate why our method is fundamentally different from and superior to recent flow-based works for general robot learning.

1) Comparison with FlowDreamer: **What FlowDreamer Does:** *FlowDreamer* [21] adopts a modular, two-stage pipeline for video prediction. It first employs a U-Net to explicitly predict 3D scene flow from RGB-D inputs. In the second stage, it uses a separate Diffusion model, conditioned on the predicted flow and the ground-truth action, to render the next frame. Essentially, it treats flow as an intermediate visual feature to guide the texture generation of a diffusion model.

Limitations of FlowDreamer: This approach has critical limitations regarding its potential as a VLA backbone:

- **Lack of Causal Reasoning:** In their own ablation study, the authors noted that when they inverted the flow input, the robot did not take contrary actions, because the generation was still dominated by the ground-truth action condition. This indicates the model treats flow merely as auxiliary information rather than a causal driver.
- **Action Dependency:** It is strictly an action-conditioned simulator ($State + Action \rightarrow NextState$) and cannot generate actions itself, making it unsuitable for autonomous decision-making.
- **Data Constraint:** It relies heavily on RGB-D data and explicit 3D flow, restricting its scalability on massive real-world datasets where perfect depth is often unavailable.

Difference with FlowDreamer: In contrast, we propose a **Unified Autoregressive Paradigm**. Instead of a multi-stage

pipeline, we tokenize optical flow and RGB images into a shared space and model them jointly. We treat optical flow not just as visual guidance, but as the causal bridge between states. Crucially, our model operates on standard RGB videos by estimating the optical flow, removing the dependency on depth or actions during inference.

Our Unique Advantages:

- **True Causal Understanding:** As demonstrated by the counterfactual intervention experiment in Section III-A, our model actively derives the predicted next frame based on the intermediate flow, effectively treating it as a critical reasoning bridge in the causal chain.
- **VLA-Ready Backbone:** By unifying perception and dynamics, our model is inherently capable of predicting actions (as tokens), serving as a decision-making brain rather than just a passive simulator.

2) Comparison with 3DFlowAction: **What 3DFlowAction Does:** *3DFlowAction* [12] proposes a hierarchical framework that first generates 3D optical flow trajectories conditioned on language. Crucially, it does not predict robot actions directly. Instead, it treats the generated flow as a geometric constraint and uses a separate optimization-based policy (involving AnyGrasp, rigid transforms, and Inverse Kinematics (IK) solvers) to calculate the robot actions required to track this flow.

Limitations of 3DFlowAction: This approach relies on a complex pipeline rather than an end-to-end learning model. We validated the fragility of this system in real-world experiments (see Figure 4):

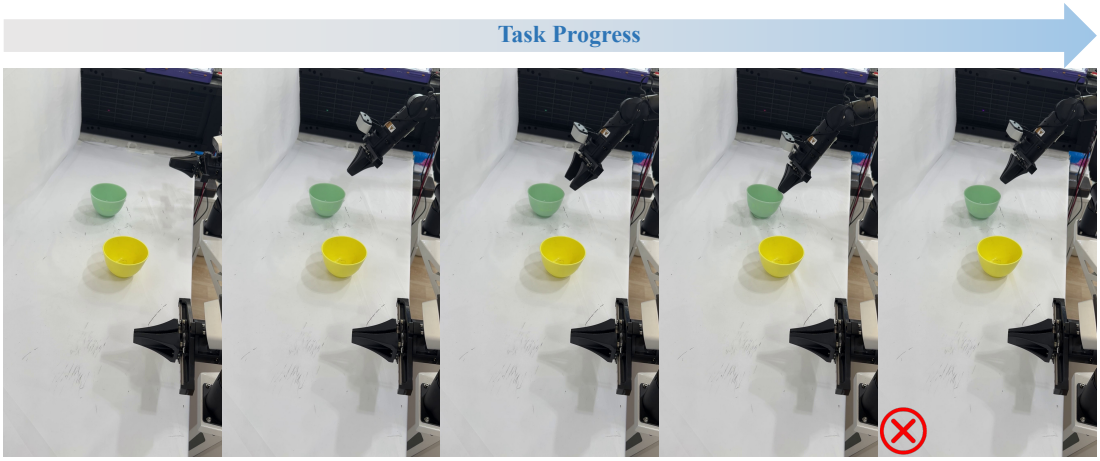
- **Fragile Multi-Stage Pipeline (Error Propagation):** The system depends on a concatenation of external modules (Sensors, Detectors, Flow Predictors, Solvers). Failure in any single module causes the entire chain to break. We validated this based on a typical task: "Grasping a green bowl and stacking it onto a yellow bowl" (Figure 4).
  - *Grasping Failure:* As shown in Figure 4(a) and (b), the reliance on raw depth maps for the initial grasp makes it highly sensitive to sensor noise, leading to inaccurate pose estimation and causing the robot to miss the object.
  - *Placement Failure:* As shown in Figure 4(c), the downstream solver treats predicted flow as a hard constraint. Slight inaccuracies in the flow prediction propagate to the Inverse Kinematics (IK) solver, causing the robot to release the object at the wrong location.
- **Lower Inference Efficiency:** The reliance on an iterative "Verify-Loop" with GPT-4o [22] and optimization solvers creates high latency, rendering it less suitable for high-frequency control scenarios.

Difference with 3DFlowAction: In contrast, we propose a **unified VLA backbone** trained under a Causal Reasoning Paradigm. Instead of separating flow generation and action solving, we tokenize optical flow, RGB images, and **actions** into a single autoregressive Transformer. This allows our model to learn the joint distribution of visual dynamics and robot control in a shared latent space, without relying on external solvers.

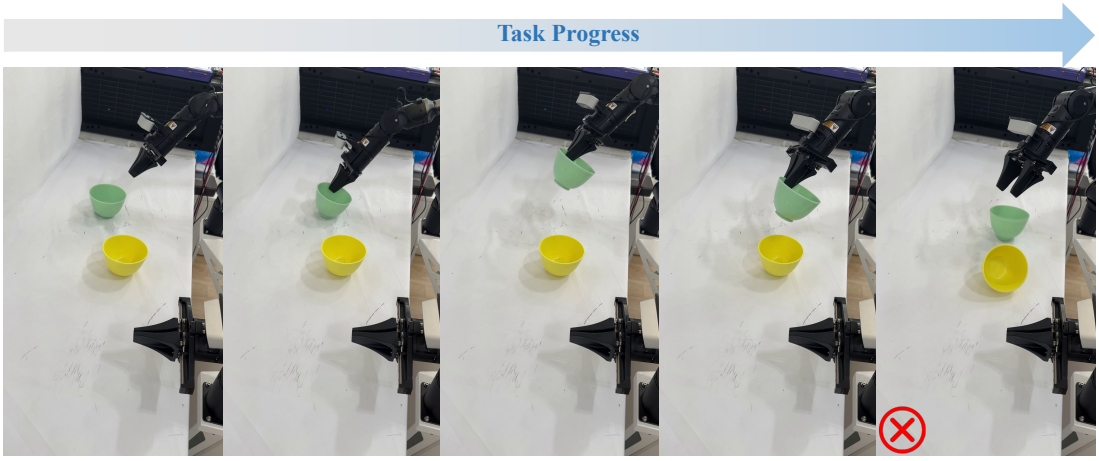
Our Unique Advantages:

- **End-to-End VLA Capability:** Unlike *3DFlowAction*'s fragmented pipeline, our model is a single Transformer that learns motion representations directly from data, enabling efficient end-to-end learning.
- **Unified Architecture for Scalability:** Unlike 3DFlowAction, which often rely on separate, heavy modules or complex 3D representations, our FlowVLA adopts a unified tokenization strategy. We discovered that flow maps can share the same VQ-GAN encoder/decoder with RGB images, allowing them to be seamlessly integrated into a single autoregressive transformer. This structural simplicity means our method adds causal supervision without breaking the standard VLM architecture, making it more accurate and robust in complex manipulation tasks.

(a) **Noisy Depth Input.** Real-world sensors often produce missing values (black regions) and noise at object boundaries.



(b) **Grasping Failure.** Relying on the noisy depth map for analytic grasping leads to estimation errors and missing the object.



(c) **Placement Failure.** Slight inaccuracies in flow prediction propagate to the rigid solver, causing the robot to drop the object.

Fig. 4: **Failure Analysis of the Modular Pipeline (3DFlowAction [12]).** We validated that the 3DFlowAction's non-end-to-end design suffers from critical error propagation. **(a)** Shows the raw depth input with significant noise. **(b)** and **(c)** show how this sensor noise and flow prediction errors propagate through the pipeline, leading to task failures in both grasping and placement stages. In contrast, as demonstrated in Figure 7(a) of the manuscript, our end-to-end model is highly robust to these imperfections and successfully completes the task.

REFERENCES

[1] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 44 776–44 791, 2023.

[2] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning*. PMLR, 2023, pp. 1723–1736.

[3] T. Dao, "Flashattention-2: Faster attention with better parallelism and work partitioning," *arXiv preprint arXiv:2307.08691*, 2023.

[4] Y. Wang, X. Li, W. Wang, J. Zhang, Y. Li, Y. Chen, X. Wang, and Z. Zhang, "Unified vision-language-action model," *arXiv preprint arXiv:2506.19850*, 2025.

[5] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7327–7334, 2022.

[6] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "OpenVLA: An open-source vision-language-action model," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=ZMnD6QZAE6

[7] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[8] M. J. Kim, C. Finn, and P. Liang, "Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success," in *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025.

[9] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[10] X. Li, K. Hsu, J. Gu, K. Pertsch, O. Mees, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani, S. Levine, J. Wu, C. Finn, H. Su, Q. Vuong, and T. Xiao, "Evaluating real-world robot manipulation policies in simulation," *arXiv preprint arXiv:2405.05941*, 2024.

[11] X. Wang, X. Zhang, Z. Luo, Q. Sun, Y. Cui, J. Wang, F. Zhang, Y. Wang, Z. Li, Q. Yu *et al.*, "Emu3: Next-token prediction is all you need," *arXiv preprint arXiv:2409.18869*, 2024.

[12] H. Zhi, P. Chen, S. Zhou, Y. Dong, Q. Wu, L. Han, and M. Tan, "3dflowaction: Learning cross-embodiment manipulation from 3d flow world model," *arXiv preprint arXiv:2506.06199*, 2025.

[13] W. Chen, S. Belkhale, S. Mirchandani, O. Mees, D. Driess, K. Pertsch, and S. Levine, "Training strategies for efficient embodied reasoning," *arXiv preprint arXiv:2505.08243*, 2025.

[14] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.

[15] J. Shen, K. Tirumala, M. Yasunaga, I. Misra, L. Zettlemoyer, L. Yu, and C. Zhou, "Cat: Content-adaptive image tokenization," *arXiv preprint arXiv:2501.03120*, 2025.

[16] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European conference on computer vision*. Springer, 2020, pp. 402–419.

[17] A. Ali, J. Bai, M. Bala, Y. Balaji, A. Blakeman, T. Cai, J. Cao, T. Cao, E. Cha, Y.-W. Chao *et al.*, "World simulation with video foundation models for physical ai," *arXiv preprint arXiv:2511.00062*, 2025.

[18] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine, "FAST: Efficient Action Tokenization for Vision-Language-Action Models," in *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025.

[19] K. Black, N. Brown, D. Driess, A. Esmail, M. R. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, L. Smith, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, "π_0: A Vision-Language-Action Flow Model for General Robot Control," in *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025.

[20] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024.

[21] J. Guo, X. Ma, Y. Wang, M. Yang, H. Liu, and Q. Li, "Flowdreamer: A rgb-d world model with flow-based motion representations for robot manipulation," *IEEE Robotics and Automation Letters*, vol. 11, no. 3, pp. 2466–2473, 2026.

[22] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.