Vision-Language Models Meet Video Diffusion: Plug-and-Play World Models for Scalable Embodied Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

World models have emerged as powerful tools for planning and decision-making in embodied agents, especially with recent advances in diffusion-based video generation that enable high-fidelity future prediction. However, challenges in scaling high-resolution, instruction-following video generation and bridging simulated predictions to real-world robot control hinder practical deployment. To address these, we introduce a novel framework that repurposes video diffusion models as zero-shot policy generators conditioned on abstract action primitives dynamically derived from language instructions via large vision-language models. By extracting 6-DoF end-effector trajectories from predicted video frames and enabling closed-loop execution with observation feedback, our method maintains interpretability while avoiding error accumulation. This modular approach leverages the inductive biases of video generation and vision-language reasoning, allowing plug-and-play generalization across tasks. Furthermore, it enables policy distillation from unannotated visual demonstrations–ranging from simulated rollouts to real-world or web-sourced videos–helping mitigate data scarcity in embodied learning.

1 Introduction

Embodied agents capable of planning and decision-making in complex environments rely heavily on internal representations of the world, commonly referred to as *world models*. Recent advances in video generation–particularly diffusion-based models–have enabled high-fidelity visual future prediction, revitalizing interest in world model-centric architectures. However, two critical challenges persist: 1) Scalability-Complexity Trade-off: High-resolution, long-horizon video generation with fine-grained instruction following demands immense computational overheads, limiting real-time deployment; 2) Zero-Shot Policy Gap: Bridging simulated visual futures to real-world robot actions without task-specific adaptation remains an open problem. On the other hand, while hierarchical architectures have demonstrated success in end-to-end vision-language-action models by decoupling high-level planning from low-level action prediction (e.g., SayCan), such structured designs remain relatively underexplored in the context of world models. Furthermore, end-to-end paradigms–such as RT-1/RT-2–that directly map vision and language to actions entangle high-level semantic reasoning with low-level motor/end-effector dynamics. This design burdens a single model with multiple abstraction levels, leading to poor data efficiency, limited explainability, and weak generalization across tasks or morphologies.

To address these limitations, we propose a modular and hierarchical framework that reinterprets video diffusion models as **primitive-level world models** for zero-shot policy generation. Rather than attempting to learn task-level action distributions directly, our method decomposes complex

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

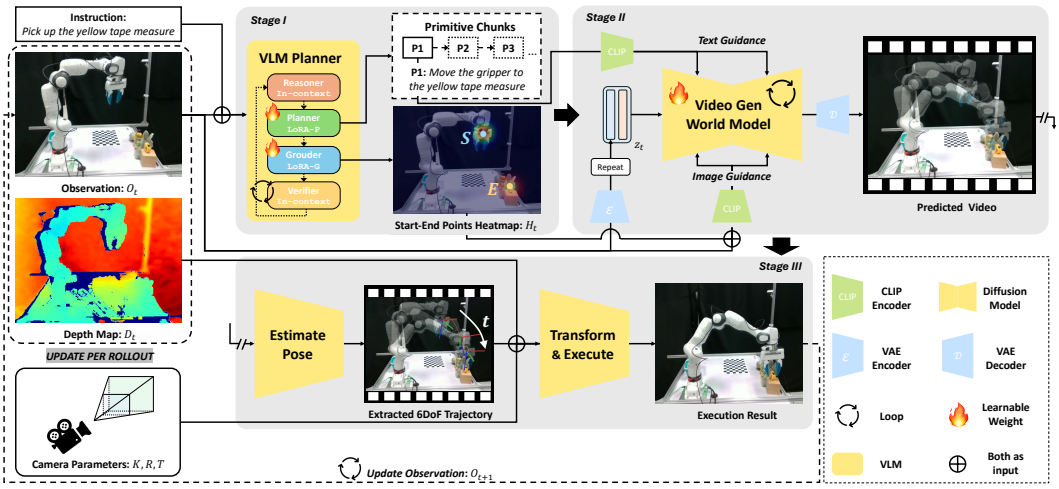

Figure 1: This workflow diagram illustrates a video generation framework structured into three stages: Stage 1 employs a VLM planner, comprising a VLM and two learnable LoRAs, along with a Reasoner and Verifier to ensure the efficiency and accuracy of the planning process. Stage 2 utilizes CLIP and a VAE encoder to generate predicted videos through the Video Generation World Model. Stage 3 focuses on pose estimation for the robotic arm, executing coordinate transformations to produce executable commands for real-world implementation.

instructions into semantically aligned atomic units–**action primitives**–and constrains video generation to short-horizon rollouts grounded in each primitive. For a quick overview of our approach, please refer to Figure 1.

At the heart of our system is a biologically inspired *cerebro-cerebellar hierarchy*: 1) A *cerebral* planner, instantiated as a vision-language foundation model (VLM), parses natural language and decomposes high-level tasks into a sequence of interpretable primitives. This narrows the predictive horizon per rollout, significantly reducing the computational burden otherwise associated with long-horizon diffusion; 2) A *cerebellar* world model, realized via a lightweight video diffusion model, predicts visual futures conditioned on each primitive, enabling controllable and interpretable rollout. This aligns the semantic granularity of visual prediction with the logical structure of instructions, allowing world models to focus on local dynamics without losing semantic context.

To operationalize this hierarchy, we leverage the grounding and spatial reasoning capabilities of the state-of-the-art VLMs. Given a high-level task instruction and current visual observation, the VLM first performs semantic decomposition, parsing the instruction into a sequence of executable action primitives. For the first primitive, the model then identifies two key spatial anchors: the pixel-space coordinate of the current gripper position (start) and the semantically inferred goal location (end). These coordinates are rendered as 2D heatmaps, Gaussian-smoothed, and used to produce a spatial guidance signal. This signal is then provided as a conditioning input to the video diffusion model, effectively anchoring its generation to a grounded, primitive-specific trajectory plan in pixel space.

Critically, our approach ensures that the video diffusion model observes the entire robot arm, including the base, joints, and full kinematic chain–unlike prior work that focuses narrowly on the end-effector or operates in cropped views. This global visibility enables the model to internalize soft physical constraints, such as joint limits, reachability, and base stability, by learning from visual data alone. As a result, the generated visual rollouts are not only semantically aligned and visually plausible, but also physically valid and executable in real environments.

Finally, instead of predicting control signals directly, we extract 6-DoF end-effector trajectories from the video frames using geometry-based visual projection. These trajectories are executed via Cartesian control, and the resulting observations are fed back to the planner, forming a semi-closed-loop control pipeline that enhances robustness, avoids error accumulation, and supports plug-and-play adaptation across embodiments.

In summary, our work advances embodied learning along three core axes–**Effectiveness**, **Explainability**, and **Efficiency** ($\mathbf{E^3}$). Our framework effectively combines the semantic parsing strengths of VLMs with the temporal modeling capacity of video diffusion, enabling policy distillation from diverse visual sources without task-specific retraining. It remains fully explainable through primitive-level decomposition, spatial grounding, and trajectory-based control, offering transparency and safety in real-world deployment. Furthermore, by limiting generation to short-horizon primitives, our system achieves significant efficiency gains over monolithic video models, supporting real-time inference with minimal overhead.

In particular, our method facilitates improved grounding between natural language and robotic execution, more efficient data utilization, and soft embodiment-aware constraints through full-arm observation–allowing the model to internalize physically feasible motion dynamics and avoid generating unrealistic or unreachable robot configurations.

2 Related Work

2.1 Video Generation as World Models

Given the overwhelming success of Diffusion Models [78, 36, 80, 79] in image generation [73, 72, 25], video diffusion models [50, 37, 35, 52, 8, 77, 87, 16, 17, 90] have emerged as a prominent research focus. With the emergence of Sora [12, 62], Transformer-based video generation has gained traction as a mainstream paradigm. Leveraging the Diffusion Transformer (DiT) architecture [68], these models significantly improve video stability and temporal consistency [101, 55, 97, 46, 71, 98, 86]. Empowered by strong visual priors, video diffusion models show promising potential as pixel-level world models. A world model learns to predict future system states, either in a low-dimensional latent space [1, 29, 47] or directly in pixel space [20, 32, 33, 66, 15]. A large body of work [45, 27, 95, 12, 6, 89, 13, 38, 102, 23, 54, 21, 38, 51, 81, 69, 100, 103, 34] has framed video generation as a new methodology for learning world models. Latest works, typically DiT-based and following a unified generation paradigm [51, 81, 100], are costly to train and lack real-time responsiveness. While hierarchical approaches [2, 23, 102] offer structure, they struggle with real-time adaptation in dynamic environments and rigid cross-modal alignment. Error accumulation in layered designs remains an open problem. Several approaches [2, 23, 102] adopt hierarchical architectures, among which [2, 23] are the closest to ours. However, these methods lack the integration of vision-language reasoning, planning, and grounding as in our approach. In contrast, our approach is distinguished by its ability to perform zero-shot policy execution, without relying on task-specific trained policies that often suffer from limited generalizability. Furthermore, our start-to-end point guidance surpasses their interactive refinement or reliance on 3D models in both efficiency and explainability. Some additional studies [75, 88, 94, 14] investigated non-diffusion approaches to modeling embodied world models.

2.2 End-to-End Vision-Language-Action Learning

End-to-end Vision-Language-Action (VLA) models constitute another major branch of embodied learning, directly mapping visual and linguistic inputs to actions without explicit dynamics modeling. Recent advances in vision-language models (VLMs) [19, 48, 3, 49] have catalyzed the development of unified VLA systems that integrate perception, grounding, reasoning, and control within a single architecture–departing from traditional modular pipelines. At the core of most VLA systems lies the intuition that pre-trained vision-language models (VLMs) can serve as high-level semantic engines. Methods like RT-2 [11], OpenVLA [44], and RoboAgent [4] inject Internet-scale knowledge into robot policies by fine-tuning VLMs or conditioning them with low-level control tokens. Based on their action modeling paradigms, VLA approaches can be broadly categorized into regression-based [53, 10, 58], generative/diffusion-based [7, 60, 83], and hybrid methods [59, 99]. While these unified policies show strong performance on short-horizon or single-task scenarios, they often struggle with generalization and scalability–particularly across long horizons, domains, or embodiments. To address these limitations, recent efforts have introduced hierarchical structures into end-to-end VLA models, either by modularizing perception and policy [57, 82, 5], or by structuring execution into multi-stage skills [76, 70]. Though often labeled hierarchical, many still preserve end-to-end differentiability or latent consistency across stages. Such designs improve sample efficiency, handle long-horizon dependencies, and provide interpretable intermediate reasoning. Despite these

advances, existing VLA systems remain heavily dependent on in-domain demonstrations. Some mitigate this via cross-modal or cross-embodiment data [39, 67], or by scaling human video demonstrations [93, 84], yet diversity and coverage remain limited. Our work takes a complementary direction: we retain an end-to-end VLA formulation while introducing primitive-level hierarchy and spatially grounded guidance. This design enables zero-shot generalization across unseen tasks and settings, while maintaining modularity, interpretability, and execution efficiency.

2.3 Datasets for Robotic Learning

High-quality and diverse data is essential for generalizable robotic learning [92]. Existing large-scale datasets are typically collected via human teleoperation [9, 24, 42], scripted pipelines [22, 31], or expert policies [74], and have largely been consolidated in OpenX-Embodiment [22]. Despite its scale, this benchmark remains costly to scale and limited in semantic and morphological diversity.

To enhance compositionality, RH20T-p [18] introduced primitive-level annotations atop RT-20T [28], but suffers from label noise and redundancy. Meanwhile, simulated benchmarks such as RLBench [40], CALVIN [65], and LIBERO [56] offer controllable environments, though often lack photorealism or task variety.

Recent works [64, 30, 96, 91] explore augmenting human demos via simulation or generative rendering, but still rely on expert trajectories. In contrast, our framework directly generates large-scale, primitive-level demonstrations in simulation without human supervision or segmentation. Combined with pre-trained vision-language models, this enables efficient sim-to-real transfer with strong generalization.

3 Method

Our approach is a modular and hierarchical system designed to bridge semantic-level planning and physically plausible action execution. It consists of three key components: a high-level semantic planner powered by a vision-language model (VLM), a primitive-conditioned video diffusion world model, and a zero-shot policy execution mechanism via trajectory projection and semi-closed-loop control. For a quick overview of our approach, please refer to Figure 1.

3.1 Hierarchical Embodied World Modeling

3.1.1 Semantic Primitive Planning via Vision-Language Models

Given a task instruction x^{text} and current observation frame x_0^{img}, we first employ a large vision-language model $\mathcal{P}$ to perform semantic parsing and primitive planning:

$$\mathcal{A}_{1:N} = \mathcal{P}(x^{\text{text}}, x_0^{\text{img}}), \tag{1}$$

where $\mathcal{A}_{1:N}$ denotes the sequence of N atomic primitives (e.g., "pick up object A", "place on B"). Each primitive is represented by an instruction a_k that is grounded locally in space and time.

For each primitive a_k, the planner identifies two pixel-space locations:

$$(s_k, g_k) = \mathcal{G}(a_k, x_0^{\text{img}}), \tag{2}$$

where $s_k \in \mathbb{R}^2$ and $g_k \in \mathbb{R}^2$ are the start and goal positions of the gripper in the image plane. These are Gaussian-blurred to produce heatmaps H_s and H_g, forming spatial guidance signals.

We leverage LoRA-based adaptation to enable reasoning and spatial grounding, following a Chain-of-LoRA [61]-style architecture. The structure is detailed in Figure 2.

3.1.2 Primitive-Conditioned Video Diffusion Generation

The video diffusion model $\mathcal{D}$ is conditioned on the current frame x_0^{img}, the text description of the primitive a_k, and the spatial heatmaps H_s, H_g. The model learns to predict a rollout of T future frames:

$$x_{1:T}^{\text{img}} \sim \mathcal{D}(x_0^{\text{img}}, a_k, H_s, H_g). \tag{3}$$

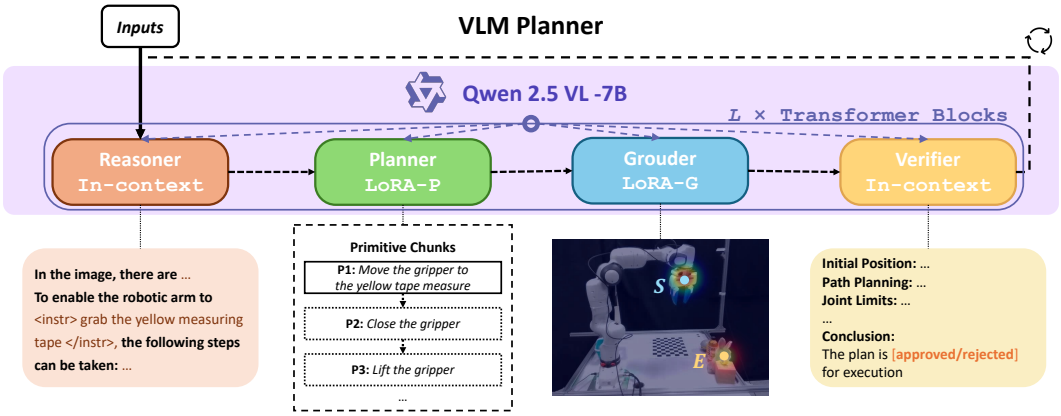

Figure 2: This workflow diagram delineates the architecture of our VLM planner, which is based on the Qwen 2.5 VL-7B model. The Reasoner generates descriptions and reasoning for the input to facilitate the functioning of subsequent components. The Planner (LoRA-P) and Grounder (LoRA-G) comprise two learnable LoRA modules: the Planner produces a series of primitive chunks, delivering modular instructions, while the Grounder provides specific descriptions of the end effector's start and goal positions. The Verifier assesses the executability of the plan, indicating approved or rejected statuses to ensure precision in planning.

Unlike prior work that only observes cropped end-effector regions, $\mathcal{D}$ is trained to observe the entire robot arm, allowing it to learn soft physical constraints such as reachability, base stability, and joint configuration limits.

The model is trained using a combination of pixel-level ℓ_2 reconstruction loss and perceptual similarity loss $\mathcal{L}_{\text{LPIPS}}$:

$$\mathcal{L}_{\text{vid}} = \sum_{t=1}^{T} \left\| x_t^{\text{img}} - \hat{x}_t^{\text{img}} \right\|_2^2 + \lambda \cdot \mathcal{L}_{\text{LPIPS}}(x_t^{\text{img}}, \hat{x}_t^{\text{img}}), \tag{4}$$

where $\hat{x}_t^{\text{img}}$ denotes the ground truth frame at time t.

Appendix Figure 5 provides a detailed overview of the video generation module and the process of extracting 6-DoF end-effector trajectories. It should be explicitly noted that for primitives involving binary gripper actions (e.g., open or close), we bypass the video generation module and execute the action directly via symbolic control. This improves execution efficiency and enables a clean architectural decoupling between discrete grasping commands and continuous 6-DoF motion planning, thereby enhancing modularity and interpretability.

3.2 Task-Agnostic Zero-Shot Policy and Execution

The generated future rollout $x_{1:T}^{\text{img}}$ for a primitive a_k lacks spatial information. To bridge the gap between pixel-space visual rollouts and real-world execution, we harness an off-the-shelf pose 6D estimation mechanism that requires only the generated RGB video as input.

Specifically, given a generated video sequence $V = \{I_1, I_2, \ldots, I_T\}$ and a reference video $V_r = \{I_{r1}, I_{r2}, \ldots, I_{rT'}\}$ of the robotic gripper, we estimate the gripper's 6-DoF pose using a model-free, RGB-based pose estimator: $\mathbf{R}, \mathbf{T} = PE(V, V_r)$, where $\mathbf{R}$ and $\mathbf{T}$ denote the rotation and translation matrices, respectively. In our implementation, we use Gen6D [63] for its simplicity and generalization capability.

The pose information of the robotic gripper, particularly its 6-DoF object pose, utilizing the generated video sequences directly for real-world robotic operations presents significant challenges.

After generating the future rollout $x_{1:T}^{\text{img}}$ for a primitive a_k, we extract a 6-DoF end-effector trajectory $\tau_k = \{p_1, \ldots, p_T\}$ using 6-DoF pose estimation [63] and geometric transformation. These trajectories are then mapped into Cartesian control space and executed by the robot.

To project the estimated poses into real-world coordinates, we correct for the scale ambiguity inherent in monocular depth predictions. Given the depth map D at the initial frame and known camera intrinsics (x_c, y_c, f_x, f_y), we compute the 3D location of any pixel (x, y) using:

$$X = (x - x_c) \cdot d / f_x, \quad Y = (y - y_c) \cdot d / f_y, \quad Z = d, \tag{5}$$

where d is the depth at pixel (x, y). Since the generated video lacks absolute depth scale, we align the scale of the predicted trajectory using the ratio between the real-world depth d_0 (from the first frame of the depth camera) and the pixel-based depth d_0^{pixel} (from COLMAP or Gen6D). This yields a fixed correction factor $\lambda = \frac{d_0}{d_0^{\text{pixel}}}$, which is applied to all subsequent predicted translations:$\mathbf{T}_{\text{real}} = \lambda \cdot \mathbf{T}_{\text{pixel}}$.

The observed result of executing τ_k is then captured and passed back to the planner $\mathcal{P}$, enabling feedback-driven refinement in subsequent primitives. This forms a semi-closed-loop execution pipeline that mitigates error accumulation and enables task-aware adaptive control without end-to-end backpropagation.

Together, this design enables zero-shot generalization to unseen tasks, objects, and morphologies, using modular, interpretable, and resource-efficient world model components.

The process of mapping the generated video through the camera's intrinsic and extrinsic parameters to obtain real-world coordinates is illustrated in Figure 3.

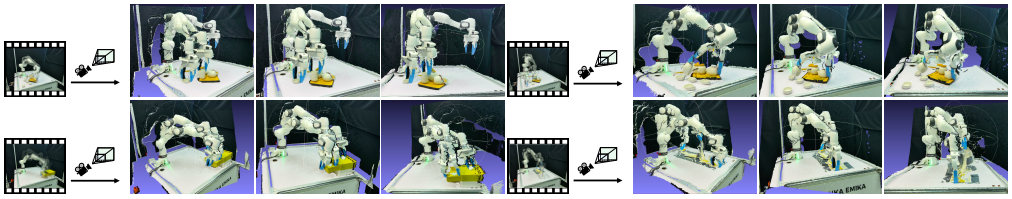

Figure 3: Mapping the generated video frames through the camera's intrinsic and extrinsic parameters to visualize the corresponding 3D object poses in real-world coordinates.

3.3 Full-Arm, Multi-View, and Primitive-Aligned Data Collection Pipeline

Model generalization in embodied learning is fundamentally governed by the quality and structure of the available data [92]. Collecting large-scale and diverse data for embodied learning, however, remains notoriously labor-intensive. Prior real-world datasets often require years of human teleoperation or scripted control to accumulate on the order of 100K-1M trajectories [9, 22, 43, 85, 23]. Despite their size, such datasets typically exhibit narrow generalization, often struggling to extrapolate beyond minor variations in object pose or scene composition [67, 64]. Moreover, the need to manually segment tasks, align modalities, and annotate action labels imposes substantial overhead, limiting diversity and scalability.

To improve generalization in embodied agents, we advocate organizing and collecting data at the level of semantically grounded primitives. This structured design enables compositional generalization by allowing the model to cover a wide task space using a compact set of atomic operations. It also facilitates more focused representation learning, as the reduced temporal scope of each primitive helps the model capture salient object properties and interaction dynamics. Moreover, the semantic granularity simplifies annotation–coarse labels can be generated using vision-language models with minimal supervision. Finally, training the video diffusion model on short-horizon primitives significantly reduces computational overhead and eases optimization compared to long-horizon video prediction.

This rationale is further illustrated in appendix Figure 8, which presents a heuristic visualization of the impact of primitive structuring on generalization.

Specifically, we construct a primitive-centric dataset $\mathcal{D}_{\text{prim}}$ tailored for training our video diffusion world model $\mathcal{W}$. Each sample in $\mathcal{D}_{\text{prim}}$ consists of a tuple $(x_0^{\text{img}}, a_k, H_s, H_g, x_{1:T}^{\text{img}})$, where x_0^{img} is the initial observation, a_k is the primitive instruction, H_s and H_g are Gaussian heatmaps denoting gripper start and goal positions, and $x_{1:T}^{\text{img}}$ are the predicted future frames. Note that the primitive data can be either segmented during collecting or automatically segmented using motion heuristics and language-guided tagging tools, minimizing human supervision.

Another notable design choice in $\mathcal{D}_{\text{prim}}$ is that the full robot arm–including the base and joint structures–is always visible in the field of view, in contrast to cropped, end-effector-only frames common in prior datasets. This enables the diffusion model to internalize soft physical constraints such as reachability, joint limits, and spatial feasibility during prediction.

4 Experiments & Analysis

4.1 Dataset Description

Our dataset consists of 7,326 simulated and 11,465 real-world primitives collected using a 5-camera synchronized setup with Deoxys [104]. By segmenting long-horizon tasks via keyframe indices, we extracted on average 5.8 primitives per session, achieving up to $29\times$ collection efficiency. As shown in Figure 4, cameras were arranged to maximize coverage of the workspace.

For annotation, 10% of the data was manually labeled, then used to fine-tune Qwen-VL-7B for auto-labeling the remainder, followed by light manual correction. To enhance visual generation quality, we mixed a small portion of diverse simulated data into the training set, including examples generated from RLBench, SayCan (PyBullet), and OpenVLA in LIBERO [56]. This hybrid strategy improves model generalization and dynamic realism, as confirmed in Section 4.4.

4.2 Baselines

We compare our framework against OpenVLA [44], a state-of-the-art end-to-end vision-language-action model that maps raw instructions and observations directly to actions. We evaluate OpenVLA under two settings: **(1) Zero-shot**: The pre-trained OpenVLA model is deployed directly on our benchmark tasks without any task-specific finetuning. This setting evaluates its raw generalization capability across unseen objects and scenes. **(2) Finetune**: The model is finetuned on 100 task-specific demonstrations per task using supervised behavior cloning, following the same protocol as in our system's policy training. This setting represents an upper-bound of OpenVLAs performance under favorable conditions. We note that OpenVLA performs poorly in the zero-shot setting and fails completely on real-robot deployment, whereas our method maintains strong performance even without task-specific finetuning or retraining of the video model.

4.3 Performance

4.3.1 Overall Success Rate

Table 1: **Overall success rate on RLBench tasks.** We compare our method against existing baselines on 9 manipulation tasks from the RLBench benchmark [40], with success rates averaged over 100 episodes. Results for Image-BC [41], UniPi* [27], and 4DWM [100] are directly cited from the 4DWM paper [100]. Our method (bottom row) achieves the highest success rate on all tasks, with consistent gains in articulated and contact-rich scenarios.

Methods	**close** box	**open** drawer	**open** jar	**open** microwave	**put** knife	**sweep to** dustpan	**lid** off	**weighing** off	**water** plants
Image-BC	53	4	0	5	0	0	12	21	0
UniPi*	81	67	38	<u>72</u>	66	49	<u>70</u>	**68**	35
4DWM	<u>88</u>	<u>80</u>	**44**	70	<u>70</u>	<u>56</u>	**73**	<u>62</u>	<u>41</u>
Ours	**93**	**84**	<u>43</u>	**78**	**72**	**63**	67	58	**56**

Table 1 presents a comprehensive comparison of overall success rates across nine manipulation tasks from the RLBench benchmark . The performance of our method is assessed against several existing baselines, including Image-BC, UniPi*, and 4DWM, with success rates averaged over 100 episodes.

Our approach consistently outperforms all baseline methods across the various tasks. Notably, we achieve a success rate of 93% in the task of closing a box, which is the highest recorded. Similarly, our method attains a success rate of 84% in opening a drawer, further demonstrating its effectiveness in articulated scenarios. While the success rate for opening a jar is comparable to the baselines at 43%, we maintain competitive performance in other tasks, such as opening a microwave (78%) and weighing plants (56%).

The results indicate that our method excels particularly in articulated and contact-rich tasks, where the nuanced handling of objects is critical. These findings underscore the robustness and generalization capabilities of our approach, establishing it as a leading solution in the domain of robotic manipulation.

4.3.2 Planning

Table 2: **Performance breakdown on three real-world tasks across planning, video generation, and primitive execution.** "Ours" denotes our method with frozen video diffusion; "OpenVLA" represents a strong end-to-end baseline. We additionally report OpenVLA's zero-shot performance without task-specific fine-tuning.

Task	Stage	Metric	Ours	OpenVLA	OpenVLA (ZS)
Pick up cup	Planning	Primitive accuracy	18 / 20	N/A	N/A
	Video Generation	Frame realism (/total)	17 / 20	N/A	N/A
	Primitive Execution	Task success	**16** / 20	12 / 20	0 / 20
Move cloth	Planning	Primitive accuracy	16 / 20	N/A	N/A
	Video Generation	Frame realism (/total)	15 / 20	N/A	N/A
	Primitive Execution	Task success	**14** / 20	10 / 20	0 / 20
Fold cloth	Planning	Primitive accuracy	15 / 20	N/A	N/A
	Video Generation	Frame realism (/total)	14 / 20	N/A	N/A
	Primitive Execution	Task success	**13** / 20	4 / 20	0 / 20

Our method performs a notable success rate, attributable to the effective decomposition of tasks into manageable primitives. Each primitive is trained on a mixed dataset, which enhances its adaptability and accuracy in real-world scenarios. Importantly, the test scenarios outlined in Table 2 were not present in the training set, yet our method achieves a high primitive accuracy of 18 out of 20 for the task of picking up a cup. Similar performance is observed across other tasks, with accuracy rates of 16 out of 20 for moving cloth and 15 out of 20 for folding cloth. This robust planning capability ensures that the robot can generate reliable sequences of actions, setting a strong foundation for successful execution.

Notably, OpenVLA fails completely in the zero-shot setting, highlighting its reliance on task-specific adaptation. In contrast, our method achieves strong generalization without retraining the video model, benefiting from modularity and spatial grounding in its design.

4.3.3 Primitive Execution

The results further highlight the effectiveness of our approach. The integration of planning and video generation significantly contributes to the overall task success rates. For instance, in the task of picking up a cup, our method achieves a task success rate of 16 out of 20, compared to only 12 out of 20 for the OpenVLA baseline. Similar trends are observed in the other tasks, with our method outperforming OpenVLA in both moving and folding cloth. Notably, the ability to succeed in these tasks–despite their absence in the training set–demonstrates the robustness and generalization capability of our method. This enhanced execution performance can be attributed to the careful segmentation of primitives and the use of high-quality video generation, which collectively facilitate precise and effective task completion.

4.3.4 Long-Horizon Tasks Planning Performance

The proposed methodology effectively addresses long-horizon tasks by decomposing them into a series of modular primitives, each designed to be plug-and-play. This hierarchical framework enables each primitive to concentrate on short-range objectives, which aligns closely with the strengths of Vision-Language Models (VLM). By constraining each component to operate within a limited contextual scope, we mitigate potential challenges such as contextual inconsistencies that may arise during long-term planning. As illustrated in Figure 3, the process of mapping generated video frames to real-world coordinates through the camera's intrinsic and extrinsic parameters underpins this modular approach. This mapping facilitates precise pose estimations for each primitive, thereby ensuring that the robotic system can execute tasks effectively without losing sight of the overarching long-

horizon goal. Consequently, our methodology leverages the inherent advantages of decomposing complex tasks into manageable segments, enhancing both efficiency and reliability in execution.

4.4 Ablation Study

Table 3: **Ablation study on model components.** Task success rate is reported over 20 trials per setting. "Primitive planner ablations" isolate the effect of VLM-based spatial grounding; "VideoGen ablations" evaluate the impact of simulation-augmented training.

Ablation Group	Variant	Pick up cup	Move cloth	Fold cloth
Primitive Planner	Full model (with start/end prompts)	**16 / 20**	**14 / 20**	**13 / 20**
	w/o start-end prompt	12 / 20	10 / 20	7 / 20
	w/o primitive planner (direct instruction-to-action)	9 / 20	5 / 20	3 / 20
Video Generation	Full model (with sim + real data)	**16 / 20**	**14 / 20**	**13 / 20**
	Trained on real-only data	12 / 20	9 / 20	5 / 20

We conduct two ablation studies to assess the contribution of (1) the VLM-based primitive planner, and (2) simulation-augmented training for video generation. As shown in Table 3, removing the start-end spatial prompts from the planner leads to a noticeable decline in performance across all tasks. This confirms that explicitly grounding primitives with spatial anchors significantly improves downstream execution. A further removal of the entire primitive decomposition step–i.e., mapping instructions directly to actions–results in severe performance degradation, especially on multi-step or spatially entangled tasks like cloth folding. For video generation, removing simulated data from the training corpus substantially weakens the model's rollout quality, leading to lower execution success rates. This suggests that simulation data offers critical diversity and coverage, allowing the model to generalize better to real-world scenes with varied configurations and object appearances.

4.5 Efficiency

As shown in appendix Table 4, our model demonstrates significant advantages in both computational efficiency and memory footprint. Compared to large-scale diffusion baselines such as Hunyuan I2V and Wan 2.1, our method achieves up to **75x faster inference** and **67x lower VRAM usage**, making it well suited for real-time robotic deployment. Despite its lightweight design, our method maintains competitive video quality and task relevance, validating its practicality in resource-constrained scenarios.

5 Conclusion

We present a modular and interpretable framework for zero-shot embodied policy learning, combining a vision-language planner with a primitive-conditioned video diffusion world model. By grounding high-level instructions into spatially guided primitives and reusing a lightweight video model across tasks, our system achieves strong performance on challenging real-world manipulation benchmarks. Extensive experiments demonstrate superior generalization, data efficiency, and runtime scalability. Our work highlights the potential of combining semantic reasoning with visual dynamics modeling for scalable, reusable, and task-agnostic robotic intelligence.

6 Limitations

While our framework shows advancements in planning and decision-making for embodied agents, several limitations warrant consideration:

Challenges in High-Resolution Generation: The framework struggles with high-resolution video generation, impacting performance in tasks needing detailed visual fidelity.

Dependency on Language Instructions: It relies on vision-language models for deriving action primitives, limiting adaptability to complex instructions not well-represented in the training data.

These limitations highlight areas for future research to enhance the framework's applicability in real-world scenarios.

References

[1] Alessandro Achille and Stefano Soatto. A separation principle for control in the age of deep learning, 2017. URL https://arxiv.org/abs/1711.03321.

[2] Anurag Ajay, Seungwook Han, Yilun Du, Shuang Li, Abhi Gupta, Tommi Jaakkola, Josh Tenenbaum, Leslie Kaelbling, Akash Srivastava, and Pulkit Agrawal. Compositional foundation models for hierarchical planning. *Advances in Neural Information Processing Systems*, 36:22304–22325, 2023.

[3] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[4] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking, 2023.

[5] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[6] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164.

[7] 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.

[8] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

[9] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[10] 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.

[11] 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.

[12] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

[13] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

[14] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

[15] Chang Chen, Yi-Fu Wu, Jaesik Yoon, and Sungjin Ahn. Transdreamer: Reinforcement learning with transformer world models. *arXiv preprint arXiv:2202.09481*, 2022.

[16] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023.

[17] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024.

[18] Zeren Chen, Zhelun Shi, Xiaoya Lu, Lehan He, Sucheng Qian, Hao Shu Fang, Zhenfei Yin, Wanli Ouyang, Jing Shao, Yu Qiao, Cewu Lu, and Lu Sheng. Rh20t-p: A primitive-level robotic dataset towards composable generalization agents. *arXiv preprint arXiv: 2403.19622*, 2024.

[19] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[20] Silvia Chiappa, Sébastien Racaniere, Daan Wierstra, and Shakir Mohamed. Recurrent environment simulators. *arXiv preprint arXiv:1704.02254*, 2017.

[21] Jaden Clark, Suvir Mirchandani, Dorsa Sadigh, and Suneel Belkhale. Action-free reasoning for policy generalization. In *https://arxiv.org/abs/2403.01823*, 2025.

[22] Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi Jim Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning

Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick Tree Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. `https://arxiv.org/abs/2310.08864`, 2023.

[23] Murtaza Dalal, Min Liu, Walter Talbott, Chen Chen, Deepak Pathak, Jian Zhang, and Ruslan Salakhutdinov. Local policies enable zero-shot long-horizon manipulation. *International Conference of Robotics and Automation*, 2025.

[24] Shivin Dass, Jullian Yapeter, Jesse Zhang, Jiahui Zhang, Karl Pertsch, Stefanos Nikolaidis, and Joseph J. Lim. Clvr jaco play dataset, 2023. URL `https://github.com/clvrai/clvr_jaco_play_dataset`.

[25] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers, 2022. URL `https://arxiv.org/abs/2204.14217`.

[26] Yilun Du and Leslie Kaelbling. Compositional generative modeling: A single model is not all you need. *arXiv preprint arXiv:2402.01103*, 2024.

[27] 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*, 36:9156–9172, 2023.

[28] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A robotic dataset for learning diverse skills in one-shot. In *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.

[29] Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. 2004.

[30] Caelan Garrett, Ajay Mandlekar, Bowen Wen, and Dieter Fox. Skillmimicgen: Automated demonstration generation for efficient skill learning and deployment, 2024. URL `https://arxiv.org/abs/2410.18907`.

[31] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. Maniskill2: A unified benchmark for generalizable manipulation skills. In *International Conference on Learning Representations*, 2023.

[32] David Ha and J
rgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

[33] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

[34] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, pages 1–7, 2025.

[35] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation, 2023. URL https://arxiv.org/abs/2211.13221.

[36] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

[37] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 8633–8646. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/39235c56aef13fb05a6adc95eb9d8d66-Paper-Conference.pdf.

[38] 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.

[39] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025. URL https://arxiv.org/abs/2504.16054.

[40] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 2020.

[41] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022.

[42] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset. 2024.

[43] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

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

[45] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B Tenenbaum. Learning to Act from Actionless Videos through Dense Correspondences. *arXiv:2310.08576*, 2023.

[46] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[47] Timothée Lesort, Natalia Díaz-Rodríguez, Jean-Franois Goudou, and David Filliat. State representation learning for control: An overview. *Neural Networks*, 108:379–392, 2018.

[48] 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.

[49] Dingzhe Li, Yixiang Jin, Yuhao Sun, Hongze Yu, Jun Shi, Xiaoshuai Hao, Peng Hao, Huaping Liu, Fuchun Sun, Jianwei Zhang, et al. What foundation models can bring for robot learning in manipulation: A survey. *arXiv preprint arXiv:2404.18201*, 2024.

[50] Muyang Li, Ji Lin, Chenlin Meng, Stefano Ermon, Song Han, and Jun-Yan Zhu. Efficient spatially sparse inference for conditional gans and diffusion models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 28858–28873. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/b9603de9e49d0838e53b6c9cf9d06556-Paper-Conference.pdf`.

[51] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.

[52] Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation, 2023. URL `https://arxiv.org/abs/2309.00398`.

[53] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*, 2023.

[54] Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation, 2024. URL `https://arxiv.org/abs/2406.16862`.

[55] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

[56] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023.

[57] Jiaming Liu, Chenxuan Li, Guanqun Wang, Lily Lee, Kaichen Zhou, Sixiang Chen, Chuyan Xiong, Jiaxin Ge, Renrui Zhang, and Shanghang Zhang. Self-corrected multimodal large language model for end-to-end robot manipulation. *arXiv preprint arXiv:2405.17418*, 2024.

[58] Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Pengju An, Xiaoqi Li, Kaichen Zhou, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation. *Advances in Neural Information Processing Systems*, 37:40085–40110, 2024.

[59] Jiaming Liu, Hao Chen, Pengju An, Zhuoyang Liu, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, Mengzhen Liu, et al. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*, 2025.

[60] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

[61] Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for long video reasoning. *arXiv preprint arXiv:2503.13444*, 2025.

[62] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

[63] Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In *Proceddings of the European Conference on Computer Vision*, pages 298–315, 2022.

[64] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations, 2023. URL `https://arxiv.org/abs/2310.17596`.

[65] 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.

[66] Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. *arXiv preprint arXiv:2209.00588*, 2022.

[67] NVIDIA, :, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Jim Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. Gr00t n1: An open foundation model for generalist humanoid robots, 2025. URL `https://arxiv.org/abs/2503.14734`.

[68] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL `https://arxiv.org/abs/2212.09748`.

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

[70] Physical Intelligence. π0.5: A vision-language-action model with open-world generalization. `https://www.physicalintelligence.company/blog/pi05`, 2025. Accessed: 2025-04-25.

[71] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev,

Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models, 2025. URL https://arxiv.org/abs/2410.13720.

[72] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

[73] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

[74] Giulio Schiavi, Paula Wulkop, Giuseppe Rizzi, Lionel Ott, Roland Siegwart, and Jen Jen Chung. Learning agent-aware affordances for closed-loop interaction with articulated objects. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5916–5922. IEEE, 2023.

[75] Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel. Masked world models for visual control. In *Conference on Robot Learning*, pages 1332–1344. PMLR, 2023.

[76] Lucy Xiaoyang Shi, Brian Ichter, Michael Equi, Liyiming Ke, Karl Pertsch, Quan Vuong, James Tanner, Anna Walling, Haohuan Wang, Niccolo Fusai, et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. *arXiv preprint arXiv:2502.19417*, 2025.

[77] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022. URL https://arxiv.org/abs/2209.14792.

[78] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. URL https://arxiv.org/abs/1503.03585.

[79] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

[80] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

[81] Aether Team, Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, et al. Aether: Geometric-aware unified world modeling. *arXiv preprint arXiv:2503.18945*, 2025.

[82] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

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

[84] Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation, 2024. URL https://arxiv.org/abs/2412.15109.

[85] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.

[86] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[87] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report, 2023. URL `https://arxiv.org/abs/2308.06571`.

[88] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, 2024.

[89] Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024.

[90] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023.

[91] Zhengrong Xue, Shuying Deng, Zhenyang Chen, Yixuan Wang, Zhecheng Yuan, and Huazhe Xu. Demogen: Synthetic demonstration generation for data-efficient visuomotor policy learning, 2025. URL `https://arxiv.org/abs/2502.16932`.

[92] Zhengrong Xue, Shuying Deng, Zhenyang Chen, Yixuan Wang, Zhecheng Yuan, and Huazhe Xu. Demogen: Synthetic demonstration generation for data-efficient visuomotor policy learning. *arXiv preprint arXiv:2502.16932*, 2025.

[93] Jiange Yang, Bei Liu, Jianlong Fu, Bocheng Pan, Gangshan Wu, and Limin Wang. Spatiotemporal predictive pre-training for robotic motor control. *arXiv preprint arXiv:2403.05304*, 2024.

[94] Jiange Yang, Bei Liu, Jianlong Fu, Bocheng Pan, Gangshan Wu, and Limin Wang. Spatiotemporal predictive pre-training for robotic motor control, 2024. URL `https://arxiv.org/abs/2403.05304`.

[95] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

[96] Sizhe Yang, Wenye Yu, Jia Zeng, Jun Lv, Kerui Ren, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Novel demonstration generation with gaussian splatting enables robust one-shot manipulation, 2025. URL `https://arxiv.org/abs/2504.13175`.

[97] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[98] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2025. URL https://arxiv.org/abs/2408.06072.

[99] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024.

[100] Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: Learning 4d embodied world models. 2025. URL https://arxiv.org/abs/2504.20995.

[101] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. URL https://arxiv.org/abs/2412.20404.

[102] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination, 2024. URL https://arxiv.org/abs/2404.12377.

[103] Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[104] Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. *6th Annual Conference on Robot Learning (CoRL)*, 2022.

A Experimental Environment

Our experimental setup is illustrated in Figure 4. The environment consists of a workstation, a FR3 (Franka Research 3) robotic arm, and five cameras. The specific configuration is as follows:

Workstation: The workstation serves as a standard experimental platform, providing a stable area to support various operational tasks.

FR3 Robotic Arm: The FR3 robotic arm is a high-precision industrial robot equipped with flexible movement capabilities and high repeatability. This robotic arm is responsible for executing various tasks and can interact in real-time with the collected visual data.

Camera System: - **Two Realsense Cameras:** These cameras are mounted on the wrist of the robotic arm and on the shelf of the workstation, respectively. They capture real-time depth information and the arm's movements, enabling a comprehensive understanding of the spatial context and enhancing the accuracy and reliability of task execution. - **Three Femto Bolt Cameras:** Positioned around the workstation, these cameras are used to capture dynamic processes at high frame rates. They provide additional visual perspectives, ensuring that comprehensive image data is collected during complex tasks for subsequent analysis and processing.

In summary, the design of this experimental environment aims to provide comprehensive visual support for robotic operations, facilitating the research and development of complex tasks. Through the collaborative operation of a multi-camera system, we can obtain high-quality visual data, thereby enhancing the performance of the robot in practical applications.

B Heuristic Intuition for Primitive-Level Data Organization Enabling Compositional Generalization under Sparse Embodied Data

Figure 8 presents a heuristic Intuition for Primitive-Level Data Organization Enabling Compositional Generalization under Sparse Embodied Data.

C An Illustration for Primitive-Level Task Execution

Figure 5 illustrates a primitive-level task execution for Pick up the yellow tape measure. Motion primitives are executed via diffusion-based visual rollout and trajectory projection, while discrete gripper actions are handled directly through symbolic execution.

D Examples of Training Video Clips

In this study, we utilize both real-world and simulated video clips capturing the operation of robotic arms to train our model. The real-world data consists of videos collected from various robotic arm tasks using a Franka Emika arm, remotely operated through a Space Mouse interface. These tasks are performed in controlled environments, showcasing different motions and interactions with objects. These videos provide valuable information about how the arm operates in realistic conditions, with varying lighting, camera angles, and object placements.

The simulated video clips, on the other hand, are generated from physics-based environments such as **RLBench** and **LIBERO**. These simulations replicate the robotic arm's movements in virtual settings, allowing for the generation of large quantities of data under controlled conditions. This enables the model to learn from a diverse range of scenarios that might be difficult or time-consuming to capture in real life.

Examples of both real-world and simulated video clips are provided in Figure 6. These images illustrate the types of data used to train the model, offering a glimpse into the varied nature of the training set.

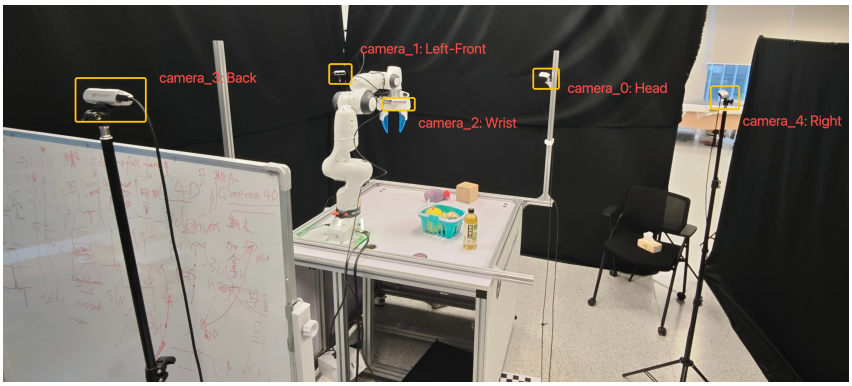

Figure 4: The workstation for data collection and real-robotic evaluation.

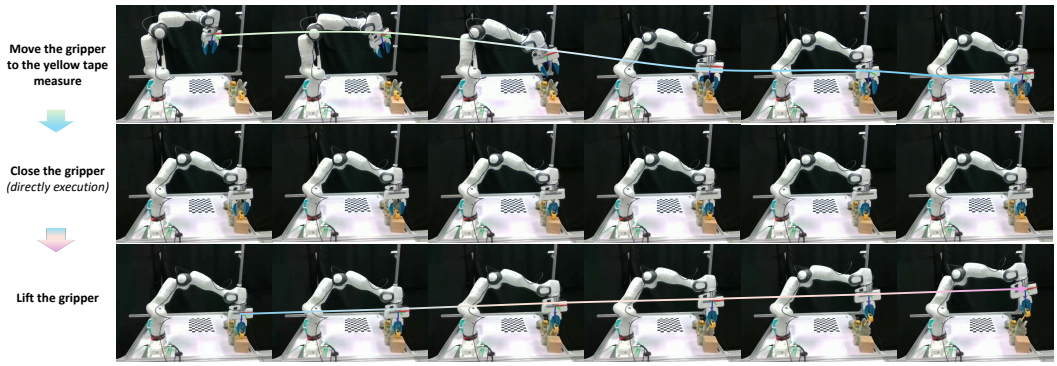

Figure 5: Illustration of primitive-level task execution for Pick up the yellow tape measure. Motion primitives are executed via diffusion-based visual rollout and trajectory projection, while discrete gripper actions are handled directly through symbolic execution.

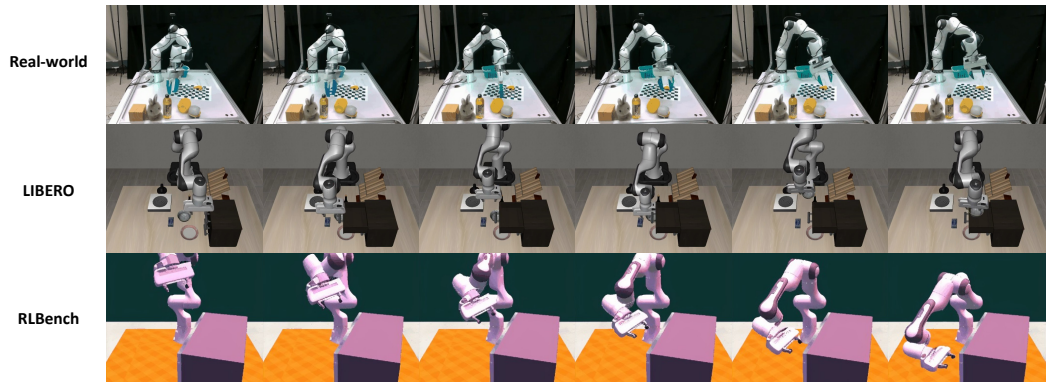

Figure 6: Sample frames from the training dataset, which consists of three components: real-world data collected using a Franka Emika robotic arm and Femto Bolt cameras, and two simulated datasets obtained from RLBench and LIBERO.

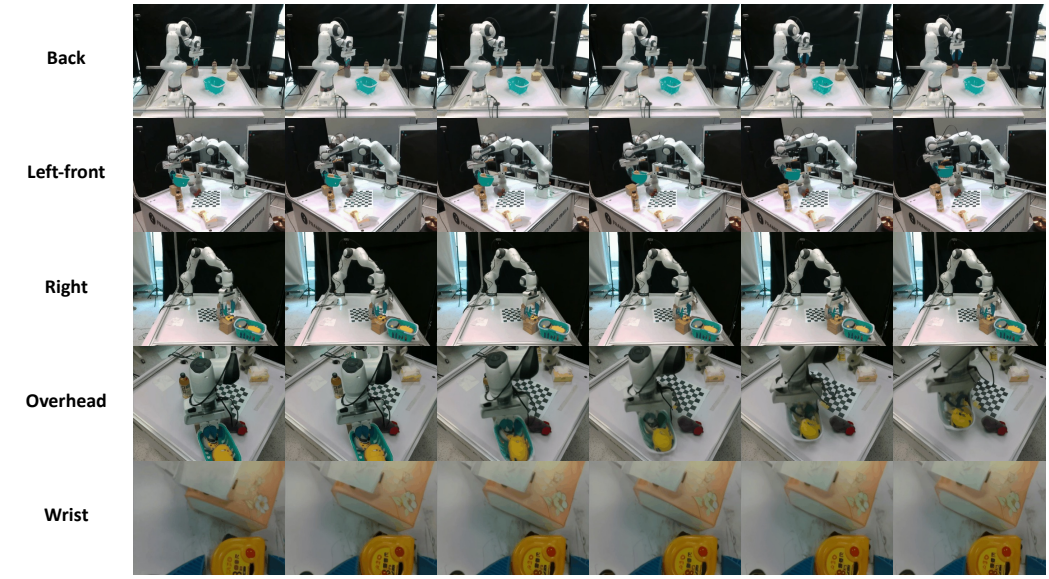

Figure 7: Sample frames from the generated video clips showing the Franka robotic arm performing manipulation tasks from five distinct viewpoints: Back, Left-front, Right, Overhead, and Wrist. Each viewpoint provides a unique perspective on the arm's movements, enhancing the model's ability to learn complex manipulation behaviors from diverse angles.

E Examples of Generated Video Clips

In addition to using recorded video data, we also generate synthetic video clips to enrich the training dataset and enhance the models generalization capability. These generated clips simulate the Franka robotic arm performing various manipulation tasks from multiple camera perspectives.

Specifically, each video sequence is rendered from five distinct viewpoints: **Back**, **Left-front**, **Right**, **Overhead**, and **Wrist**. These perspectives are selected to comprehensively capture the arm's kinematics, end-effector trajectories, and object interactions from both global and local contexts. The Back and Overhead views provide an overall understanding of the workspace and arm configuration, while Left-front and Right offer lateral angles that help reveal occluded motions. The Wrist view, positioned close to the end-effector, offers detailed observation of grasping and manipulation actions.

Representative frames from these generated clips are shown in Figure 7, illustrating how each viewpoint contributes to a multi-faceted understanding of the robotic operation.

F Efficiency Comparisons

Table 4 compares the video generation efficiency.

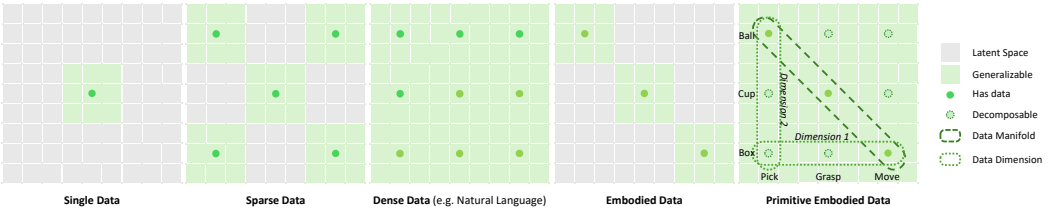

Figure 8: While densely distributed data can enable broad generalization, embodied data often suffer from sparsity [26, 92]. The rightmost schematic highlights how organizing embodied data at the primitive level–along orthogonal dimensions such as action and object–supports compositional generalization even under limited data availability.

Table 4: **Efficiency comparison of video generation models.** We compare video generation speed, memory footprint, and runtime on A100 GPUs. Ours achieves the best speed-VRAM trade-off under realistic deployment conditions.

Model	Resolution	VRAM (A100)	Time / Frames	FPS
Hunyuan I2V	480p	6079 GB	50 min / 81 frames (local)	0.027
4DWM (CogVideoX1.5-5B-I2)	480p	20 GB	18m20s / 49 frames	0.045
Wan 2.1 I2V (14B)	720p	76.7 GB	2715s / 81 frames	0.03
Ours	480p	**11 GB**	**16s / 32 frames**	**2.0**

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discuss the limitations of the work in section "Limitations".

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: A theoretical proof of the methods used is provided in the "Method" section.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper fully discloses all the necessary information to reproduce the main experimental results, including detailed descriptions of the methodologies, parameters, and conditions that affect the main claims and conclusions.

   Guidelines:

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

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

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

Answer: [No]

Justification: The paper does not currently provide open access to the data and code, but this will be addressed and included later.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: In the description of the experiment, this article specifies all the training and test details.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [No]

Justification: The experiments in this paper are not suitable for quantitative statistical analysis, so there are no error bars or similar components.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The paper provides sufficient information on the computer resources required to reproduce each experiment

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the paper conforms to the NeurIPS Code of Ethics in all respects, ensuring that ethical considerations were prioritized throughout the study.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The discussion is included in the sections on Conclusion and Limitations of the paper.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The creators or original owners of the assets used in the paper are properly credited, and the license and terms of use are explicitly mentioned and respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

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

13. **New assets**

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

    Answer: [No]

    Justification: The paper does not currently provide open access to the new assets, but this will be addressed and included later.

    Guidelines:

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

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

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

    Answer: [NA]

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

    Guidelines:

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

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

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

    Answer: [NA]

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

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: The paper describes the usage of LLMs as a crucial component of the core methods, highlighting their importance and originality in the research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.