

BagelVLA: Enhancing Long-Horizon Manipulation via Interleaved Vision-Language-Action Generation

Yucheng Hu^{*†§}, Jianke Zhang^{*†§}, Yuanfei Luo^{†‡}, Yanjiang Guo^{*}, Xiaoyu Chen^{*}, Xinshu Sun^{*}
Kun Feng^{*}, Qingzhou Lu^{*}, Sheng Chen[†], Yangang Zhang[†], Wei Li[†] and Jianyu Chen^{*}

[†]Equal Contribution, [§]Joint project lead

^{*}Tsinghua university

Email: {huyc24,zhangjk24}@mails.tsinghua.edu.cn

[†]ByteDance Seed

Email: liwei.85@bytedance.com

Abstract—Equipping embodied agents with the ability to reason about tasks, foresee physical outcomes, and generate precise actions is essential for general-purpose manipulation. While recent Vision-Language-Action (VLA) models have leveraged pre-trained foundation models, they typically focus on either linguistic planning or visual forecasting in isolation. These methods rarely integrate both capabilities simultaneously to guide action generation, leading to suboptimal performance in complex, long-horizon manipulation tasks. To bridge this gap, we propose BagelVLA, a unified model that integrates linguistic planning, visual forecasting, and action generation within a single framework. Initialized from a pretrained unified understanding and generative model, BagelVLA is trained to interleave textual reasoning and visual prediction directly into the action execution loop. To efficiently couple these modalities, we introduce Residual Flow Guidance (RFG), which initializes from current observation and leverages single-step denoising to extract predictive visual features, guiding action generation with minimal latency. Extensive experiments demonstrate that BagelVLA outperforms existing baselines by a significant margin on multiple simulated and real-world benchmarks, particularly in tasks requiring multi-stage reasoning. Demonstrations and code are available at <https://cladernyjorn.github.io/BagelVLA.github.io>

I. INTRODUCTION

The pursuit of generalist robots capable of performing complex manipulation tasks in unstructured environments remains a central goal in robotics. A robust embodied agent must possess three fundamental capabilities: understanding what to do based on instructions, predicting what will happen next, and executing the necessary motions. While recent Vision-Language-Action (VLA) models [4, 38, 27] have made progress by incorporating vision language models (VLMs) [22, 2, 17] or visual generation models [11, 16, 23, 35], they often treat these capabilities as separate modules. Some methods focus on high-level planning [17, 12] but lack visual forecasting, while others focus on visual prediction [11, 16, 5] but struggle with the logical reasoning required for complex tasks[26]. A unified framework that seamlessly integrates reasoning, prediction, and control remains a key challenge.

Meanwhile, the field of multimodal learning has witnessed the emergence of unified understanding and generation models [9, 37, 36, 44, 45]. Architectures like Bagel [9] employ a

single transformer backbone to jointly process and generate text and images, exhibiting emergent abilities in multimodal reasoning. These models provide an appealing prior for embodied agents: the model can “think” about the next step in text and “imagine” the outcome in pixels. However, such general-purpose models are not designed for embodied domain reasoning and continuous real-time control.

To make unified multimodal priors actionable for long-horizon manipulation, we propose BagelVLA, a unified VLA framework that integrates linguistic planning, visual forecasting, and action generation. Rather than treating these as isolated modules, BagelVLA interleaves them within a unified transformer architecture. The model first generates a textual plan to decompose the instruction (e.g., identifying the next object to manipulate), then predicts the future visual state, and finally generates the action. This design combines the logical reasoning of language models with the predictive power of visual generation, providing rich visual dynamics aligned with instruction to guide low-level control for long-horizon tasks.

Realizing this interleaved behavior requires a suitable training architecture and data, for which we design a two-stage training strategy to inject embodied multi-modal planning capabilities into the model. In the first stage, we construct a massive hybrid dataset combining general multimodal data[42, 30, 25] with large-scale robotic datasets[20, 1, 18, 43]. Robotic datasets from diverse embodiments are annotated to transfer the model’s general reasoning and visual predictive abilities to embodied settings. In the second stage, we introduce the action expert and fine-tune the full model to couple language, predicted visual dynamics, and control. This progressive approach ensures the model retains its high-level reasoning capabilities while acquiring precise low-level control policies. To address the high latency in combining visual generation, we introduce Residual Flow Guidance (RFG). This substantially reduces foresight cost by starting from the current observation and focusing the denoising step on action-induced residual changes rather than synthesizing the full scene [23, 10].

We validate BagelVLA through extensive experiments in both simulation and real-world environments. Results indicate

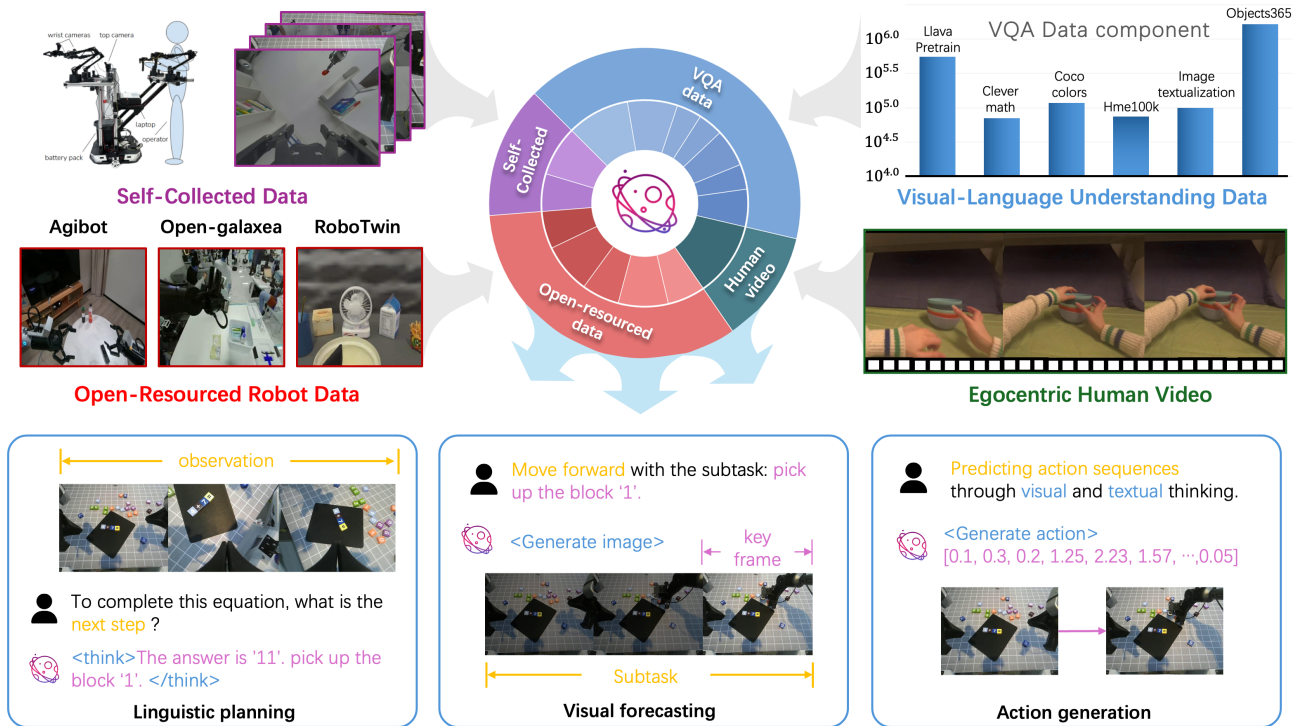


Fig. 1. **Overview of our framework.** We present BagelVLA, a unified model that integrates linguistic planning, visual forecasting, and action generation within a single framework. We construct a massive hybrid dataset combining general multimodal data with large-scale robotic datasets. Robotic datasets with sub-tasks and keyframes are annotated to transfer the foundation model’s general reasoning and visual generation abilities to embodied settings.

that BagelVLA is robust and generalizes to unseen instructions and diverse object arrangements where baseline methods often fail. Our contributions are as follows:

- We propose BagelVLA, which integrates linguistic planning, visual forecasting, and action generation into a single architecture. By explicitly modeling the transition from language to visual dynamics, our approach enhances reasoning and control in long-horizon tasks.
- By exploring various schemes for learning action representations from interleaved planning, we introduce Residual Flow Guidance (RFG), which uses the current observation as a structural prior and applies single-step denoising to capture future visual dynamics with minimal latency.
- BagelVLA substantially outperforms existing baselines in simulation benchmarks and demonstrates strong generalization to diverse instructions and environments in real-world experiments.

II. RELATED WORKS

1) *Vision-Language-Action Models:* Vision-Language-Action (VLA) models aim to enhance policy generalization to linguistic instructions and visual scenes by integrating vision-language models (VLMs) with action prediction. For example, methods like RT-2 [3] and OpenVLA [21] employ discrete action tokens compatible with VLMs, allowing direct mapping from vision-language representations to executable actions, though this can limit expressiveness in continuous control. In contrast, approaches such as

Octo [39], 3D Diffuser Actor [19], and π_0 [2] utilize continuous action representations via diffusion models to capture multimodal distributions, better handling fine-grained manipulations. However, these methods—whether discrete or continuous, overlook the alignment gap between VLM pre-training and VLA fine-tuning, resulting in degraded vision-language capabilities during adaptation. To mitigate this gap, other approaches [48, 16, 14, 35, 23] introduce visual prediction tasks as a bridge to map vision-language signals to action signals. For instance, VPP [16] proposes a video prediction policy that conditions robot actions on future visual representations derived from video diffusion models. Cosmos Policy [23] directly fine-tunes a large pretrained video model to serve as a robot policy. Although pre-training with pixel prediction can be easily aligned with the robot observations, the absence of a dedicated VLM backbone often leads to poor instruction-following performance, particularly in tasks requiring complex reasoning.

2) *Unified understanding and generation models:* In multi-modal learning, recent efforts [9, 37, 36, 44] have developed unified architectures for joint understanding and generation across modalities. For example, Bagel [9] uses a single transformer to process and generate text and images, trained on interleaved datasets for emergent reasoning. Chameleon [37] employs a token-based framework for mixed-modal input/output, supporting tasks like question answering and image generation. LMFusion [36] integrates language and vision in a fused transformer, focusing on efficient cross-modal

alignment, while Show-o [44] emphasizes unified multimodal understanding and generation, including text-conditioned image generation and editing for enhanced scene comprehension. These models, trained on diverse datasets including generation, QA, and editing, demonstrate strong capabilities in multimodal reasoning that can extend to embodied agents. Inspired by these, several VLA works [49, 33] have introduced action experts to transfer their capabilities to embodied scenarios. However, the lack of explicit embodied vision-language interleaved reasoning means these approaches only retain a subset of the original model’s capabilities, failing to implement step-by-step multimodal chain-of-thought reasoning. This deficiency is deemed critical for complex long-horizon tasks. In contrast, our proposed methods successfully incorporate the multi-modal reasoning capability into robotic manipulation via a complete data processing pipeline and a progressive training paradigm.

III. METHODOLOGY

A. Preliminaries: Interleaved planning for Robot Control

For classic language-conditioned manipulation settings, a policy is typically learned from a demonstration dataset $\mathcal{D} = \{L_i, \tau_i\}_{i=1}^N$, where each trajectory $\tau_i = \{(v_1, l_1, a_1), \dots, (v_T, l_T, a_T)\}$ consists of observations v_t (images and proprioception), stage-specific language descriptions l_t , and action chunks a_t . Conventional VLA models simplify this by conditioning purely on the global instruction L , learning a direct mapping policy $p_\theta(a_t|v_t, L)$. However, this formulation is insufficient for long-horizon tasks where a global instruction (e.g., stacking blocks in a specified order (red→yellow→blue→green)) implicitly entails a sequence of distinct stages. We address this by modeling the problem as **Interleaved Planning**. Instead of a black-box mapping, we require the model to explicitly reason through the causal chain of the task.

Formally, given the global instruction L and current observation v_t , BagelVLA models the joint distribution of the current subtask l_t , the future outcome (keyframe) v_{t+k} , and the action a_t . This joint distribution $p_\theta(a_t, v_{t+k}, l_t|v_t, L)$ is factorized based on the logical dependency of manipulation:

- 1) **Linguistic Planning**: The model first identifies the immediate textual objective l_t from the global instruction. We consider task decomposition to be the primary semantic capability of VLM-based architectures.
- 2) **Visual Forecasting**: Conditioned on this subtask, the model acts as a world model to predict the physical outcome v_{t+k} .
- 3) **Action Generation**: Finally, the action a_t is generated, grounded in both the textual plan and the visual forecast.

Consequently, our objective is formulated as the maximization of the following factorized likelihood:

$$\begin{aligned} \mathcal{J} &= -(\mathcal{L}_l + \mathcal{L}_v + \mathcal{L}_a) \\ &= \max_{\theta} \mathbb{E}_{\mathcal{D}} \log p_\theta(l_t|v_t, L) \cdot p_\theta(v_{t+k}|v_t, L, l_t) \\ &\quad \cdot p_\theta(a_t|v_t, L, l_t, v_{t+k}) \end{aligned}$$

B. Data Engine

To construct a large-scale pretraining dataset for subtask planning and keyframe prediction in embodied scenarios, we leverage diverse sources of manipulation demonstrations and apply tailored processing pipelines to four major data categories in Fig. 1 according to their characteristics.

1) **Robot Data**: The robot data comprises self-collected expert demonstrations and publicly available data from diverse embodiments. For proprietary data, we manually annotate and segment videos to obtain l_t , ensuring high-quality planning and keyframe prediction. For public datasets lacking fine-grained labels, we utilize Seed-1.5-VL-thinking [13] to synthesize l_t and identify temporal boundaries (start and end frames). These samples are then filtered to retain high-quality instances. These two components are used exclusively for pretraining to transfer the model’s fundamental planning and prediction capabilities to the embodied domain.

2) **General Data**: General Data includes egocentric human videos and large-scale image–text VQA data. For the former, we similarly employ Seed-1.5-VL-thinking to generate language annotations; however, due to the complexity of human-centered scenes, we do not annotate subtasks and instead predict only the final frame of each operation. These two data sources are mainly used to preserve the base model’s original understanding and generation capabilities.

Details of all data annotations and components are provided in the Supplementary Materials.

C. Model Architecture

To address the *interleaved planning* problem defined in Sec. III-A, we propose **BagelVLA**, a unified framework for understanding, prediction, and action generation. As illustrated in Fig. 2, BagelVLA is designed to process data across three modalities simultaneously. To leverage pre-existing large-scale multimodal data, we employ a Mixture of Transformers (MoT) architecture to orchestrate experts managing different modalities: specifically, an LLM expert, a generation expert, and an action expert, all connected via self-attention mechanisms.

We initialize the LLM and generation experts using Bagel [9], a unified MoT model for understanding and generation. On top of this foundation, we incorporate a smaller transformer to serve as the action expert. Distinct from prior MoT-based VLA architectures [33, 49], BagelVLA benefits from robust pre-training initialization for both its language and vision components and employs a novel dual flow-matching mechanism (detailed in Sec. III-D).

1) **Understanding Expert & Generation Expert**: The understanding and generation experts adopt the architecture of Qwen2.5-LLM-7B [46]. Following Bagel’s configuration, we utilize two distinct visual encoders responsible for visual-language understanding and goal image prediction, respectively. Each input observation view v_t is encoded by the SigLIP2 [40] and concatenated with the text instructions L (and l_t) to serve as input for the LLM Expert.

We also utilize the VAE from FLUX [24] to encode images. For **linguistic planning**, the understanding expert attends to ViT

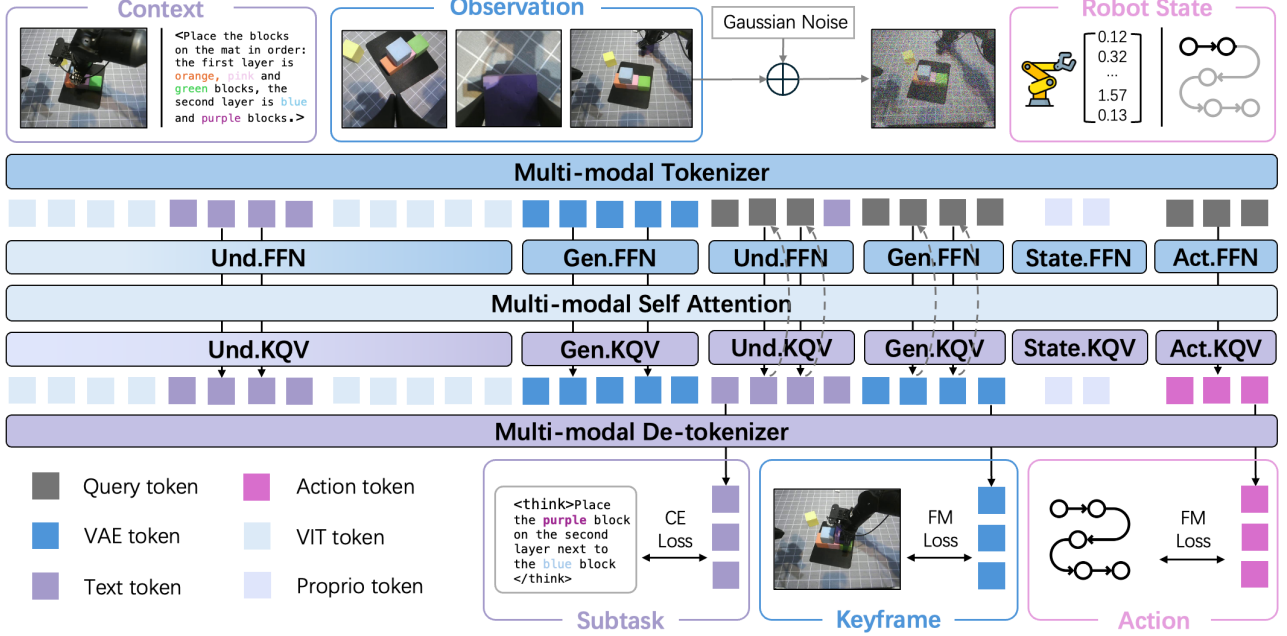


Fig. 2. **Illustration of the BagelVLA framework.** BagelVLA utilizes a Mixture-of-Transformers (MoT) architecture, comprising three independent transformers specialized for linguistic, visual, and action modalities. To tackle long-horizon tasks and semantic generalization, we formulate language-conditioned action learning as a long-sequence interleaved planning problem. As shown, we structure these modalities into a unified sequence, enabling the model to generate predictions across all three modalities based on the interleaved context. To support this architecture, we have designed specific mechanisms to facilitate interaction among multiple flow-matching experts and to enhance inference efficiency.

features when autoregressively generating the subtask l_t . We optimize textual-planning task using an autoregressive Cross-Entropy (CE) loss: $\mathcal{L}_l = -\log p_\theta(l_t|v_t, L)$. For **visual forecasting**, the generation expert, while denoising the keyframe image, attends to all input views' VAE and ViT features, and relevant textual information. It generates keyframe by iteratively denoising input VAE noise using Flow Matching [28, 32], denoted as: $\mathcal{L}_v = -\log p_\theta(v_{t+k}|v_t, L, l_t)$.

2) **Action Expert:** We employ an independent transformer connected via the MoT framework as the action expert, which is responsible for processing proprioceptive and action modalities. The action expert shares a similar architecture with the Qwen2.5 LLM; however, we reduce the intermediate size of the MLP to 1/5th of the original, resulting in 2B parameters. This compact size facilitates higher execution frequency during inference through KV-cache and asynchronous action generation.

For **action planning**, we similarly employ Flow Matching to learn action chunks, denoted as $\mathcal{L}_a = -\log p_\theta(a_t|v_t, L, l_t, v_{t+k})$. During the denoising process, the action sequence can attend to the VAE and ViT features of the input views, the global instruction L , the generated subtask l_t , and also the proprioceptive state input to the action expert. Notably, the action expert attends to the intermediate latent states of the image currently being generated. This involves handling the asymmetric information interaction between the dual Flow Matching modules of the generation and action experts. We detail the various schemes we explored in Sec. III-D and ablate these methods in Sec. IV-C.

D. Conditioning Schemes in Dual Flow-Matching

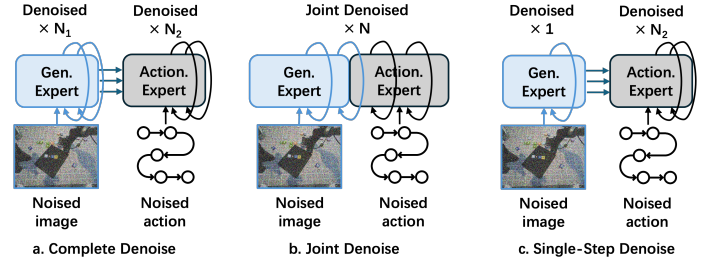


Fig. 3. **Illustration of different types of dual denoising schemes.** (a) Complete Denoise: Image prediction and action generation are performed separately, requiring a total of $N_1 + N_2$ denoising steps. (b) Joint Denoise: Image prediction and action generation are performed simultaneously, denoising together for N steps. (c) Single-Step Denoise: Action generation is conditioned directly on the context from the first denoising step of the image prediction.

In this section, we detail the computation of $\mathcal{L}_v$ and $\mathcal{L}_a$ within a unified interleaved input sequence, ensuring consistency with the inference context. As illustrated in Fig. 3, we propose three interaction mechanisms for the Flow Matching (FM) of keyframe prediction and action generation.

1) **Scheme 1: Complete Denoise:** As shown in Fig. 3(a), Complete Denoise prioritizes the full denoising of the keyframe image by the generation expert. The generated image is then fed back as context for action generation. During training, to ensure the action expert observes a fully denoised image, we append the ground truth keyframe subsequent to the denoising sequence. The loss functions are defined as follows:

$$\begin{aligned}
\mathcal{L}_v &= \mathbb{E} [\|\mathbf{v}_{v,\theta}(L, v_t, l_t, \tau, v_{t+k}^\tau) - (v_{t+k}^1 - v_{t+k}^0)\|_2^2] \\
\mathcal{L}_{a1} &= \mathbb{E} [\|\mathbf{v}_{a,\theta}(L, v_t, l_t, v_{t+k}^{\tau=1}, \tau, a_t^\tau) - (a_t^1 - a_t^0)\|_2^2] \quad (1) \\
v_{t+k}^\tau &= (1 - \tau)v_{t+k}^0 + \tau v_{t+k}^1, \quad v_{t+k}^1 = v_{t+k} \\
a_t^\tau &= (1 - \tau)a_t^0 + \tau a_t^1, \quad a_t^1 = a_t
\end{aligned}$$

where $\mathbf{v}$ denotes the velocity predicted by the model for the corresponding modality. L and l_t represent the global instruction and sub-task plan, v_t is the current input observation, v_{t+k} is the target keyframe and a_t is the action chunk. τ denotes the denoising timestep (where $\tau = 0$ represents initial noise and $\tau = 1$ represents the ground truth target).

This approach effectively combines a World Model (WM) with an Inverse Dynamics Model (IDM) [11]. While theoretically sound for leveraging the WM, it suffers from high inference latency (total denoising steps $N_1 + N_2$) and the potential accumulation of visual errors. To mitigate these issues, we propose alternative schemes.

2) **Scheme 2: Joint Denoise:** As shown in Fig. 3(b), we synchronize the denoising processes of the keyframe and the action. Here, the action generation attends to the noisy image currently undergoing denoising. The computation for the action FM loss in Eq. 1 is modified as:

$$\mathcal{L}_{a2} = \mathbb{E} [\|\mathbf{v}_{a,\theta}(L, v_t, l_t, v_{t+k}^\tau, \tau, a_t^\tau) - (a_t^1 - a_t^0)\|_2^2]$$

During training, we append the action denoising block directly after the keyframe denoising sequence, allowing the action component to attend to the intermediate noisy keyframes. During inference, the model generates both keyframes and actions within N steps, significantly reducing latency.

3) **Scheme 3: Single-step Denoise:** To further minimize the computational cost of action inference imposed by image denoising, we propose single-step denoise. In this scheme, action generation attends only to the KV-cache from the initial denoising step of the keyframe. This implies the model generates actions while conditioning on the initial noise as the keyframe input:

$$\mathcal{L}_{a3} = \mathbb{E} [\|\mathbf{v}_{a,\theta}(L, v_t, l_t, v_{t+k}^{\tau=0}, \tau, a_t^\tau) - (a_t^1 - a_t^0)\|_2^2]$$

Furthermore, based on Scheme 3, we introduce a variant of Single-step Denoise where we inject current frame information into the initial image noise to provide stronger priors for both keyframe and action generation:

$$\text{Naive Single-step Denoise : } v_{t+k}^{\tau=0} \sim \mathcal{N}(0, I) \quad (2)$$

$$\text{Residual Flow Guidance (RFG) : } v_{t+k}^{\tau=0} \sim \mathcal{N}(v_t, I) \quad (3)$$

We provide an ablation study of these methods in Sec. IV-C. Based on the results, we select the Single-step Denoise (RFG) as our default setting for BagelVLA. Notably, we observe that RFG, which incorporates the initial frame v_t prior, significantly reduces the required denoising steps as shown in Fig. 5. We hypothesize that this allows the WM to focus on modeling robot manipulation changes rather than reconstructing static background details. Further quantitative comparisons are available in Sec. IV-C1.

E. Training and Inference Strategy

BagelVLA requires the simultaneous alignment of three distinct planning tasks: linguistic planning, visual forecasting, and action generation. To achieve this, we divide the training process into two stages, maximizing the utilization of general multimodal data and large-scale embodied data without action labels.

1) **Stage 1: Pretraining - Finetuning Linguistic Planning and Learning Visual Dynamics:** In this stage, we exclusively finetune the understanding and generation experts to acquire capabilities in textual planning and keyframe prediction. To preserve the model’s general linguistic proficiency, we co-train with general Question-Answering (QA) data.

2) **Stage 2: Finetuning - Learning Action Planning:** In this stage, we introduce downstream robot data containing action labels for finetuning. We finetune the entire model on all three planning tasks simultaneously to obtain an interleaved planning model that performs robustly in specific scenarios. For the four scenarios used in our experiments, we employ the following finetuning strategies:

- Calvin (Visual & Action Planning): ABC split dataset.
- Robotwin (Linguistic, Visual & Action Planning): 50 tasks with 50 episodes each, totaling 2.5k episodes.
- ALOHA Basic Tasks (Visual & Action Planning): 3k episodes.
- ALOHA Long-horizon Tasks (Linguistic, Visual & Action Planning): 1.5k episodes.

Inference Strategy During inference, the model generates textual plans, keyframes, and actions in an interleaved manner. At each denoising step, only a single expert is activated (7B model for text and keyframe or 2B model for action generation). The single-step denoise scheme further enhances execution frequency. Specifically, we concatenate the current frame, instruction context, and a pure noise image to compute the KV pairs of the understanding and generation experts, which then condition the action generation. This mechanism enables BagelVLA to infer at a speed of 1.2 seconds per chunk on a single RTX 5090 GPU (yielding a real-world action frequency of 40Hz with a chunk size of 48).

We also introduce **Asynchronous Execution** [47, 8] to further boost inference speed. During training, we randomly replace the current frame with a preceding image. This allows us to reduce the updating frequency of the KV contexts of understanding and generation experts during inference, updating only the proprioceptive inputs to output new action chunks. Under this setting, our policy can achieve an execution frequency of 72Hz.

IV. EXPERIMENT

We conducted extensive experiments to evaluate the interleaved planning capabilities of BagelVLA across a diverse range of manipulation tasks. These experiments encompass two simulation environments, Calvin [34] and Robotwin [7], as well as a basic tasks suite containing 9 skills of 30 tasks, and a long-horizon task suite performed on the AgileX dual-arm robot system.

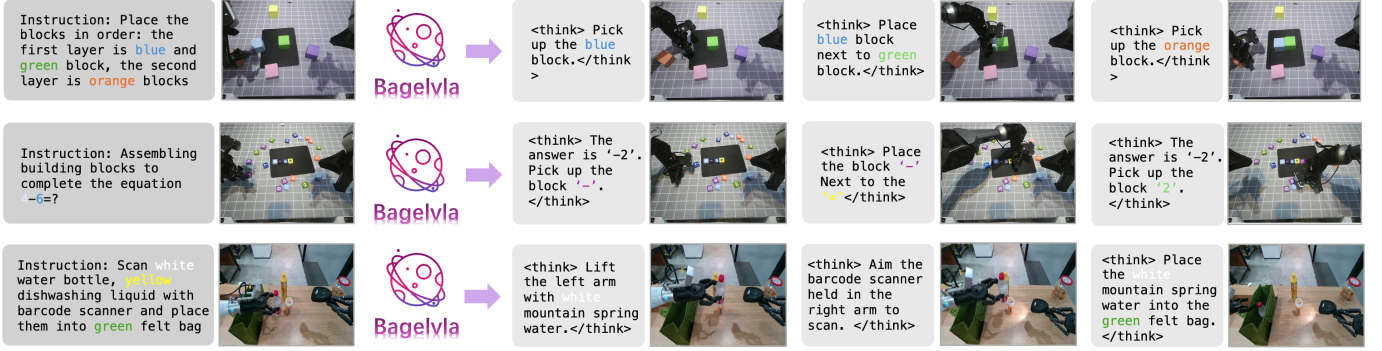


Fig. 4. **Visualization of interleaved planning results on a real-world robotic task.** Given a global instruction and the current observation, BagelVLA leverages the context to identify the immediate subtask, predicts a goal image for that subtask, and subsequently generates actions. The figure illustrates the interleaved planning results for *Stack Cubes in Requested Order*, *Calculate and Place Symbol Blocks* task, and a task from Agibot dataset [1].

A. Evaluation in Simulation Environment

We benchmark BagelVLA against π_0 [2], RDT [31] and two VLA models that incorporate future prediction capabilities, UP-VLA [48] and VPP [16], in the Calvin and Robotwin environments.

In the Calvin environment, models are trained on the ABC split and evaluated in the D environment. For Robotwin, we utilize a training dataset consisting of 50 clean demonstrations for each of the 50 tasks. All models are then tested in both Clean and Randomized settings using unseen instructions. To verify the efficacy of interleaved planning, we conduct experiments with BagelVLA trained and tested both with and without interleaved planning.

TABLE I
RESULTS ON CALVIN AND ROBOTWIN2.0 BENCHMARKS. *Since Calvin consists exclusively of single-step tasks, we did not evaluate BagelVLA’s performance under the textual planning setting in this domain.*

Model	Calvin	Robotwin	
	ABC-D	Clean	Randomized
π_0	3.648	46.42	16.34
RDT	-	34.50	13.72
UP-VLA	4.078	52.92	15.16
VPP	4.329	-	-
w/o Textual-planning	-	54.00	19.20
w/o Keyframe-forecasting	3.345	56.72	17.97
BagelVLA	4.405	75.26	20.87

As presented in Table I, BagelVLA outperforms all baselines on both the Calvin ABC-D split and the Robotwin tasks. BagelVLA achieves an average completion length of 4.41 on the Calvin ABC-D benchmark. This indicates that models leveraging only visual prediction as an auxiliary task can effectively generalize from in-domain training to Out-of-Distribution (OOD) scenarios involving background and color variations, while maintaining high manipulation accuracy.

On the Robotwin benchmark, BagelVLA without textual-planning surpasses π_0 in both Clean and Randomized settings, achieving success rates comparable to UP-VLA, which similarly employs visual prediction as an auxiliary task. This suggests that the visual prediction component within our interleaved planning framework yields consistent gains

across different VLM backbones. However, when incorporating textual-planning, BagelVLA achieves state-of-the-art performance in both in-domain and out-of-domain settings on Robotwin, demonstrating the substantial effectiveness of the proposed interleaved planning scheme.

B. Real-world Experiments

We evaluated BagelVLA on the Aloha-AgileX bimanual robot platform across two categories of dual-arm manipulation tasks. These tasks were designed to assess the model’s performance on both basic tasks and long-horizon tasks that require planning. Specifically, we collect 3,000 trajectories categorized as basic tasks, covering 9 distinct skills ranging from short-horizon tasks such as pick-and-place to medium-horizon tasks such as sweeping rubbish. Furthermore, we designed two types of Long-Horizon planning tasks that necessitate subtask planning, for which we gathered 1,500 demonstrations. All collected data are manually annotated with subtasks and corresponding keyframes. We then fine-tune the pretrained BagelVLA on all trajectories and evaluate its multi-task learning capabilities. We compare BagelVLA with π_0 [2] and VPP [16]. A visualization of interleaved plans generated by BagelVLA for the real-world tasks is illustrated in Fig. 4.

1) *Basic Task Experiments:* Table II presents the performance of BagelVLA in a multi-task setting without the use of subtask planning. BagelVLA achieved the highest average success rate across the 9 categories of tasks, which demonstrates its outstanding multi-task learning capabilities. Additionally, we tested the model on pick-and-place tasks involving unseen objects. As shown, BagelVLA significantly outperforms VPP and π_0 in the OOD setting. This advantage stems from the powerful semantic features preserved during VLA fine-tuning, which are inherited from the pre-training of our understanding and generation experts.

2) *Long-Horizon Planning Task Experiments:* We collected data for two categories of long-horizon tasks that require planning. In the colored block stacking task, shown in the first row of Fig. 4, the model is instructed to stack cubes in the order specified by the instruction. This task challenges both the model’s visual-language interleave planning ability and its capacity to follow instructions at the action level. In the

TABLE II
RESULTS ON REAL-WORLD BASIC TASKS WITHOUT USING SUBTASK PLANNING. We run 20 times for each task.

Model	Pick&Place Seen	Pick&Place Unseen	Water Flower	Stack Cubes	Put Flowers in Vase	Stack Bowls	Pour Fries	Sweep Rubbish	Press Button	Drawer Close	Success Average
π_0	95	55	50	65	40	70	35	55	90	95	65.0
VPP	85	45	60	50	50	55	30	45	75	100	59.5
BagelVLA	95	85	60	80	35	90	45	80	90	95	75.5

TABLE III
RESULTS ON REAL-WORLD LONG-HORIZON PLANNING TASKS. We run 20 times for each task. Difficulty settings for the tasks are defined as follows. For Stack Cubes: Easy (2-3 cubes, 1-2 layer), Middle (3-4 cubes, 2 layers), and Hard (3-5 cubes, 3 layers). For Calculate and Place Symbols: Easy (1-2 blocks for single-digit addition), Middle (2 blocks for the answer of a double-digit addition), and Hard (3-5 blocks within a double-digit addition). The Success Rate column shows the average success rate, while Planning Accuracy indicates correct motion trends across 20 different intermediate states.

Tasks	Stack Cubes in Requested Order					Calculate and Place Symbol Blocks				
Difficulty	Easy	Middle	Hard	Success Rate↑	Planning Accuracy↑	Easy	Middle	Hard	Success Rate↑	Planning Accuracy↑
π_0	75	35	10	40.0	55	70	25	0	31.7	40
VPP	60	15	0	25.0	45	60	10	0	23.3	30
w/o Keyframe-forecasting	90	45	25	53.3	80	70	50	30	50.0	75
w/o Textual-planning	75	40	15	43.3	70	65	25	10	33.3	50
BagelVLA	95	65	60	73.3	95	80	65	45	63.3	85

arithmetic equation arrangement task, shown in the second row, we require the model first to compute an arithmetic expression and then place the corresponding symbolic blocks in a single sequence. The objective of this task is to verify whether the model can retain reasoning capabilities (such as performing simple addition) during the planning process. Table III displays the performance of the three models on these two long-horizon planning tasks. It is evident that although all three models were trained on the exact same action data, BagelVLA, with its interleaved planning capabilities, exhibits a significant advantage in planning-oriented tasks. In addition to the average task success rate, we also measured the correctness of the motion trend for each subtask to assess the model’s semantic understanding and action-following fidelity. Overall, BagelVLA achieved a planning accuracy of nearly 90%, which implies that its multi-modal planning is correct and possesses strong generalization abilities. Concurrently, we observed a gap between task success rate and the planning accuracy, suggesting deficiencies in action mapping due to limitations in both the model and the dataset, specifically concerning the precision of fine-motor control.

C. Ablation Study

We conduct comprehensive ablation studies on the various modules of BagelVLA in both simulated and real-world environments. Through these experiments, we aim to answer the following questions:

- 1) What is the optimal interaction mechanism between the generation experts and the action experts?
- 2) How does RFG outperform naive single-step denoising?
- 3) What is the effect of BagelVLA’s pre-training on action execution performance?
- 4) Does each modality within the interleaved planning framework contribute positively to the action generation process?

1) Comparison of Different Conditioning Schemes in Dual Flow-Matching: We evaluate the three dual flow-matching interaction schemes described in Sec. III-D within the Calvin ABC-D environment. For complete denoise and joint denoise, the image noise initialization follows the formulation in Eq. 2. We utilize single-view inputs and train each method for 10k steps for testing. We denoise 50 times for image generation and 10 times for action generation. Table IV reports the average task completion length and the inference latency per action chunk for each interaction method, measured 20 times on a single NVIDIA A800 GPU.

TABLE IV
ABLATION ON DIFFERENT CONDITIONING SCHEMES AND RFG MECHANISM ON CALVIN SINGLE-VIEW SETTING.

Dual Flow-Matching Schemes	Latency $\downarrow$	ABC-D $\uparrow$
Complete Denoise	6.12s	2.480
Joint Denoise	2.94s	2.038
Single-step Denoise (Eq. 2)	1.41s	3.345
RFG (Eq. 3)	1.41s	3.600

The results indicate that the *single-step denoising* strategy not only significantly outperforms the other two approaches in terms of task success rate but also achieves superior inference speed. This performance gap can be attributed to the domain shift introduced during testing, where the model encounters scenes with altered color schemes. Under these conditions, models employing *complete denoising* or *joint denoising* are prone to encountering out-of-distribution (OOD) intermediate states during the flow-matching phase of the generation expert. This consequently leads to a substantial degradation in action execution performance. Based on these empirical findings, we adopt *single-step denoising* as the default interaction mechanism for the dual flow-matching framework in BagelVLA across all subsequent scenarios and tasks.

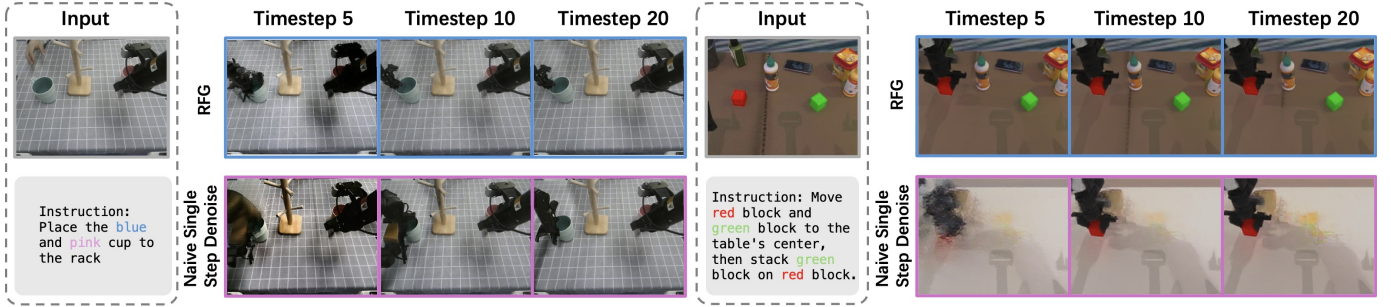


Fig. 5. **Predicted images using different denoising steps.** The figure displays the generation results for the naive single-step denoise (Eq. 2) and RFG (Eq. 3) across varying denoising steps in real-world basic tasks and Robotwin randomized (unseen) scenarios. RFG demonstrates the capability to preserve backgrounds and achieve high-quality generation with very few steps. This provides strong support for reducing the inference latency of interleaved generation.

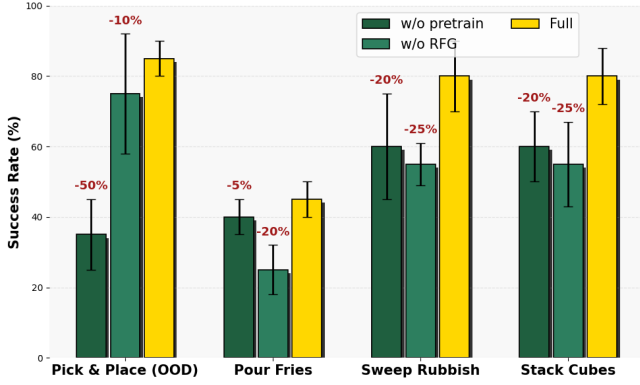


Fig. 6. **Ablation on Real-World Basic Tasks**

2) Advantages of RFG over Naive Single-Step Denoising:

In contrast to the conventional naive single-step denoising approach, which employs Eq. 2 for noise initialization, RFG utilizes Eq. 3. We compare these two methods across both the Calvin simulation environment and real-world basic tasks. In the real-world basic tasks shown in Fig. 6, RFG demonstrates significantly superior performance compared to naive single-step denoising on several tasks. Concurrently, as shown in Table IV, RFG achieves faster action learning convergence while maintaining the low inference latency characteristic of naive single-step denoising. This improvement stems from the fact that in the naive approach, action generation relies on intermediate features derived from a single denoising step on pure Gaussian noise. Conversely, RFG incorporates the initial frame into the noise initialization, thereby providing stronger prior information for action generation.

Furthermore, we observe that RFG offers a distinct advantage in keyframe prediction, even though fully denoising the keyframe is not strictly required for action generation. Fig. 5 visualizes the predicted keyframes for both RFG and naive single-step denoising across different denoising steps. It is evident that RFG is capable of generating high-quality future frames with very few denoising steps (e.g., 10 steps). We hypothesize that this phenomenon arises because the inclusion of the first frame in Eq. 3 allows the model to focus its capacity on the dynamic regions, rather than learning complex static background information.

3) Effectiveness of Large-Scale Language Planning and Visual Dynamics Pre-training:

In the real-world basic tasks, we evaluate the impact of pre-training. By comparing the w/o pretrain variant with baseline in Fig. 6, it is evident that the pre-trained baseline achieves a significantly higher success rate on pick&place (OOD) tasks. This indicates that pre-training solely on linguistic planning and visual forecasting is sufficient to enhance the model’s semantic generalization capabilities. Furthermore, on three medium-horizon tasks (including sweep rubbish, pour fries and stack cubes), the model utilizing joint pre-training exhibits higher accuracy. We attribute this improvement to the planning capabilities acquired from the language planning tasks during pre-training. During the subsequent action fine-tuning phase, the model retains these state prediction and planning capabilities, thereby enabling it to perform implicit subtask planning even without explicitly utilizing interleaved planning during inference.

4) Effectiveness of Visual and Language Modalities in Interleaved Planning:

To verify the effectiveness of the interleaved planning mechanism, we investigate the performance impact of omitting textual planning and keyframe forecasting, respectively.

- **Linguistic Planning:** The results in Table I demonstrate that employing textual planning with BagelVLA in RoboTwin environment improves the success rate by 21%. Similarly, in the two categories of real-world long-horizon tasks shown in Table III, the use of textual planning also yields substantial performance gains. These two sets of experiments conclusively prove that incorporating language planning significantly benefits long-horizon tasks.
- **Visual Forecasting:** The keyframe variant in Table I and III illustrates the impact of using keyframe prediction as a training objective. The results indicate that visual planning markedly improves the accuracy of action planning in both simulation and real-world environments.

The aforementioned experiments confirm that both visual planning and language planning play crucial roles within the interleaved planning framework.

V. CONCLUSION

We presented BagelVLA, a unified Vision-Language-Action framework for long-horizon manipulation by interleaving linguistic planning, visual forecasting, and action generation within a single transformer system. Building on Bagel’s unified multimodal backbone, we introduce an action expert and adopt a two-stage training recipe to progressively transfer multimodal reasoning and visual dynamics into embodied planning, then couple these representations with control. To address the latency of visual foresight, we further propose Residual Flow Guidance (RFG), which captures task-relevant future dynamics with substantially reduced computational costs. Overall, our results suggest that explicitly coupling linguistic planning with predictive visual representations can improve robustness and instruction-following in long-horizon manipulation.

REFERENCES

- [1] AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyan Gao, Xindong He, Xuan Hu, Xu Huang, Shu Jiang, Yuxin Jiang, Cheng Jing, Hongyang Li, Jialu Li, Chiming Liu, Yi Liu, Yuxiang Lu, Jianlan Luo, Ping Luo, Yao Mu, Yuehan Niu, Yixuan Pan, Jiangmiao Pang, Yu Qiao, Guanghui Ren, Cheng Ruan, Jiaqi Shan, Yongjian Shen, Chengshi Shi, Mingkan Shi, Modi Shi, Chonghao Sima, Jianheng Song, Huijie Wang, Wenhao Wang, Dafeng Wei, Chengen Xie, Guo Xu, Junchi Yan, Cunbiao Yang, Lei Yang, Shukai Yang, Maoqing Yao, Jia Zeng, Chi Zhang, Qinglin Zhang, Bin Zhao, Chengyue Zhao, Jiaqi Zhao, and Jianchao Zhu. Agibot world colosseum: A large-scale manipulation platform for scalable and intelligent embodied systems, 2025. URL <https://arxiv.org/abs/2503.06669>.
- [2] 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.
- [3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huang Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. URL <https://arxiv.org/abs/2307.15818>.
- [4] 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.
- [5] 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, 2024. URL <https://arxiv.org/abs/2410.06158>.
- [6] Chilam Cheang, Sijin Chen, Zhongren Cui, Yingdong Hu, Liqun Huang, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Xiao Ma, Hao Niu, Wenxuan Ou, Wanli Peng, Zeyu Ren, Haixin Shi, Jiawen Tian, Hongtao Wu, Xin Xiao, Yuyang Xiao, Jiafeng Xu, and Yichu Yang. Gr-3 technical report,

2025. URL <https://arxiv.org/abs/2507.15493>.
- [7] Tianxing Chen, Zanzin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.
- [8] Can Cui, Pengxiang Ding, Wenxuan Song, Shuanghao Bai, Xinyang Tong, Zirui Ge, Runze Suo, Wanqi Zhou, Yang Liu, Bofang Jia, et al. Openhelix: A short survey, empirical analysis, and open-source dual-system vla model for robotic manipulation. *arXiv preprint arXiv:2505.03912*, 2025.
- [9] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.
- [10] 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.
- [11] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in Neural Information Processing Systems*, 36, 2024.
- [12] Huang Fang, Mengxi Zhang, Heng Dong, Wei Li, Zixuan Wang, Qifeng Zhang, Xueyun Tian, Yucheng Hu, and Hang Li. Robix: A unified model for robot interaction, reasoning and planning, 2025. URL <https://arxiv.org/abs/2509.01106>.
- [13] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, Jingji Chen, Jingjia Huang, Kang Lei, Liping Yuan, Lishu Luo, Pengfei Liu, Qinghao Ye, Rui Qian, Shen Yan, Shixiong Zhao, Shuai Peng, Shuangye Li, Sihang Yuan, Sijin Wu, Tianheng Cheng, Weiwei Liu, Wenqian Wang, Xianhan Zeng, Xiao Liu, Xiaobo Qin, Xiaohan Ding, Xiaojun Xiao, Xiaoying Zhang, Xuanwei Zhang, Xuehan Xiong, Yanghua Peng, Yangrui Chen, Yanwei Li, Yanxu Hu, Yi Lin, Yiyuan Hu, Yiyuan Zhang, Youbin Wu, Yu Li, Yudong Liu, Yue Ling, Yujia Qin, Zhanbo Wang, Zhiwu He, Aoxue Zhang, Bairen Yi, Bencheng Liao, Can Huang, Can Zhang, Chaorui Deng, Chaoyi Deng, Cheng Lin, Cheng Yuan, Chenggang Li, Chenhui Gou, Chenwei Lou, Chengzhi Wei, Chundian Liu, Chunyuan Li, Deyao Zhu, Donghong Zhong, Feng Li, Feng Zhang, Gang Wu, Guodong Li, Guohong Xiao, Haibin Lin, Haihua Yang, Haoming Wang, Heng Ji, Hongxiang Hao, Hui Shen, Huixia Li, Jiahao Li, Jialong Wu, Jianhua Zhu, Jianpeng Jiao, Jiashi Feng, Jiaze Chen, Jianhui Duan, Jihao Liu, Jin Zeng, Jingqun Tang, Jingyu Sun, Jia Chen, Jun Long, Junda Feng, Junfeng Zhan, Junjie Fang, Junting Lu, Kai Hua, Kai Liu, Kai Shen, Kaiyuan Zhang, Ke Shen, Ke Wang, Keyu Pan, Kun Zhang, Kunchang Li, Lanxin Li, Lei Li, Lei Shi, Li Han, Liang Xiang, Liangqiang Chen, Lin Chen, Lin Li, Lin Yan, Liying Chi, Longxiang Liu, Mengfei Du, Mingxuan Wang, Ningxin Pan, Peibin Chen, Pengfei Chen, Pengfei Wu, Qingqing Yuan, Qingyao Shuai, Qiuyan Tao, Renjie Zheng, Renrui Zhang, Ru Zhang, Rui Wang, Rui Yang, Rui Zhao, Shaoqiang Xu, Shihao Liang, Shipeng Yan, Shu Zhong, Shuaishuai Cao, Shuangzhi Wu, Shufan Liu, Shuhan Chang, Songhua Cai, Tenglong Ao, Tianhao Yang, Tingting Zhang, Wanjuan Zhong, Wei Jia, Wei Weng, Weihao Yu, Wenhao Huang, Wenjia Zhu, Wenli Yang, Wenzhi Wang, Xiang Long, Xiangrui Yin, Xiao Li, Xiaolei Zhu, Xiaoying Jia, Xijin Zhang, Xin Liu, Xincheng Zhang, Xinyu Yang, Xiongcai Luo, Xiuli Chen, Xuandong Zhong, Xuefeng Xiao, Xujing Li, Yan Wu, Yawei Wen, Yifan Du, Yihao Zhang, Yining Ye, Yonghui Wu, Yu Liu, Yu Yue, Yufeng Zhou, Yufeng Yuan, Yuhang Xu, Yuhong Yang, Yun Zhang, Yunhao Fang, Yuntao Li, Yurui Ren, Yuwen Xiong, Zehua Hong, Zehua Wang, Zewei Sun, Zeyu Wang, Zhao Cai, Zhaoyue Zha, Zhecheng An, Zhehui Zhao, Zhengzhuo Xu, Zhipeng Chen, Zhiyong Wu, Zhuofan Zheng, Zihao Wang, Zilong Huang, Ziyu Zhu, and Zuquan Song. Seed1.5-vl technical report, 2025. URL <https://arxiv.org/abs/2505.07062>.
- [14] Yanjiang Guo, Yucheng Hu, Jianke Zhang, Yen-Jen Wang, Xiaoyu Chen, Chaochao Lu, and Jianyu Chen. Prediction with action: Visual policy learning via joint denoising process. *Advances in Neural Information Processing Systems*, 37:112386–112410, 2024.
- [15] Ryan Hoque, Peide Huang, David J Yoon, Mouli Siva-
purapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.
- [16] 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.
- [17] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Es-
mail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. pi0.5: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.
- [18] Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianning Cui, Xiao Liu, Shuiqi Cheng, Jiyang Gao, Huazhe Xu, and Hang Zhao. Galaxea open-world dataset and g0 dual-system vla model, 2025. URL <https://arxiv.org/abs/2509.00576>.
- [19] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragki-
adaki. 3d diffuser actor: Policy diffusion with 3d scene representations, 2024. URL <https://arxiv.org/abs/2402.10885>.
- [20] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ash-
win Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yun-
liang 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, Vitor Guizilini, David Antonio Herrera, Minh Heo, Kyle Hsu, Jiaheng Hu, Muhammad Zubair Irshad, Donovan 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.
- [21] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model, 2024. URL <https://arxiv.org/abs/2406.09246>.
- [22] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.
- [23] Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.
- [24] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- [25] Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. Zebra-cot: A dataset for interleaved vision language reasoning, 2025. URL <https://arxiv.org/abs/2507.16746>.
- [26] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabriel Levine, Wensi Ai, Benjamin Martinez, Hang Yin, Michael Lingelbach, Minjune Hwang, Ayano Hiranaka, Sujay Garlanka, Arman Aydin, Sharon Lee, Jiankai Sun, Mona Anvari, Manasi Sharma, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Yunzhu Li, Silvio Savarese, Hyowon Gweon, C. Karen Liu, Jiajun Wu, and Li Fei-Fei. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation, 2024. URL <https://arxiv.org/abs/2403.09227>.
- [27] Wei Li, Renshan Zhang, Rui Shao, Jie He, and Liqiang Nie. Cogvla: Cognition-aligned vision-language-action model via instruction-driven routing & sparsification, 2025. URL <https://arxiv.org/abs/2508.21046>.
- [28] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.
- [29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL <https://arxiv.org/abs/2304.08485>.
- [30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. URL <https://arxiv.org/abs/2310.03744>.
- [31] 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.
- [32] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.
- [33] Qi Lv, Weijie Kong, Hao Li, Jia Zeng, Zherui Qiu, Delin Qu, Haoming Song, Qizhi Chen, Xiang Deng, and Jiangmiao Pang. F1: A vision-language-action model bridging understanding and generation to actions. *arXiv preprint arXiv:2509.06951*, 2025.
- [34] 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.
- [35] Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint arXiv:2512.15692*, 2025.
- [36] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Lmfusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024.
- [37] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- [38] Gemini Robotics Team, Abbas Abdolmaleki, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac,

- Montserrat Gonzalez Arenas, Ashwin Balakrishna, Nathan Batchelor, Alex Bewley, Jeff Bingham, et al. Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. *arXiv preprint arXiv:2510.03342*, 2025.
- [39] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy, 2024. URL <https://arxiv.org/abs/2405.12213>.
- [40] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.
- [41] 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.
- [42] Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. Finevision: Open data is all you need, 2025. URL <https://arxiv.org/abs/2510.17269>.
- [43] Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, YINUO Zhao, Zhiyuan Xu, Guang Yang, Shichao Fan, Xinhua Wang, Fei Liao, Zhen Zhao, Guangyu Li, Zhao Jin, Lecheng Wang, Jilei Mao, Ning Liu, Pei Ren, Qiang Zhang, Yaoyu Lyu, Mengzhen Liu, Jingyang He, Yulin Luo, Zeyu Gao, Chenxuan Li, Chenyang Gu, Yankai Fu, Di Wu, Xingyu Wang, Sixiang Chen, Zhenyu Wang, Pengju An, Siyuan Qian, Shanghang Zhang, and Jian Tang. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation, 2025. URL <https://arxiv.org/abs/2412.13877>.
- [44] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.
- [45] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models, 2025. URL <https://arxiv.org/abs/2506.15564>.
- [46] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [47] Jianke Zhang, Yanjiang Guo, Xiaoyu Chen, Yen-Jen Wang, Yucheng Hu, Chengming Shi, and Jianyu Chen. Hirt: Enhancing robotic control with hierarchical robot transformers. *arXiv preprint arXiv:2410.05273*, 2024.
- [48] Jianke Zhang, Yanjiang Guo, Yucheng Hu, Xiaoyu Chen, Xiang Zhu, and Jianyu Chen. Up-vla: A unified understanding and prediction model for embodied agent. *arXiv preprint arXiv:2501.18867*, 2025.
- [49] Jianke Zhang, Yucheng Hu, Yanjiang Guo, Xiaoyu Chen, Yichen Liu, Wenna Chen, Chaochao Lu, and Jianyu Chen. Unicod: Enhancing robot policy via unified continuous and discrete representation learning. *arXiv preprint arXiv:2510.10642*, 2025.

TABLE V
PARAMETERS OF MODEL ARCHITECTURE

Modules	Understanding Expert	Generation Expert	Action Expert
Size	7B	7B	2B
Input Modality	Image/Text	Image	Proprio/Action
Output Modality	Text	Image	Action
Encoder	ViT+MLP	VAE+MLP	MLP
Image Resolution	256×256	256×256 (VAE)	—
Hidden size	3584	3584	3584
Intermediate size	18944	18944	3584
Layers	28	28	28
Loss Type	CE	MSE(FM)	MSE(FM)
FM Timestep Distribution	—	LogitNormal(0,1)	Beta(1.5,1)

VI. APPENDIX

A. Details of Model Architecture

The architecture of each expert in BagelVLA is detailed in the table below.

B. Dual Denoise Flow-Matching Implementation Details

Here, we demonstrate how to implement the three dual flow-matching methods mentioned in Sec. III-D. Specifically, this requires designing a unified multi-task attention mask for training, enabling a single input sequence to be used for the simultaneous computation of multiple task losses. When designing the corresponding interleaved sequences, we must not only prevent information leakage between different modalities but also align the training setup with special conditions encountered during inference, such as the time-sampling discrepancies that arise from varying numbers of denoising steps. We visualize the masking strategy used in our experiments in Fig. 7.

C. Data Details

1) *Stage 1: Pretraining - Finetuning Language Planning and Learning Visual Dynamics*: In this stage, we exclusively finetune the Understanding and Generation Experts to acquire capabilities in sub-task planning and keyframe prediction. To preserve the model’s general linguistic proficiency, we co-train with general Question-Answering (QA) data. Specifically, the pretraining dataset comprises:

- General VQA (Language Co-training): 2.56M QA pairs.
- Human-hand Data (Visual Dynamics): 310k episodes.
- Open-source Robot Data (Language Planning & Visual Dynamics): 382k episodes.
- Self-collected Real Robot Data (Language Planning & Visual Dynamics): 4.5k episodes.

2) *Stage 2: Finetuning - Learning Action Planning*: In this stage, we introduce downstream robot data containing action labels for finetuning. We finetune the entire model on all three planning tasks simultaneously to obtain an interleaved planning model that performs robustly in specific scenarios. For the four scenarios used in our experiments, we employ the following finetuning strategies:

- **Calvin (Visual Dynamics & Action Planning)**: ABC dataset.

TABLE VI
DETAILS OF DATA COMPONENT.

Task Name	Dataset Name	Number of Samples
General VQA	LLaVA-Pretrain[29]	558k
	FineVision[42]	2M
Open-source Robot Data	AgibotWorld[1]	120k
	GR[6]	80k
	Galaxea Open-World[18]	99k
	Bridge[41]	55k
	Robotwin[7]	27.5k
Human-hand Data	Egodex[15]	310k
Self-collected Data	Aloha	4.5k

- **Robotwin (Language Planning, Visual Dynamics & Action Planning)**: 50 tasks with 50 episodes each, totaling 2.5k episodes.
- **Aloha Short-horizon Tasks (Visual Dynamics & Action Planning)**: 3k episodes.
- **Aloha Long-horizon Tasks (Visual Dynamics & Action Planning)**: 1.5k episodes.

3) *Implementation Details about task Annotation*: For open-source robotic datasets without subtask annotations, such as Bridge, we use the prompt template in Fig. 12 and apply Seed-1.5-VL-thinking to process videos (or image sequences) solely for pretraining. For datasets that do not provide the overall task descriptions(e.g., EgoDex, AgiBot), we adopt the prompt template in Fig. 13 to extract a global task description used for planning or keyframe prediction.

D. Training and Evaluation Details

For all our experiments, we used a learning rate of 1e-5 and employed packed datasets within the FSDP framework to maximize resource utilization. Pre-training was conducted on 64 A800 GPUs with a batch size of approximately 1600 for 20,000 steps. For action fine-tuning and evaluation, we adopted different settings for various downstream scenarios:

- **Calvin ABC-D Simulation Environment**: We trained on 8 A800 GPUs (effective batch size 192) for 30,000 steps. We used an action chunk size of 10, did not include proprioceptive input, and used two camera views as input to predict only the third view. For evaluation, we tested on 1,000 tasks of length 5 from the D-split and reported the mean task completion length.
- **Robotwin Simulation Environment**: We trained on 8 A800 GPUs using 2,500 clean demonstrations (effective batch size 128) for 60,000 steps. We used an action chunk size of 16, sampling one action every 3 steps (effective action horizon of 48). All three camera views were used as input, and we predicted the primary view image. For evaluation, we tested 100 times on 50 tasks in both Clean and Randomized settings using unseen instructions and reported the success rate.
- **Real-Robot Tasks**: We trained on 32 A800 GPUs (effective batch size 512) for 50,000 steps. We used an action chunk size of 24, inputting three views (primary, left

wrist, right wrist) and predicting the primary view image. For evaluation, we tested each task type 20 times with randomized initial positions and added distractor objects. For OOD tasks, we included unseen target objects.

E. Detailed Results in Simulation Environments

Table VIII and VII present more detailed experimental results from the simulation environments.

F. Evaluation Demos of Real-World Tasks

In this section, we detail the setup for two categories of real-robot tasks: Basic Tasks and Long-Horizon Planning Tasks. We also present demo videos of BagelVLA performing on each task type.

1) *Basic Tasks*: During testing, we incorporate several kinds of randomness to evaluate robustness and generalization: *Novel Objects*: Adding unseen objects. *Distractors*: Operating in the presence of irrelevant distractor objects. *Visual Variations*: Adapting to changes in background color and object color. The task suite for the basic tasks on the 14-DOF dual arm includes:

- **Pick & Place**: Grasping and placing a wide range of objects. The training set includes toy fruits, a computer mouse, colorful blocks, toy phones, and so on. The placed targets include colorful plates, baskets, boxes, and so on.
- **Pick & Place Unseen**: Grasping and placing unseen objects to unseen targets. We tested picking up OOD objects such as pears, peaches, a purple block, and placing to novel targets, like pink plates, transparent plates, pink blocks, and so on. We found that although the training set scenes did not involve numerous distractor objects or unseen items, the model still robustly generalizes to new objects and targets with the correct semantics.
- **Water Flower**: This task involves grasping the handle of a toy watering can to simulate the pouring action of watering a plant. It rigorously tests the model’s fine-grained manipulation capabilities, as any action error could easily result in a failure to grasp the handle or align with the flowerpot.
- **Stack Cubes**: The training data includes blocks of four different colors. The instructions require stacking several of these blocks together (up to three high), but without a specific order.
- **Put Flowers in Vase**: Grasp a bouquet lying flat on the table and insert it into a vase. This task requires the model to precisely grasp the thin stems of the bouquet and align them with the opening of the vase, testing the accuracy of the manipulation.
- **Stack Bowls**: Stack bowls of three different colors according to a specified color sequence. This task evaluates the model’s robustness to object positions and its ability to follow language instructions.
- **Pour Fries**: Open the lid of a carton and pour the toy fries from inside it onto a plate. This is a relatively long-horizon task that requires the model to autonomously determine the next action based on its current progress. It tests both manipulation accuracy and long-horizon task capabilities.
- **Sweep Rubbish**: Grasp a toy broom, sweep the randomly placed tissue paper trash on the table into a dustpan, and then put down the broom. This is a task that combines both long-horizon planning and dynamic control. The model must not only assess its current progress but also increase the sweeping speed to ensure the tissue paper rolls into the dustpan.
- **Press Button**: Press different buttons in a specified color sequence. This is a simple long-horizon task that also tests the model’s semantic following capabilities.
- **Drawer Operation**: Opening and closing a drawer. This task primarily evaluates the accuracy of the manipulation.

Fig. 8 illustrates several test scenarios for the basic tasks and presents video recordings of the model’s performance.

2) *Long-Horizon Planning Tasks*: We designed two distinct types of long-horizon planning tasks: (1) Stack Cubes in Requested Order and (2) Calculate and Place Symbol Blocks. We will now detail the setup for each and showcase corresponding demonstration videos.

a) *Stack Cubes in Requested Order*: This task requires the model to stack scattered, multi-colored cubes from the tabletop into a structure that matches a specified shape and sequence given by a language instruction. The target structures can range from one to three layers, with each layer containing one to three cubes. An example instruction is: *Place the cubes in order: the first layer is a blue and a green block, the second layer is an orange block*. The model must perform interleaved planning at each step based on this high-level command. This task involves a very long sequence of actions, posing a significant semantic-following challenge for conventional methods that do not employ explicit planning. In our experiments in Sec. IV-B2, we demonstrate that our method holds a distinct advantage on such long-horizon tasks.

b) *Calculate and Place Symbol Blocks*: This task requires the model to assemble scattered number and symbol blocks to form an arithmetic equation specified by a language instruction, such as: *Assemble the building blocks to complete the equation $21+3=?$* . The initial scene may already contain partially arranged blocks, forcing the model to autonomously decide which block to grasp and place next. It must also place the correct blocks representing the calculated result. Similar to the stacking task, this task also involves long-horizon operational planning. Beyond that, it introduces an additional layer of complexity by requiring a Chain-of-Thought (CoT) process: the model must first leverage the mathematical reasoning capabilities of the general-purpose VLM to compute the result, and then map this result back to the planning and action space. We use this task to validate the effectiveness and generalization capabilities of our interleaved planning framework on long-horizon reasoning tasks.

Fig. 9 illustrates several test scenarios for the long-horizon planning tasks and presents video recordings of the model’s performance.

TABLE VII
EVALUATION ON ROBOTWIN 2.0 SIMULATION (CLEAN VS RANDOMIZED, 50 TASKS). THE TABLE SHOWS SUCCESS RATES IN PERCENT (%) FOR VARIOUS MODELS. MODELS ARE TRAINED USING 50 CLEAN DEMOS PER TASK, AND EVALUATED USING UNSEEN INSTRUCTIONS.

Robotwin Tasks	π_0		RDT		UP-VLA		w/o Textual		w/o Keyframe		BagelVLA	
	Clean	Random	Clean	Random	Clean	Random	Clean	Random	Clean	Random	Clean	Random
Adjust Bottle	90	56	81	75	100	17	100	7	99	4	100	14
Beat Block Hammer	43	21	77	37	66	16	63	18	80	13	87	16
Blocks Ranking Rgb	19	5	3	0	38	0	32	2	46	25	84	4
Blocks Ranking Size	7	1	0	0	21	0	19	0	23	5	45	2
Click Alarmclock	63	11	61	12	69	41	84	60	95	43	85	20
Click Bell	44	3	80	9	54	72	78	60	98	29	100	35
Dump Bin Bigbin	83	24	64	32	81	35	67	26	87	41	91	51
Grab Roller	96	80	74	43	99	28	100	63	97	37	99	41
Handover Block	45	8	45	14	4	0	0	0	18	1	38	0
Handover Mic	98	13	90	31	45	0	76	0	44	3	75	8
Hanging Mug	11	3	23	16	4	0	6	0	2	1	12	1
Lift Pot	84	36	72	9	20	0	0	0	64	7	87	32
Move Can Pot	58	21	25	12	48	0	51	0	9	2	78	0
Move Pillbottle Pad	21	1	8	0	51	7	60	2	22	3	92	1
Move Playingcard Away	53	22	43	11	79	13	86	6	64	31	92	30
Move Stapler Pad	0	2	2	0	8	0	5	0	6	1	27	0
Open Laptop	85	46	59	32	86	21	57	13	62	3	96	37
Open Microwave	80	50	37	20	2	7	0	5	8	14	0	0
Pick Diverse Bottles	27	6	2	0	52	18	74	22	15	11	83	34
Pick Dual Bottles	57	12	42	13	82	31	89	33	33	9	93	56
Place A2b Left	31	1	3	1	74	4	59	7	50	15	79	12
Place A2b Right	27	6	1	1	56	1	53	6	55	19	81	11
Place Bread Basket	17	4	10	2	63	20	71	29	42	17	90	29
Place Bread Skillet	23	1	5	1	71	16	82	26	62	2	91	26
Place Burger Fries	80	4	50	27	97	26	95	56	55	2	99	11
Place Can Basket	41	6	19	6	20	0	37	1	8	0	63	0
Place Cans Plasticbox	34	2	6	5	66	24	23	40	46	6	94	5
Place Container Plate	88	45	78	17	86	48	97	71	82	55	100	58
Place Dual Shoes	15	0	4	4	45	0	36	12	21	0	57	0
Place Empty Cup	37	11	56	7	74	27	94	34	76	35	97	34
Place Fan	20	10	12	2	31	1	15	3	18	2	62	5
Place Mouse Pad	7	1	1	0	27	0	14	10	18	12	46	14
Place Object Basket	16	2	33	17	56	1	44	1	40	6	66	3
Place Object Scale	10	0	1	0	36	4	46	7	31	8	71	0
Place Object Stand	36	11	15	5	76	24	77	35	45	27	87	21
Place Phone Stand	35	7	15	6	32	0	48	0	33	9	61	2
Place Shoe	28	6	35	7	76	12	63	15	44	23	90	29
Press Stapler	62	29	41	24	79	56	59	50	93	52	94	58
Put Bottles Dustbin	54	13	21	4	7	0	12	0	10	0	42	10
Put Object Cabinet	68	18	33	18	7	0	45	4	21	1	52	0
Rotate Qrcode	68	15	50	5	56	2	68	3	72	4	81	21
Scan Object	18	1	4	1	47	23	66	22	38	3	77	32
Shake Bottle Horizontally	99	51	84	51	100	68	99	84	87	60	100	73
Shake Bottle	97	60	74	45	98	54	98	82	83	44	100	74
Stack Blocks Three	17	0	2	0	8	0	15	0	5	2	45	5
Stack Blocks Two	42	1	21	2	61	0	59	2	29	31	95	6
Stack Bowls Three	66	24	51	17	42	1	42	7	37	12	63	13
Stack Bowls Two	91	41	76	30	69	12	70	21	88	48	90	52
Stamp Seal	3	4	1	0	34	2	29	1	23	8	77	8
Turn Switch	27	23	35	15	43	26	37	14	52	10	49	30
Average	46.42	16.34	34.50	13.72	52.92	15.16	54.00	19.20	56.72	15.92	75.26	20.87

G. More Interleaved Planning Visualizations on diverse robotic Tasks

Similar to Fig. 4, in Fig. 10 we provide additional results of interleaved planning in real-world scenarios for reference.

H. More Comparison using RFG and Naive Single-Step Denoising

In Fig. 11, we provide additional comparison using RFG and naive single-step denoising for reference.

I. Usage of LLMs

In the final stages of preparing this manuscript, the authors used a Large Language Model (LLM) solely for grammar checking and language polishing. The model assisted in improving sentence structure and correcting grammatical errors to enhance readability.

TABLE VIII
DETAILED RESULTS OF EVALUATION ON THE CALVIN ABC→D BENCHMARK.

Method	Tasks completed in a row					Avg. Len $\uparrow$
	1	2	3	4	5	
π_0^*	0.937	0.832	0.740	0.629	0.510	3.65
UP-VLA	0.928	0.865	0.815	0.769	0.699	4.08
VPP	0.965	0.909	0.866	0.820	0.769	4.33
w/o Keyframe-forecasting	0.909	0.792	0.676	0.546	0.422	3.35
BagelVLA (Ours)	0.993	0.954	0.893	0.824	0.741	4.41

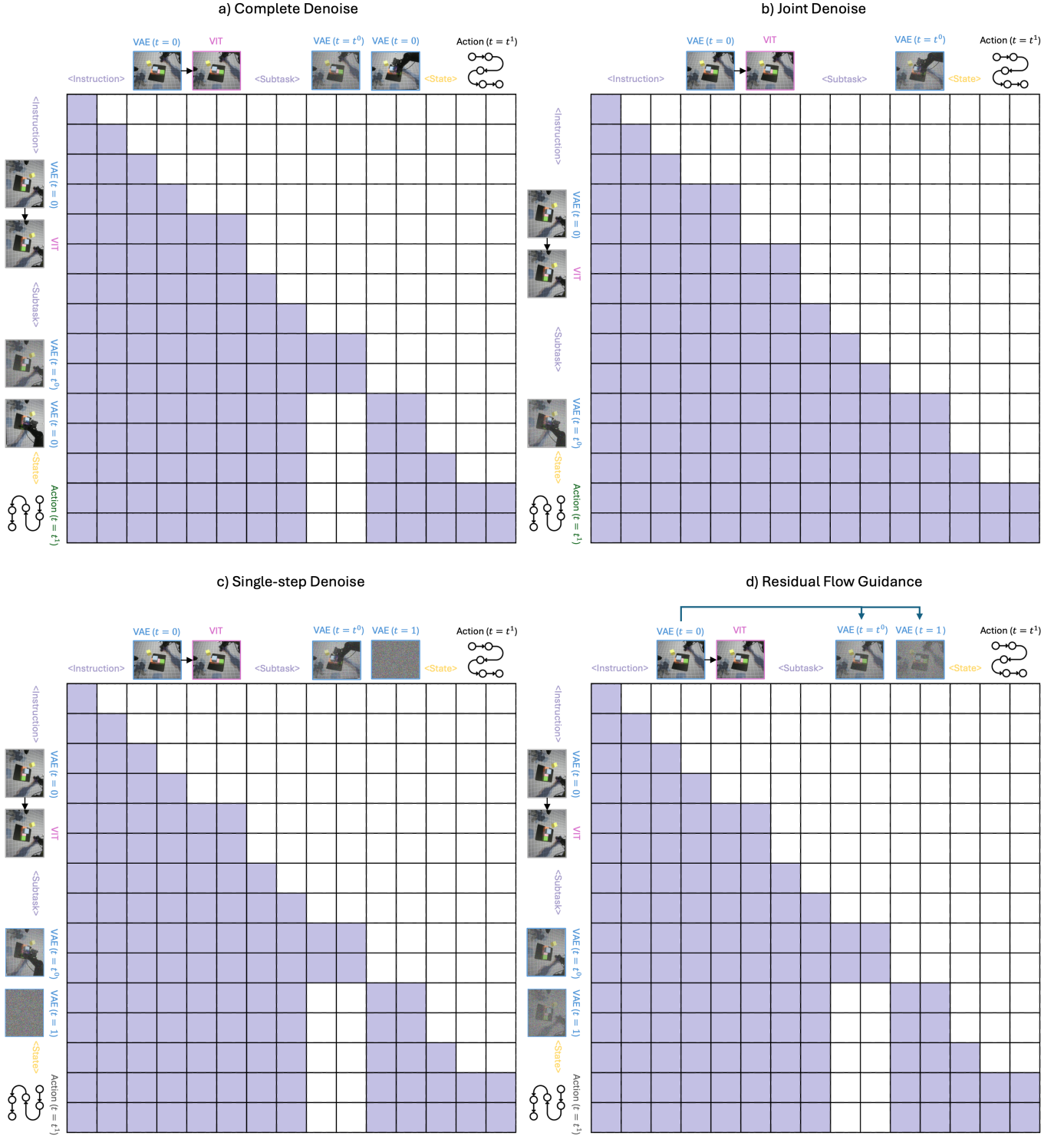


Fig. 7. **Attention Mask used for Different Conditioning Schemes.** (a) Complete Denoise: Image prediction and action generation are performed separately, requiring a total of $N_1 + N_2$ denoising steps. (b) Joint Denoise: Image prediction and action generation are performed simultaneously, denoising together for N steps. (c) Single-Step Denoise: Action generation is conditioned directly on the context from the first denoising step of the image prediction. Further implementation details, including the construction of the concatenated sequence and the attention mask are provided in Appendix VI-B.



Fig. 8. Demos videos of BagelVLA on Basic Tasks.

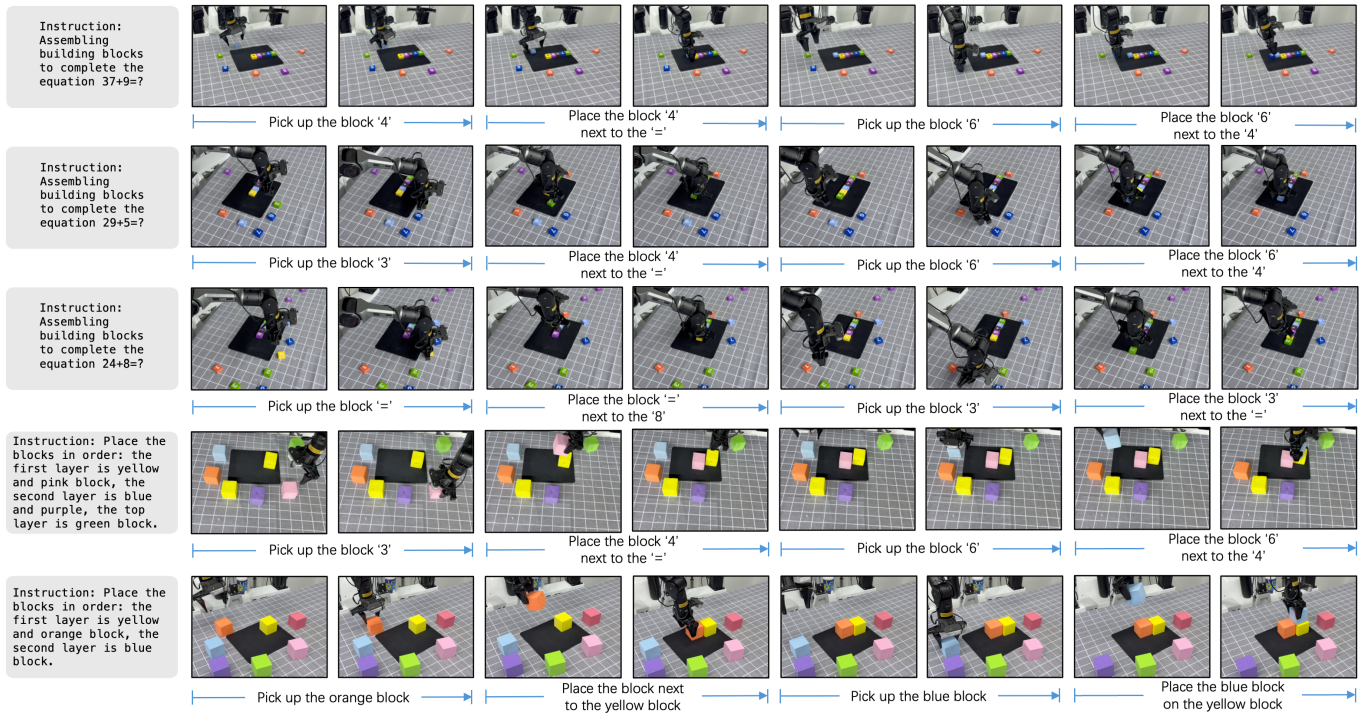


Fig. 9. Demos videos of BagelVLA on Long-Horizon Planning Tasks.



Fig. 10. **Visualizations of interleaved planning results on diverse robotic tasks.** Given a global instruction and the current observation, BagelVLA leverages the context to identify the immediate subtask, predicts a goal image for that subtask.

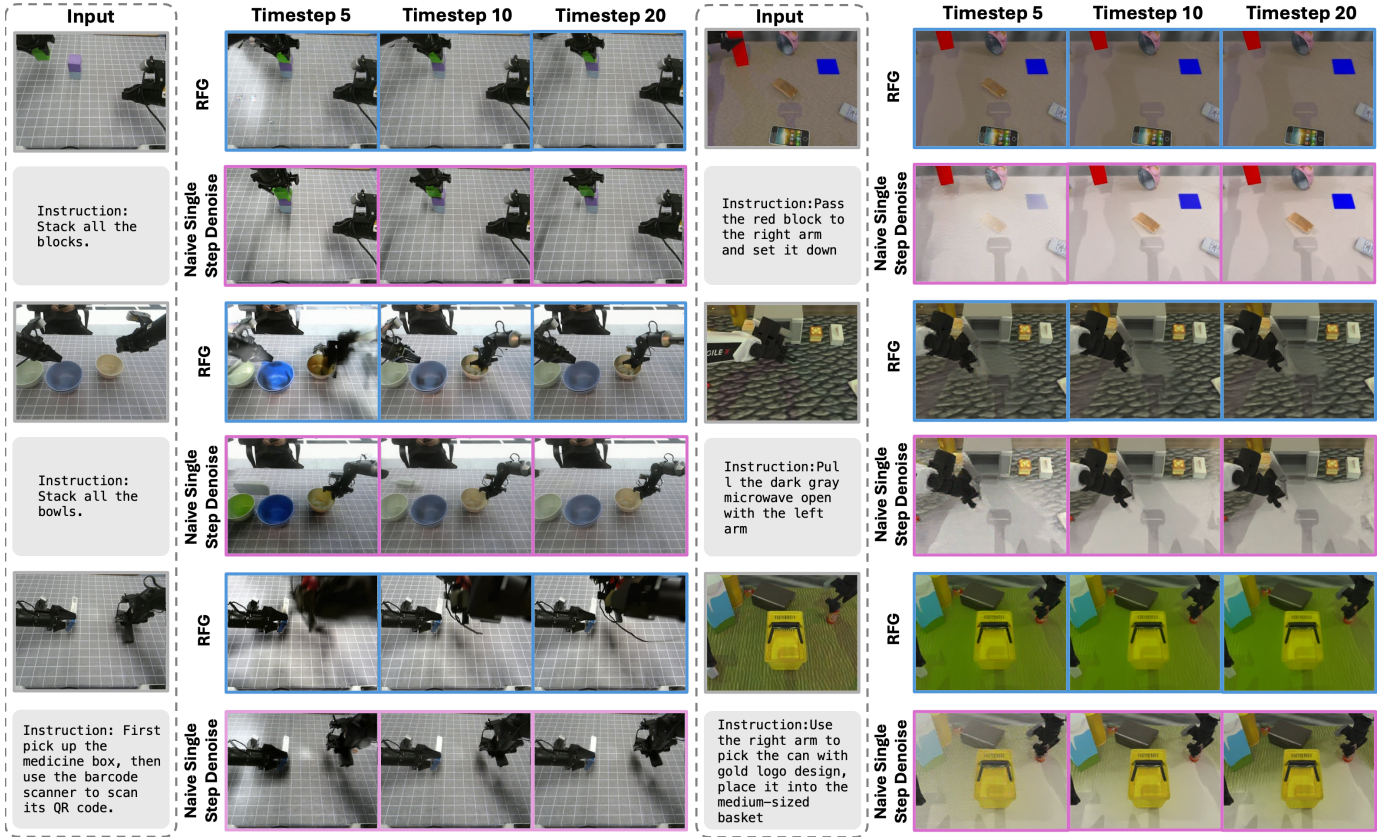


Fig. 11. **Predicted images using different denoising steps.** The figure displays the generation results for the naive single-step denoise (Eq. 2) and RFG (Eq. 3) across varying denoising steps in real-world tasks and simulation scenarios. RFG demonstrates the capability to preserve backgrounds and achieve high-quality generation with very few steps. This provides strong support for reducing the inference latency of interleaved generation.

Subtask prompt

Prompt Example 1

**** System Prompt ****

You are an expert in video analysis and robotic task understanding.

**** Task Description ****

You will analyze an image sequence of a robotic arm performing a specific task with an overall task description. Your task is to make the overall task description more detailed, with the help of the video clip, extract the necessary steps to complete the task, and specify the frame range for each step.

**** Target ****

1.Step Extraction: Extract the ****key steps**** required to complete the overall task, ensuring that each step is clearly described and logically ordered. Each step should include: 2.Specific actions (e.g., tightening screws, stirring mixtures, pressing buttons, etc.) which decomposed from video and overall description 3.Frame window: Specify the start and end frame for each step

**** Output Format ****

Return your output in the following JSON structure:

1.task summary: A ****string**** summarizing the primary task in the video, without mentioning the subject — the robotic arm.

2.steps: An ****array**** where each element includes:

- 'step description': A concise ****verb phrase**** describing the action performed, without mentioning the robotic arm.

- 'start frame': An integer between '0' and 'frame_num' indicating the start frame.

- 'end frame': An integer between '0' and 'frame_num' indicating the end frame.

**** Example Input ****

- 'task description': "Moving colored blocks into a container."

- 'video': ****an image sequence with 'n'****

**** Example Output ****

```
{ "tasksummary": "Moving the red and yellow blocks into a container.",  
  "steps": [ "stepdescription": "pick the red block.", "startframe": 0,  
             "endframe": 6, "stepdescription": "place the red block into container.",  
             "startframe": 7, "endframe": 12, "stepdescription": "pick the yellow  
block.", "startframe": 13, "endframe": 15, "stepdescription": "place  
the yellow block into container.", "startframe": 16, "endframe": frame-  
num-1, ] }
```

Fig. 12. Prompt for Seed-1.5-VL-thinking.

Task description prompt

Prompt Example 1

**** System Prompt ****

You are an expert in video analysis and robotic task understanding.

**** Task Description ****

You will analyze a video of a robotic arm performing a specific task.
Your task is to summary the overall task description.

**** Requirements ****

1. The summary description needs to be as brief as possible but be more detailed than task annotation.
2. Critical task in the video need to be included in this overall task description.
3. The output overall task description must not contradict the simple task annotation.

**** Output Format ****

Return your output by using a simple sentence.

**** Example Input ****

A video clip or image sequence.

**** Example Output ****

Pick up and place the red,blue block, then move the cucumber into the plastic bag

Fig. 13. Prompt for Seed-1.5-VL-thinking.