RoboVIP: Multi-View Video Generation with Visual Identity Prompting Augments Robot Manipulation

Anonymous CVPR submission

Paper ID 9993

Abstract

The diversity, quantity, and quality of manipulation data are critical for training effective robot policies. However, due to hardware and setup constraints, collecting large-scale real-world manipulation data remains difficult to scale across diverse environments. Recent work uses text-prompt conditioned image diffusion models to augment manipulation data by altering the backgrounds and tabletop objects in the visual observations. However, these approaches often overlook the practical need for multi-view and temporally coherent observations required by state-of-the-art policy models. Further, text prompts alone cannot reliably specify the scene setup. To provide the diffusion model with explicit visual guidance, we introduce visual identity prompting, which supplies exemplar images as conditioning inputs to guide the generation of the desired scene setup. To this end, we also build a scalable pipeline to curate a visual identity pool from large robotics datasets. Using our augmented manipulation data to train downstream vision-language-action and visuomotor policy models yields consistent performance gains in both simulation and real-robot settings.

1. Introduction

High-quality and diverse visual data remains fundamental to progress in robotic manipulation and policy learning [7, 24, 38]. However, collecting such data in the real world is notoriously challenging: each episode requires precise mechanical setups, calibrated camera rigs, and reliable synchronization across sensing devices. These constraints make it difficult to scale manipulation datasets in both quantity and environmental diversity. As a complementary solution, recent work has turned to generative models [17, 37] to synthesize additional data, offering a promising alternative to labor-intensive data collection [10, 48, 49].

A growing line of work [10, 48, 49] augments visual observations in manipulation data while keeping the underlying action trajectory fixed. They segment the robot arm and the interacted objects, then apply text prompt-guided image generative models [35] to **inpaint** masked regions, diversifying backgrounds and table-top contents. However, these approaches typically operate in a single-frame, single-view setting, which diverges from the needs of modern policy models [7, 11, 24, 38]. First, many manipulation tasks inherently require reasoning over longer temporal histories, rather than relying solely on a single observation. Consider a policy model executing a *push-button* task where the pre-interaction and post-interaction states of the button appear visually identical. With only one historical frame, the policy model cannot tell whether it has already pushed the button or is about to, often producing indecisive behaviors, even action loops. Second, multi-view observations are increasingly adopted in visuomotor policy models [11, 52] and VLA systems [7, 24, 38], and are usually provided in recent robotics datasets [13, 23], as they provide richer spatial cues and better cross-view generalization. As a result, a practical augmentation framework should operate at the ***video level*** and support ***multi-view*** generation.

In this work, we present a multi-view video generation augmentation framework—including dynamically moving wrist-camera views—that enables diversifying backgrounds and tabletop scenes in a fully **plug-and-play** manner, requiring only raw videos as input. This framework necessitates an automatic segmentation pipeline to segment out both the robot and the objects being interacted with. In practice, the object of interaction may be invisible in early wrist-camera frames; coupled with rapid camera motion, narrow field of view, long trajectories, and limited robotics-specific training, directly applying off-the-shelf models like vision-language models [4, 6] often fail to localize the target object reliably. To address these issues, we propose an automated segmentation pipeline that leverages action information to mitigate segmentation failures from the off-the-shelf model, especially on the wrist-camera. Specifically, we use the 1D gripper state to identify the time window in which the robot actually interacts with the target object, which narrows the search space for the off-the-shelf model.

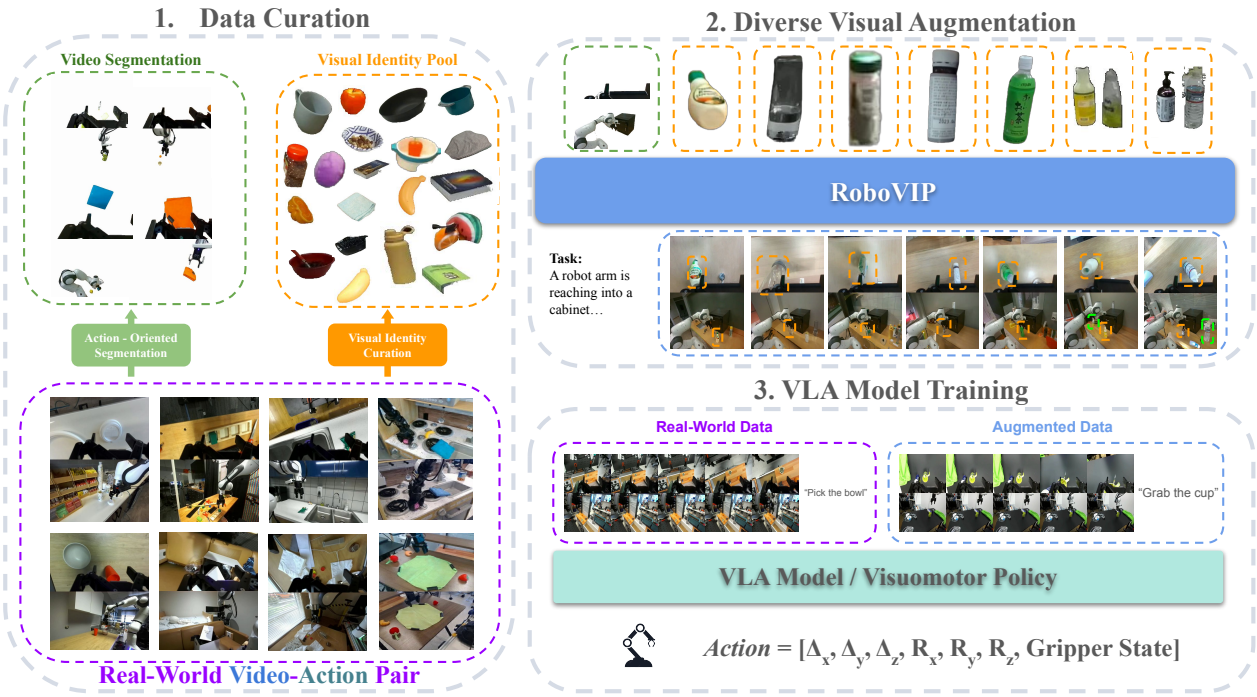

Figure 1. **Overview of our RoboVIP Workflow.** (1) We capture observation videos with corresponding actions to build segmented sequences for inpainting-based augmentation. (2) A large-scale pool of visual identity prompts is curated from robotics datasets and used as conditioning inputs for our generative model RoboVIP to synthesize diverse setups. (3) The augmented videos, paired with maintained action data, are finally utilized for downstream VLA and visuomotor policy training.

Furthermore, we observe that solely relying on text-prompt–guided generation, as done in prior work [2, 3, 10, 48, 49], imposes limitations. The text prompts provided by existing datasets [13, 23] are typically overly simplistic and lack detailed table-top descriptions. Even if we apply SoTA vision-language-model [6] to caption, the generated descriptions frequently suffer from hallucinations and misalignment. More importantly, text prompts can not capture low-level details. To mitigate these issues, we introduce ***visual identity prompting***, conditioning the video diffusion model on one or more exemplar images to synthesize both semantically and low-level consistent content in the inpainted regions. This conditioning will force the model to enrich the table-top and background contents. In parallel, we further modify our video diffusion model to incorporate visual identity prompting conditions in the multi-view paradigm and propose an efficient scheme to embrace multiple identities at the same time. To preserve the plug-and-play nature of our framework, visual identities are *not manually* provided by humans, as in general video generation methods [15, 30]. Instead, we propose an agentic curation and filtering pipeline that automatically constructs a million-scale visual identity pool from large-scale robotics datasets.

In summary, we present ***RoboVIP***, a multi-view inpainting video diffusion model with visual identity prompting to augment the visual observations of the robotics manipulation data. Our approach integrates: (i) an action-oriented pipeline for multi-view robot and object segmentation, (ii) a plug-and-play visual identity pool constructed by a scalable curation process, and (iii) a video diffusion model capable of generating temporally consistent multi-view sequences with visual identity conditions. To assess our effectiveness at scale, we augment 12K BridgeV2 [41] trajectories for training mainstream VLA models, including π_0 [7] and Octo [38]. We further evaluate RoboVIP on a real-world robot dataset (100 trajectories) for training a visuomotor policy, Diffusion Policy [11]. Across both simulation and real-robot evaluations, RoboVIP delivers consistent gains in success rate, demonstrating its practicality for large-scale VLA training as well as low-data policy learning.

2. Related Works

2.1. Conditioned Video Generation

Video generation synthesizes realistic, temporally coherent sequences conditioned on text, images, or videos [2, 26, 42, 47]. Video-to-video models transform an input clip into another consistent sequence, enabling style transfer, enhancement, and content editing [22, 26, 55]. Beyond pixel-aligned cues, identity references [15, 30, 43] have emerged as a way to inject explicit visual attributes into generation. Although video generation is increasingly used for robot planning [12, 19, 28] and controllable video generation is

gaining traction [44, 56], its use for visual augmentation remains underexplored: most existing approaches rely on image-based diffusion [48, 49] or support only single-view conditioning [2], leaving multi-view and richer conditioning largely unexamined.

2.2. Visual Augmentation on Robotics

Traditional augmentations like cropping, rotation, and resizing offer limited benefit for robot policy training and do not resolve data scarcity. As a result, practitioners often resort to manual dataset edits to increase visual diversity. GreenAug [39] augments the dataset by manually setting the real-world robot manipulation environment with the green screen and generating the background by post-effects. ReBot [14] and RoboSplat [46] conduct a real-to-sim-to-real paradigm, which converts the real-world robot information to a simulation environment and then manually adds objects, changes views, or poses of the objects to create a hand-crafted dataset. However, these methods require significant manual effort and do not generalize to new tasks or environments in a plug-and-play manner. To achieve plug-and-play, Cosmos-Transfer [2, 3] and RoboTransfer [31] estimate pixel-aligned dense cues—edges, depth, segmentation—to drive appearance synthesis from text prompts. Although effective, this design limits the ability to introduce new semantic content. Instead, Rosie [48] and RoboEngine [49] apply masking to the background and tabletop contents and then apply a generative model to inpaint masked areas, which unleash the pixel-aligned constraints on the generated contents. Beyond these works, we introduce accurate multi-view masks and enable finer controllability over the masked regions by visual identity prompting.

2.3. Manipulation Models

Research on robot manipulation has progressed from early visuomotor policies to unified VLA architectures. Classic visuomotor systems [11, 27] map images to actions with supervised or reinforcement learning; however, task-specific demonstrations limit generalization. Large vision-language models [1, 33] broadened representational capacity, which in turn motivated VLAs that jointly encode vision, language, and action. Two design axes are now prominent: first, temporal conditioning, where models range from short history windows (RT-1 [8], Octo[38], OpenVLA[38]) to sequence encoders; second, viewpoint, where many systems use a single egocentric stream, whereas newer ones incorporate multi-view inputs to improve 3D reasoning (e.g., π_0[7]). As a result, training increasingly requires temporally coherent and cross-view-aligned data, yet such data remain scarce because real-world collection is slow, costly, and rarely long-horizon or synchronized [9]. This gap motivates data augmentation that generates additional supervision while preserving frame-to-frame and view-to-view consistency.

3. Method

3.1. Problem Formulation

As shown in Fig. 1, starting from a robotic manipulation episode with multi-view video sequences and its corresponding action information on the end-effector state, we segment the robot arm and the objects being interacted with to preserve the fundamental 6-DoF Cartesian end-effector delta pose and gripper-state information. For the remaining masked regions—including background, foreground, and tabletop objects—we adopt a video diffusion model that is conditioned on the text prompt $\mathbf{y}$, the masked multi-view videos $\mathbf{M} = \{M_0, ..., M_N\}$, and the proposed visual identity prompting $\mathbf{f} = \{\mathbf{f_1}, ..., \mathbf{f_k}\}$. The model is trained to learn the conditional joint distribution $p_\theta(I_0, ..., I_N \mid M_0, ..., M_N, y, f_1, ..., f_k)$, where $\mathbf{I} = \{I_0, ..., I_N\}$ denotes the generated latent video frames that adhere to all conditioning signals. The inpainted multi-view videos serve as augmented observations for policy model training, while the action sequences are directly reused from the original episodes. Meanwhile, original episodes are mixed into the training set.

3.2. Action-guided Segmentation of Robots and Interacted Objects

The action information generally includes the 6-DoF pose of the end-effector Δx, Δy, Δz, $\Delta roll$, $\Delta pitch$, Δyaw, and a 1-D gripper state. The gripper state indicates when the robot arm closes or opens, which provides decisive cues for most robot manipulation tasks. In a long video sequence, an effective grasp by the robot arm occurs only within a very short time window and the moment when the gripper state changes can be used to localize the interacted object from the wrist-view perspective. Specifically, we first identify the frame range corresponding to gripper-closure intervals in the wrist view, which marks the preparation and execution of interaction. The extracted video clip is then fed into a video-reasoning VLM [4] to infer the object's semantic label, enabling accurate object naming directly from the wrist view.

When processing other third-person camera views, the object name obtained from the wrist view is directly reused. The identified object name is fed into an open-vocabulary segmentation model [49, 51] to obtain a reliable mask for the corresponding frame. We separately extract the masks for the robot and the interacted object, followed by median blurring to filter out outlier pixels. At this stage, we obtain accurate frame indices and mask locations. To further refine temporal consistency, we perform k-means sampling on the masks, and the sampled points will be taken as prompting for the video segmentation model [34] to track the complete video segmentation for the robot and interacted objects. The robot and the mask of the object being inter-

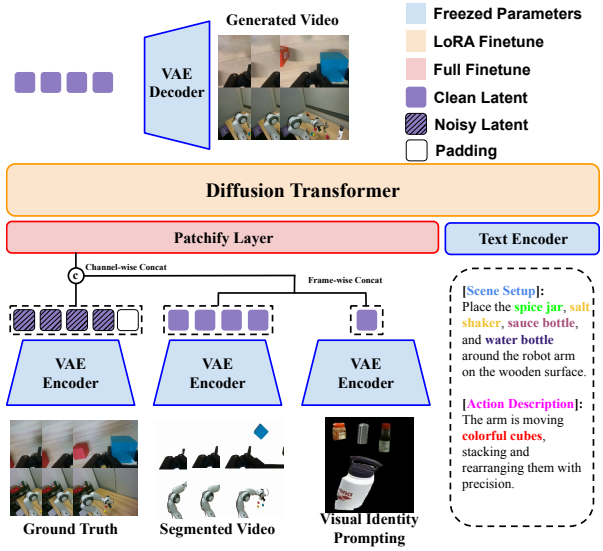

Figure 2. **Video Diffusion Model Architecture on Training.** Our video diffusion model is conditioned on the segmented video sequence, structured text prompt, and proposed visual identity prompting.

acted with are processed independently and then merged into one at the end. The resulting video segmentation provides high-quality mask conditions that can be directly used in the training process.

3.3. Multi-View Inpainting Video Diffusion Model

We aim to transfer the generation quality and condition-alignment capability of state-of-the-art video diffusion models to robotic tasks. To this end, our base model is the Wan2.1 [42] image-to-video variant with 14 billion parameters. However, directly fine-tuning such a large model is computationally infeasible. More critically, it leads to severe overfitting collapse, causing the model to rapidly forget its original visual generation stability. To address this, we adopt a Low-Rank Adaptation [18] (LoRA) strategy to enable feasible and memory-efficient fine-tuning.

Recent video diffusion models are predominantly built upon the Diffusion Transformer [32] architecture, where attention blocks serve as the primary computational units. LoRA injects trainable low-rank adapters into the linear projections, typically applied to the query and value matrices within attention layers. Apart from the attention blocks, the patchification layer—implemented as a convolutional layer that transforms latent images into patches—is fully trainable but not part of LoRA fine-tuning. Since our training objective shifts from single-image conditioning in the base model to the masked video sequences as conditions, we enable the patchification encoder for training as well. Empirically, we find that fine-tuning this additional layer beyond the LoRA setup tends to yield slight improvements in performance.

For multi-view inputs, inspired by [21], we adopt a structured vertical stitching strategy that concatenates masked frames from different views at the same timestamp. The ground-truth sequences are processed in the same manner, ensuring that the learning objective remains view-aligned and encourages the video diffusion model to capture cross-view spatial consistency and correspondence. Accordingly, we modify the base model's input structure by replacing the original single-image padding with channel-wise concatenation of the full video sequence—to achieve a minimally invasive yet effective formulation of a video-conditioned objective. The overall model structure can be found in Fig. 2. The visual identity prompting part is in Sec. 3.4.

3.4. Visual Identity Prompting

For robotic downstream tasks, we aim for a pipeline that can autonomously select appropriate and necessary visual identities without any human intervention. To achieve this, we design an agentic inference pipeline, as shown in Fig. 4, that automatically constructs a massive, rich, and diverse visual identity pool. We find that adopting a panoptic segmentation [25] approach is the most straightforward way to achieve this goal. Panoptic segmentation simultaneously provides mask localization and corresponding label classification. Based on the classification label, we select common objects that are needed and do not consider background-related large objects, like the table and the wall. Using these labels, we can naturally classify both tabletop objects and background elements, ultimately forming a comprehensive visual identity pool.

We observe that objects obtained by straightforward segmentation are often of suboptimal quality. Moreover, some of these segmented objects are partially occluded and thus cannot serve as semantically complete visual identity references. To address this, we crop the corresponding visual identity image predicted by the panoptic segmentation model [20] and then apply several filtering criteria, including image quality assessment [45], sharpness clarity assessment, CLIP-based text–image scoring [33], and resolution size filtering. The CLIP text embedding is derived from the panoptic segmentation class label, serving as an effective proxy to assess the semantic completeness of each object. The filtered visual identities constitute a high-quality million-scale visual identity pool.

Unlike previous approaches that inject only a single identity reference per frame, we adopt a packing scheme to efficiently accommodate multiple visual identity references within a single frame, thereby reducing computational overhead. To prevent overfitting to fixed scale ratios, each identity image is randomly resized before encoding. During training under multi-view supervision, all visual identity references are sampled from a single view to avoid ambiguity in identity prompting.

To incorporate visual identity prompting into the video

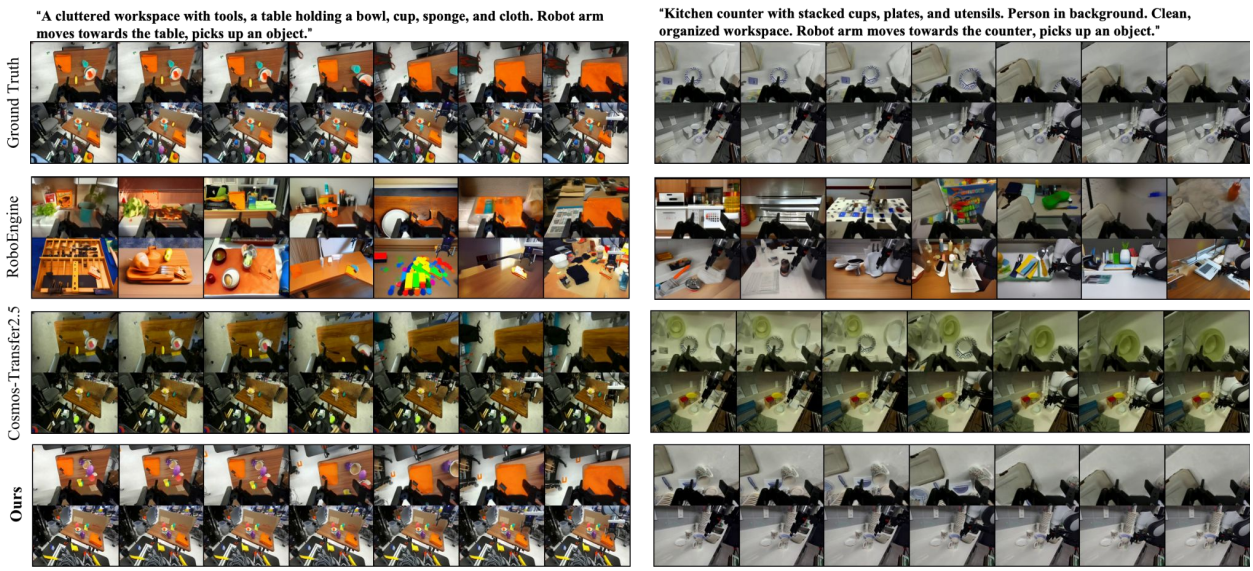

Figure 3. Qualitative comparison of generative augmentation for robot manipulation. Our method produces temporally consistent and visually diverse results, outperforming RoboEngine [49], which is a single-image-based method, and Cosmos-Transfer2.5 [3], which struggles to generalize beyond appearance-level edge conditioning. **Zoom in** for the best view.

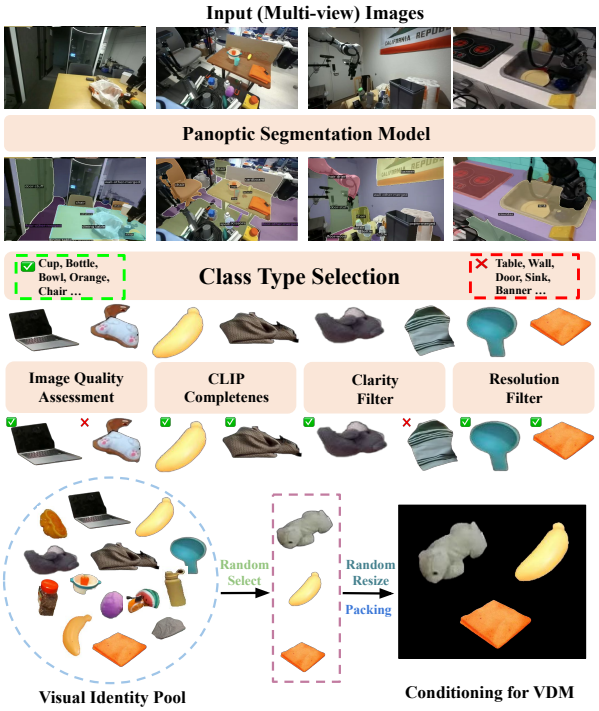

Figure 4. **Visual Identity Curation and Processing Pipeline.** Our visual identity is curated by panoptic segmentation from the large-scale robotics dataset [13, 23, 41], followed by several scoring criteria filters. In augmentation, we randomly select some from the pool and pack them into one image frame to serve as conditioning for our video diffusion model.

diffusion model, we adopt a frame-wise concatenation strategy, following the design of [43, 54]. As illustrated in Fig. 2, before entering the video diffusion transformer, the packed identity images are first encoded by a shared causal VAE encoder [42] and concatenated with the latent video segmentation inputs along the frame dimension. The noisy frame latent is zero-padded for temporal alignment and then channel-wise concatenated with the conditional inputs. After the diffusion transformer processes all layers, the identity tokens are dropped and excluded from loss computation, ensuring they serve purely as contextual guidance rather than optimization targets. During inference, newly encoded identity images are injected at each diffusion timestep to continuously guide generation.

4. Experiment

4.1. Video Diffusion Model Implementation Details

Data Curation. For all videos, we first discard sequences that are too short (fewer than 25 frames). For overly long sequences (more than 550 frames), we perform temporal cropping to mitigate segmentation failures induced by excessively extended inputs. Since the original dataset captions are often noisy, we re-caption all videos using Qwen2.5-VL 32B [6], employing a multi-view vertical stitching strategy to ensure consistent and accurate textual descriptions across different viewpoints. The text prompt is composed of the scene setup and the action description. Then, for the robot and object segmentation described in the method section, we adopt OneFormer [20] for the panoptic segmentation. The open-vocab segmentation model is the EVF-SAM [51] model from RoboEngine [49]. The video segmentation is by SAM2 [34].

Training Details. We train our video diffusion models on

CVPR #9993

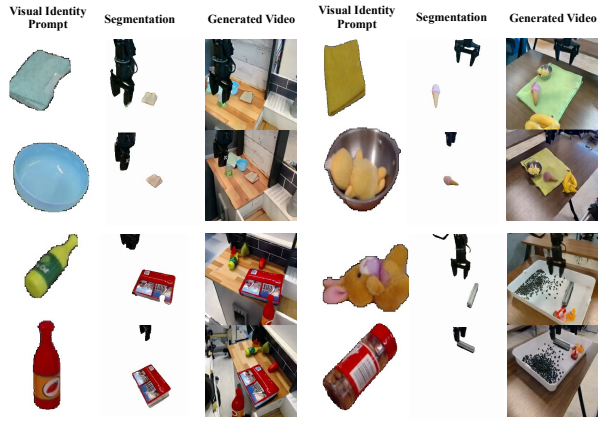

Figure 5. **Augmented BridgeV2 Data by our RoboVIP for VLA Training**. Our visual identity prompting enriches tabletop contents and introduces additional distractors to create more challenging settings for the policy model. The visual identity is randomly selected from our proposed pools. **Zoom in** for the best view.

Bridge V1 [13] and V2 [41] for the following downstream VLA tasks, which provide one to three third-person views. Further, we train Droid [23] for the visual quality comparisons and real-robot augmentation, which includes a wrist-mounted camera and two third-person views. To support variable-length sequences, we adopt a batch size of 1 per GPU and use gradient accumulation to achieve an effective batch size of 4 per GPU, which costs around 70GB per GPU in training. This strategy enables dynamic frame sampling without incurring unnecessary computation from padding or attention masking overhead. Since current VLA models cannot condition on long observation histories, we train on at most 49 frames. We train for 15K iterations on 8 GPUs whose per-GPU memory is 144GB, resulting in a total cumulative batch size of 32. Each view is trained at 256×256 resolution for Bridge and 320×416 for Droid. When an instance contains only a single view, the conditioning input for the missing view is zero-padded. To avoid ambiguity between padding and segmentation masks, we represent masked regions using **white** pixels rather than zeros.

In practice, we observe that generative quality correlates with the model's pretrained resolution; therefore, the cumulative stitched width and height for multi-view must be lower than the pretrained setting. Guided by this finding, we employ the Wan2.1-I2V [42] 720p variant to support diverse generation settings, rather than the lower-resolution 480p model. We set the LoRA [18] rank to 128 and 256 for the Bridge and Droid configurations, respectively. To maximize data utilization, we randomly sample two views from the three available in Bridge V2. For Droid, we fix the wrist-mounted camera as the first view and select the second view from the two third-person perspectives.

4.2. Video Generation Results

Our goal is to develop a scalable, plug-and-play multi-view inpainting generative framework that serves as an effective augmentation solution for robotic manipulation. To this end, we evaluate our method against Cosmos-Transfer2.5 [3], a video diffusion model designed for real-to-real generation, and RoboEngine [49], an inpainting-based approach similar to ours, on the held-out test subset of the Droid [23] dataset consisting of 300 test cases. We consider both a wrist-mounted view and a third-person view for augmentation. All models are conditioned on the same Qwen2.5-VL [6] captioned text prompts. For RoboEngine, we use identical robot and object segmentation masks as our method for a fair comparison, since inaccurate masks—e.g., those covering substantially larger regions than the true object—can artificially enlarge the ground-truth area and spuriously improve distribution-based metrics like FID [16] (detailed in supplementary). For Cosmos-Transfer2.5, we evaluate its edge-conditioned variant, using its native 720p setup. For our RoboVIP, we apply visual identity prompting. For all methods, we generate at most 49 frames per episode, starting from the first frame.

We report standard generative video metrics: Fréchet Inception Distance [16] (FID), Fréchet Video Distance [40] (FVD), and Learned Perceptual Image Patch Similarity [50] (LPIPS). FID measures single-frame visual quality via distributional differences, while FVD captures temporal coherence and video-level dynamics. LPIPS quantifies image-level perceptual similarity between generated outputs and ground truth in deep feature space rather than pixel space. These metrics measure on vertically stitched inputs due to the multi-view setting. To reflect the multi-view nature of our setting, we follow prior works [5, 31] and evaluate cross-view correspondence by counting matched feature points between two generated views (MV-Mat.). A higher count indicates better spatial consistency and generative stability. We use GIM [36] as the correspondence model, keeping all confidence thresholds and hyperparameters identical to its demo configuration.

As shown in Tab. 2, our method consistently outperforms prior approaches on most quantitative metrics. The improvement can be attributed to the fact that RoboEngine operates under a single-frame, single-view setting, while Cosmos-Transfer2.5 overlooks the requirements of multi-view generation. In supplementary, we will also include a human study for the visual identity prompting-oriented comparisons. As shown in Fig. 3, compared to RoboEngine, our RoboVIP performs distinguished temporal consistency. Compared to Cosmos-Transfer2.5, ours unleashes diverse scene generation, which is not limited by the pixel-aligned conditions, like edges or depth. This is thanks to our inpainting design choices. Further, none of the methods achieve multi-view consistent generation.

CVPR #9993

Table 1. Comparison of evaluation results on the WidowX robot in SIMPLERENV. We evaluate two variants of our RoboVIP: a text-prompt–conditioned multi-view inpainting video diffusion model, and another version with additional visual identity prompting conditions (denoted as ID). Each task is performed on 100 trials. Each entry shows *Grasp/Put*, where Put is the conditional success rate given a successful Grasp (Put = Success/Grasp), and the *Success* column reports overall task success. **Bold** and underlined numbers in the Average Success column indicate the best and second-best performance.

Model	Put spoon on towel		Put carrot on plate		Stack green cube on yellow cube		Put eggplant in basket		Average	
	Grasp/Put	Success	Grasp/Put	Success	Grasp/Put	Success	Grasp/Put	Success	Grasp/Put	Success
Octo [38] (Zero-Shot)	34% / 26%	9%	35% / 20%	7%	28% / 0%	0%	65% / 51%	33%	40.5% / 30.1%	12.2%
Octo (Bridge V2 SFT)	52% / 41%	29%	32% / 37%	14%	47% / 4%	3%	60% / 11%	5%	47.5% / 23.0%	12.8%
Octo+RoboEngine [49]	67% / 21%	14%	43% / 37%	16%	43% / 5%	2%	0% / 0%	0%	38.2% / 20.9%	8.0%
Octo+**RoboVIP** (Text prompt)	59% / 7%	4%	69% / 46%	32%	55% / 9%	5%	63% / 17%	11%	**61.5%** / 21.1%	13.0%
Octo+**RoboVIP** (Text prompt with ID)	59% / 63%	37%	37% / 62%	23%	47% / 15%	7%	37% / 19%	7%	45.0% / **41.1%**	**18.5%**
π_0 [7] (Zero-Shot)	52% / 63%	33%	0% / 0%	0%	28% / 7%	2%	31% / 42%	13%	27.75% / 43.2%	12%
π_0 (Bridge V2 SFT)	57% / 63%	36%	44% / 43%	19%	30% / 7%	2%	29% / 41%	12%	40% / 43.1%	17.25%
π_0+RoboEngine [49]	61% / 70%	43%	33% / 30%	10%	61% / 11%	7%	31% / 45%	14%	46.5% / 39.8%	18.5%
π_0+**RoboVIP** (Text prompt)	74% / 84%	62%	52% / 40%	21%	49% / 14%	7%	36% / 64%	23%	52.75% / **55.0%**	**29%**
π_0+**RoboVIP** (Text prompt with ID)	73% / 64%	47%	49% / 41%	20%	52% / 13%	7%	53% / 70%	37%	**56.75%** / 48.9%	27.75%

Table 2. **Generative Model Comparisons** on 300 test cases of Droid [23]. Cosmos refers to Cosmos-Transfer2.5 [3]. The best is **highlighted**.

Method	FID↓	FVD↓	LPIPS↓	MV-Mat.↑
Cosmos [3]	47.43	325.4	**0.353**	1583.4
RoboEngine [49]	62.77	1788.8	0.598	1301.9
RoboVIP (Ours)	**39.97**	**138.4**	0.409	**2242.1**

4.3. Simulation Results

To evaluate performance in a reproducible and scalable way, we employ the simulation environment suite SimplerEnv [29], which shows realistic textures in the simulation environment like the real-world and has been shown to correlate well with real-world robot manipulation performance [53]. SimplerEnv enables consistent benchmarking of generalist manipulation policies under common robot setups.

We evaluate our pipelines on two recent vision-language-action models: Octo-base [38] and π_0 [7]. Specifically, we adopt the Octo-base model as a **multi-frame conditioned** policy (with 2 history frames) and the π_0 model as a **single-frame** VLA policy. We fine-tune both Octo and π_0 on 8 NVIDIA GPUs with 48GB of memory each. All experiments use a global seed of 42 and identical data processing and augmentation settings following their official preprocessing pipeline.Both Octo and π_0 are evaluated under three training regimes:

- **Zero-shot**: both Octo-base and π_0 are directly deployed without any further supervised fine-tuning in the SimplerEnv tasks.
- **Supervised Fine-Tuning (SFT) on BridgeDataV2**: we fine-tune each model using the open-ended instruction-conditioned dataset BridgeDataV2 [41] and then deploy.
- **Mixed-policy baseline vs. our method**: We mix BridgeDataV2 with the augmented data produced by **RoboEngine** [49] and our proposed **RoboVIP**. We evaluate two variants of our multi-view inpainting video diffusion model. The first variant uses only text prompts as the generative condition. The second variant augments the same architecture with our visual identity prompting, which provides exemplar images as additional conditioning signals as shown in Fig. 5.

Tab. 1 presents quantitative comparisons across the SimplerEnv tasks. For the Octo family, our RoboVIP (Text+ID) achieves an average success rate of 18.5%, improving upon Octo zero-shot experiment (12.2%) and Bridge SFT version (12.8%), and our text-prompt-only variant (13.0%). For π_0, our RoboVIP (Text-only) configuration yields the highest overall success at 29.0%, outperforming both the SFT baseline (17.25%) and RoboEngine (18.5%). The Text+ID variant performs similarly at 27.75%, confirming that both visual identity prompting and temporally consistent generative augmentation contribute to stronger policy generalization.

A closer look at the decomposition into *Grasp* and conditional *Put* success (*Put = Success / Grasp*) reveals the source of these gains. For Octo, our RoboVIP (Text+ID) attains the best average *Put* success at 41.1%, significantly higher than the 23.0% achieved by Octo SFT. In π_0, the text-only RoboVIP obtains the highest *Put* success of 55.0%, exceeding the SFT baseline (43.1%) and RoboEngine (39.8%). These results indicate that our method not only improves task initiation (grasping) but also strengthens the more challenging post-grasp "Put" phase, demonstrating enhanced closed-loop control and task completion reliability.

The observed improvements arise from RoboVIP's abil-

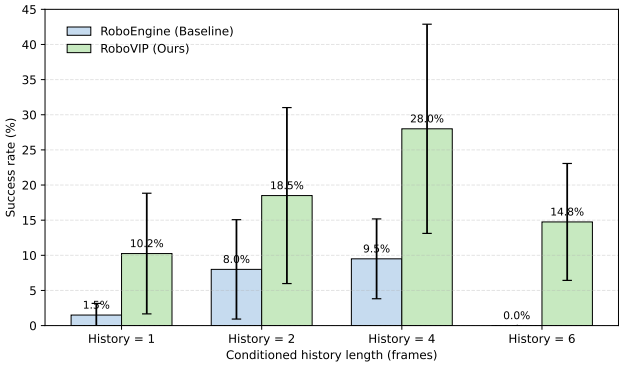

Figure 6. Policy success rate vs. conditioned history length (frames). Bars show task-averaged success rates for RoboEngine (baseline) and our RoboVIP (Text prompt with ID) on Octo [38]. The error bars denote standard deviation across tasks. The bar plot indicates that RoboVIP consistently outperforms RoboEngine, whose average success falls to zero at six history frames.

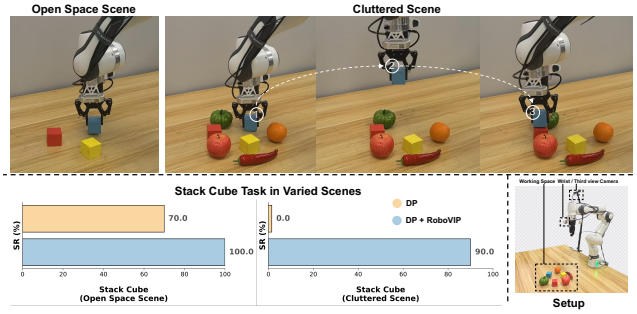

Figure 7. **Real Robot Experiment on Diffusion Policy.** Both policies were trained using identical parameter settings and measured over 10 trials.

ity to generate temporally consistent, multi-view scenes that closely approximate real data distributions. For models such as Octo, which condition on multiple frames, our generated sequences provide realistic motion continuity that mitigates frame inconsistency issues present in RoboEngine. For π_0, the multi-view setup aligns with its pretraining configuration, reducing the gap between synthetic and real data distributions. Moreover, the use of visual identity prompts enriches scene diversity (as shown in Fig. 5) and introduces controlled clutter, which empirically benefits learning in visually complex settings. As a result, generative data from RoboVIP can closely approach—or even surpass—the effectiveness of real fine-tuning data.

Further, we compare the influence of history length on VLA success in Fig. 6. We retrained and tested Octo on different numbers of history observation frames. Across all historical lengths, our RoboVIP maintains consistently higher success rates than RoboEngine (baseline). Notably, while RoboEngine's performance collapses to nearly zero under six-frame conditioning, our RoboVIP still preserves meaningful success rates, underscoring its robustness to longer temporal contexts. This trend suggests that video-level generative augmentation—not image diffusion—is a more scalable and forward-looking direction, which is applicable for future long-horizon needs on VLA training.

4.4. Real-World Robot Results

To specifically validate the effectiveness of our RoboVIP augmentation pipeline against real-world background distractors, we conduct experiments using a 6-DoF Franka Research 3 robotic arm equipped with a Robotiq gripper. We design a cube stacking task, which requires grasping a blue cube and stacking it onto the red cube. All experiments are conducted using Diffusion Policy (DP) [11]. We established two experimental settings to test robustness against background distractors: **Open space:** A clean background with no distractors. **Cluttered:** A scene with 4 different distractor objects. We compare the performance of two policies:

- **DP:** A DP model trained solely on 100 real-world demonstration trajectories.
- **DP + RoboVIP (Text+ID) :** A DP model trained on a mixed dataset of 200 trajectories, consisting of the 100 original demonstrations and 100 additional trajectories augmented by our RoboVIP framework.

As shown in Figure 7, the baseline DP model's success rate drops from 7/10 in the **Open space** setting to 0/10 in the **Cluttered** setting. In contrast, the **DP + RoboVIP** model achieves perfect 10/10 success in the open space setting and maintains a robust 9/10 success rate in the cluttered setting. This demonstrates that our augmentation pipeline significantly enhances the policy's generalization and robustness to real-world visual distractors. More experiments are presented in the supplementary.

5. Conclusion

In this work, we introduce RoboVIP, a multi-view inpainting video diffusion model with visual identity prompting to augment visual observations of the robotic manipulation data in a plug-and-play manner. We augment large-scale data and demonstrate its effectiveness in both vision-language-action and visuomotor policy models on both the simulation environment and the real-world robot deployment.

Limitation. Although our method can automate large-scale visual data augmentation and we prove its effectiveness on VLA and visuomotor policy learning, several limitations stem from current tools. State-of-the-art video segmentation [34] still struggles with gripper localization and flickering; VLM reasoning [4, 6] often fails to identify interactive objects; and the open-vocabulary segmentation [49, 51] frequently produces incorrect masks and does not produce consistent results in multi-view inputs.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Jin, et al. Flamingo: A visual language model for few-shot learning. In *NeurIPS*, 2022. 3

[2] Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, et al. Cosmos-transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492*, 2025. 2, 3

[3] Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025. 2, 3, 5, 6, 7

[4] Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025. 1, 3, 8

[5] Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. *arXiv preprint arXiv:2412.07760*, 2024. 6

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2, 5, 6, 8

[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. 1, 2, 3, 7

[8] Anthony Brohan, Yevgen Chen, Karol Hausman, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 3

[9] Anthony Brohan, Noah Brown, Daniel Rifai, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023. 3

[10] Zoey Chen, Zhao Mandi, Homanga Bharadhwaj, Mohit Sharma, Shuran Song, Abhishek Gupta, and Vikash Kumar. Semantically controllable augmentations for generalizable robot learning. *The International Journal of Robotics Research*, 44(10-11):1705–1726, 2025. 1, 2

[11] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44 (10-11):1684–1704, 2025. 1, 2, 3, 8

[12] 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. 2

[13] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021. 1, 2, 5, 6

[14] Yu Fang, Yue Yang, Xinghao Zhu, Kaiyuan Zheng, Gedas Bertasius, Daniel Szafir, and Mingyu Ding. Rebot: Scaling robot learning with real-to-sim-to-real robotic video synthesis. *arXiv preprint arXiv:2503.14526*, 2025. 3

[15] Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025. 2

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. 6

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4, 6

[19] 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. 2

[20] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2989–2998, 2023. 4, 5

[21] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through neural trajectories. *arXiv e-prints*, pages arXiv–2505, 2025. 4

[22] Xuan Ju, Tianyu Wang, Yuqian Zhou, He Zhang, Qing Liu, Nanxuan Zhao, Zhifei Zhang, Yijun Li, Yuanhao Cai, Shaoteng Liu, et al. Editverse: Unifying image and video editing and generation with in-context learning. *arXiv preprint arXiv:2509.20360*, 2025. 2

[23] 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. 1, 2, 5, 6, 7

[24] 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. 1

[25] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 4

[26] Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024. 2

[27] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. In *Journal of Machine Learning Research*, pages 1–40, 2016. 3

[28] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025. 2

[29] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024. 7

[30] Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*, 2025. 2

[31] Liu Liu, Xiaofeng Wang, Guosheng Zhao, Keyu Li, Wenkang Qin, Jiaxiong Qiu, Zheng Zhu, Guan Huang, and Zhizhong Su. Robotransfer: Geometry-consistent video diffusion for robotic visual policy transfer. *arXiv preprint arXiv:2505.23171*, 2025. 3, 6

[32] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 4

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3, 4

[34] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3, 5, 8

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[36] Xuelun Shen, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, and Cheng Wang. Gim: Learning generalizable image matcher from internet videos. *arXiv preprint arXiv:2402.11095*, 2024. 6

[37] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[38] 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. 1, 2, 3, 7, 8

[39] Eugene Teoh, Sumit Patidar, Xiao Ma, and Stephen James. Green screen augmentation enables scene generalisation in robotic manipulation. *arXiv preprint arXiv:2407.07868*, 2024. 3

[40] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

[41] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023. 2, 5, 6, 7

[42] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 4, 5, 6

[43] Boyang Wang, Xuweiyi Chen, Matheus Gadelha, and Zezhou Cheng. Frame in-n-out: Unbounded controllable image-to-video generation. *arXiv preprint arXiv:2505.21491*, 2025. 2, 5

[44] Boyang Wang, Nikhil Sridhar, Chao Feng, Mark Van der Merwe, Adam Fishman, Nima Fazeli, and Jeong Joon Park. This&that: Language-gesture controlled video generation for robot planning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12842–12849. IEEE, 2025. 3

[45] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023. 4

[46] 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. *arXiv preprint arXiv:2504.13175*, 2025. 3

[47] 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. 2

[48] Tianhe Yu, Ted Xiao, Austin Stone, Jonathan Tompson, Anthony Brohan, Su Wang, Jaspiar Singh, Clayton Tan, Jodilyn Peralta, Brian Ichter, et al. Scaling robot learning with semantically imagined experience. *arXiv preprint arXiv:2302.11550*, 2023. 1, 2, 3

[49] Chengbo Yuan, Suraj Joshi, Shaoting Zhu, Hang Su, Hang Zhao, and Yang Gao. Roboengine: Plug-and-play robot data augmentation with semantic robot segmentation and background generation. *arXiv preprint arXiv:2503.18738*, 2025. 1, 2, 3, 5, 6, 7, 8

[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pages 586–595, 2018. 6

[51] Yuxuan Zhang, Tianheng Cheng, Lianghui Zhu, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Evf-sam: Early vision-language fusion for text-prompted segment anything model. *arXiv preprint arXiv:2406.20076*, 2024. 3, 5, 8

[52] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023. 1

[53] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*, 2024. 7

[54] Yong Zhong, Zhuoyi Yang, Jiayan Teng, Xiaotao Gu, and Chongxuan Li. Concat-id: Towards universal identity-preserving video synthesis. *arXiv preprint arXiv:2503.14151*, 2025. 5

[55] Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2535–2545, 2024. 2

[56] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: A fine-grained world model for robot manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9834–9844, 2025. 3