2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

ViSA-Flow: Accelerating Robot Skill Learning via Large-Scale Video Semantic Action Flow

Changhe Chen*[1], Quantao Yang*[2], Xiaohao Xu[1], Nima Fazeli[1], and Olov Andersson[2]

***Abstract*— One of the central challenges preventing robots from acquiring complex manipulation skills is the prohibitive cost of collecting large-scale robot demonstrations. In contrast, humans are able to learn efficiently by watching others interact with their environment. To bridge this gap, we introduce *semantic action flow* as a core intermediate representation capturing the essential spatio-temporal manipulator-object interactions, invariant to superficial visual differences. We present ViSA-Flow, a framework that learns this representation self-supervised from unlabeled large-scale video data. First, a generative model is pre-trained on semantic action flows automatically extracted from large-scale human-object interaction video data, learning a robust prior over manipulation structure. Second, this prior is efficiently adapted to a target robot by fine-tuning on a small set of robot demonstrations processed through the same semantic abstraction pipeline. We demonstrate through extensive experiments on the CALVIN benchmark and real-world tasks that ViSA-Flow achieves state-of-the-art performance, particularly in low-data regimes, outperforming prior methods by effectively transferring knowledge from human video observation to robotic execution. Videos are available at `https://visaflow-web.github.io/ViSAFLOW`.**

I. Introduction

Robot imitation learning has achieved remarkable success in enabling robots to acquire complex manipulation skills, ranging from basic object manipulation [1], [2] to intricate assembly procedures [3]. However, the scalability of imitation learning approaches is fundamentally limited by the need for extensive, carefully curated robot datasets that are costly to collect. This has become a critical bottleneck in developing robots capable of performing diverse real-world tasks.

In contrast, humans demonstrate an extraordinary ability to learn new skills by observing others. From in-person observation to videos, humans naturally focus on semantically relevant components. For instance, when learning tennis, we naturally attend to the player's body movements, racquet handling techniques, and ball trajectories, while effectively filtering out irrelevant background information. This selective attention to meaningful elements enables efficient skill acquisition and transfer. The vast repository of publicly available videos on the internet similarly represents an untapped resource for robot learning, offering diverse demonstrations of human skills across countless domains. However, effectively leveraging this resource requires addressing several key challenges, particularly in bridging the gap between human demonstrations in unconstrained videos and robot execution in the real world.

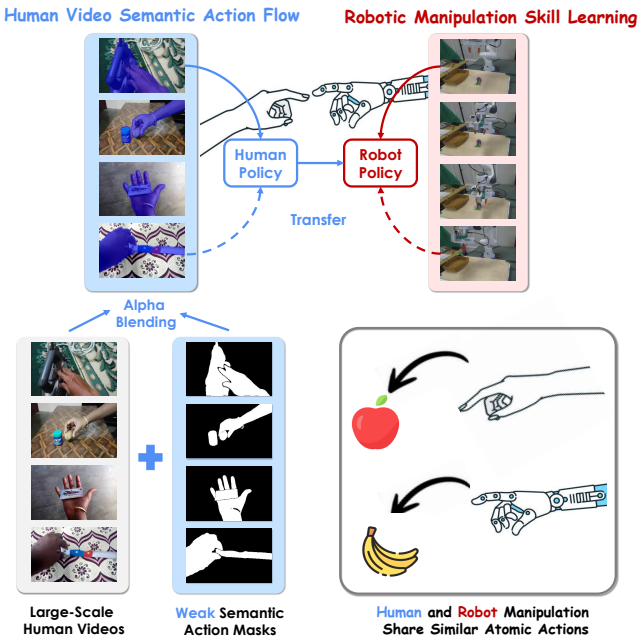

Fig. 1: Humans and robots often share underlying atomic actions for similar tasks. Our framework leverages large-scale, unlabeled human videos by extracting weakly supervised semantic action flow priors (ViSA-Flow). This knowledge is distilled into a human policy and efficiently transferred to learn a corresponding robot policy.

Recent research [4]–[6] has investigated how robots can acquire skills by directly observing unstructured human videos. These methods have shown promising generalization, enabling robots to adapt to new tasks beyond their training data. However, most existing approaches [7]–[9] rely heavily on motion flow as a conditional input for policy learning. While effective in some settings, this low-level representation often overlooks the higher-level semantic cues that humans naturally attend to when learning new skills. In real-world scenarios, when humans acquire a skill, we rarely process the entire visual scene indiscriminately. Instead, we focus selectively on the interaction between the hand (or arm) and the relevant object, while disregarding irrelevant background elements or distractions. Mimicking this selective attention mechanism could make robot learning from videos more

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

*These authors contributed equally.

[1]University of Michigan. (changhec@umich.edu).
[2]Division of Robotics, Perception and Learning (RPL), KTH Royal Institute of Technology, Sweden. (quantao@kth.se).

efficient and robust.

Drawing inspiration from this observation, we propose a novel framework that learns robot skills by extracting temporally consistent semantic action flows from large-scale human manipulation videos (Fig. 1). Unlike prior works that condition policy learning on motion flow, our approach leverages semantic representations—capturing object interactions, body poses, and motion patterns—that can be consistently transferred from human demonstrations to robotic actions. By focusing on these meaningful semantic structures, our method enables more efficient and generalizable skill learning from videos. Our key contributions are threefold:

1) We propose **ViSA-Flow**, a framework for pre-training generative policies using large-scale **Vi**deo **S**emantic **A**ction **Flow**, capturing spatio-temporal manipulator-object interactions from diverse human video demonstrations. This enables efficient knowledge transfer from large-scale human video data to robotic manipulation policies and offers a new perspective on representing action sequences for cross-domain transfer.
2) We refine the pretrained policy using robot-specific semantic actions from few expert demonstrations by tracking hand-object interactions in both human videos and robot data, enabling robust semantic alignment for improved policy adaptation.
3) We evaluate ViSA-Flow in both simulated and real-world robotic manipulation tasks, demonstrating substantial performance improvements over SOTA baselines, particularly in low-data regimes.

II. Related Work

Visual Imitation Learning. Recent advancements [10]–[14] in visual feature-based imitation learning have significantly improved the efficiency, generalization of learning from visual demonstrations. VIEW [15] introduces a trajectory segmentation approach that extracts condensed prior trajectories from demonstrations, allowing robots to learn manipulation tasks more efficiently. Similarly, K-VIL [16] enhances efficiency by extracting sparse, object-centric keypoints from visual demonstrations, reducing redundancy and improving learning speed. Beyond efficiency, generalization remains a critical challenge, particularly in adapting to diverse visual environments. Stem-OB [17] addresses this issue by leveraging diffusion model inversion to suppress low-level visual differences, improving robustness against variations in lighting and texture. In addition, goal-oriented approaches have been developed to improve policy learning and adaptation. Visual hindsight self-imitation learning [18] introduces hindsight goal re-labeling and prototypical goal embedding, enhancing sample efficiency in vision-based tasks.

Video-Based Robot Learning. Recent advancements [19]–[22] in robot learning have demonstrated the effectiveness of large-scale video datasets for pre-training models and improving generalization. Methods such as Time-Contrastive Networks (TCN) [23] have pioneered the extraction of temporally consistent features to align human demonstrations with robot actions. Building on this foundation, video pre-training [24] has shown that large-scale video data can be used to pretrain robust visual representations for downstream manipulation tasks. More recent works [25] have further leveraged large-scale video datasets to enhance manipulation performance. Similarly, Vid2Robot [26] directly translates video demonstrations into robot actions using cross-attention for alignment. [6] highlights the potential of leveraging partially-annotated data to enhance robot policy learning by integrating multi-modal information. [27] transforms human video demonstrations into robot-compatible observation-action pairs by inpainting the human arm, and overlaying a rendered robot to achieve visual domain alignment. Beyond task-specific learning, Ye et al. [28] explored scaling up robot learning via Internet videos, investigating how web-scale human video datasets can enhance policy learning efficiency.

Flow-Guided Imitation Learning. A growing body of work has explored representing manipulation trajectories in more flexible and generalizable forms to enhance policy learning. Wen et al. [7] introduced Any-point Trajectory Modeling (ATM), which allows policies to query and generate trajectory states at arbitrary temporal points. This continuous-time representation enables efficient interpolation and improved temporal flexibility, reducing dependence on fixed-horizon action sequences. Bharadhwaj et al. [29] proposed Track2Act, which leverages point tracks extracted from Internet videos to learn generalizable manipulation skills. By grounding policies in persistent motion tracks, their method facilitates robust transfer across embodiments and visual domains without requiring paired robot demonstrations.

Another emerging line of research focuses on using *flow*-based representations as a domain-agnostic interface for manipulation. Xu et al. [8] introduced a method that predicts dense optical flow fields as an intermediate representation, enabling the transfer of skills across domains by decoupling perception from control. Extending this idea, Yuan et al. [9] proposed general flow as a unified affordance representation for scalable robot learning, demonstrating its capacity to generalize across object categories and manipulation tasks. Flow-based interfaces effectively capture motion intent while enabling cross-domain consistency. Ren et al. [30] introduced Motion Tracks, an agent-agnostic trajectory representation for few-shot human-to-robot transfer, enabling rapid adaptation to new tasks with minimal robot-specific data.

All these works leverage dense optical flow or motion tracks as intermediate representations, enabling skill transfer across domains. These approaches have shown strong results using non-robot videos to improve robot policies. Our method differs in the type of representation: rather than relying on flow fields, we use weak hand–object segmentation masks amplified by temporal tracking (Sec. III), aiming to capture higher-level semantic interactions.

III. Method

Our approach facilitates learning robot manipulation policies from limited *target-domain* data by leveraging knowledge distilled from large-scale *source-domain* (human)

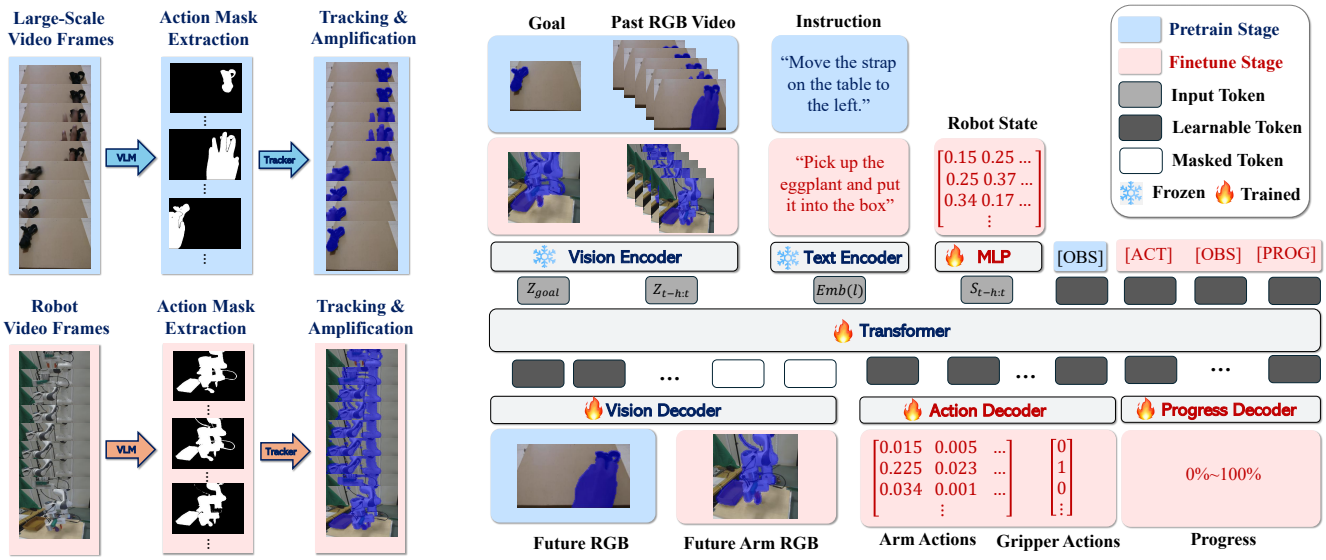

Fig. 2: **ViSA-Flow Architecture and Policy Learning Framework.** (**a**) During pretraining, hand-object interaction masks are extracted from large-scale video frames and amplified via tracking to generate semantic flow representations. (**b**) In the finetuning stage, a multi-modal Transformer architecture conditions on the goal image, a sequence of RGB observation frames enhanced with pre-trained ViSA-Flow, language instructions and robot state. The Transformer predicts future visual states, low-level robot actions, and task progress using dedicated decoders.

videos. This is achieved through the introduction and utilization of **Video Semantic Action Flow (ViSA-Flow)**, a structured intermediate representation designed for cross-domain transfer. We first formulate the conceptual properties of ViSA-Flow and motivate its suitability for transfer learning, then detail its concrete implementation via our two-stage learning framework.

A. Problem Definition

Our objective is to pretrain a policy model π_θ by utilizing human-object interactions from a large dataset of human manipulation videos, $D_v = \{v_i\}^M$. This pretraining aims to facilitate learning on a target robotic task using only a small dataset of robot demonstrations, $D_\tau = \{\tau_j\}^N$, where $N \ll M$. The target task involves controlling a robot based on language instructions, observations, and proprioceptive state. We define the robot's observation space as O, its proprioceptive state space as S, and its action space as A. Given a language instruction l, our goal is to learn a policy $\pi_\theta(a_t|l, o_{t-h:t}, s_{t-h:t})$ that outputs an action $a_t \in A$ based on the instruction l, a history of recent observations $o_{t-h:t} \in O$, and recent states $s_{t-h:t} \in S$. This policy is learned primarily by imitating the demonstrations in D_τ, leveraging the pretraining from D_v.

B. ViSA-Flow Representation

We propose ViSA-Flow as an intermediate representation $z_t \in Z_{\text{ViSA-Flow}}$ obtained by mapping an observation o_t and context l through a function $f : O \times L \to Z_{\text{ViSA-Flow}}$. We design $Z_{\text{ViSA-Flow}}$ to preserve task-relevant interactions while mitigating domain-specific nuisance factors, enabling cross-domain transfer.

a) Semantic Entity Grounding.: Given the initial observation frame o_0 and context l, we utilize a pre-trained Vision-Language Model (VLM) to ground textual descriptions of the manipulator (e.g., 'hand', 'gripper') and task-relevant objects (e.g., 'red block') identified from l. A segmentation model (e.g., SAM [31]) then generates initial segmentation masks for these grounded entities, including manipulators and objects, *i.e.*, $\{m_{M,0}, m_{O_k,0}\}$.

b) Hand-Object Interaction Tracking.: Due to the instability of semantic segmentation across sequential frames, we propose tracking the correctly segmented hand-object interaction mask over time. Specifically, we instantiate a robust point tracker (e.g., CoTracker [32]) with points densely sampled within the initial masks. The tracker estimates the 2D image trajectories $P_t = \{p_{j,t}\}_{j=0}^J$ for these points across the sequence $\{o_t\}_{t=0}^T$. These trajectories P_t represent the extracted raw flow information.

c) Flow-Conditioned Feature Encoding.: To produce the final VISA-Flow representation z_t, we encode the flow information P_t into a rich feature vector while retaining visual context. We first apply a perceptual enhancement process directly on the raw observation frame o_t. Using tracked point trajectories P_t, we generate a spatially-localized amplification mask $M_t(x, y)$ with parameterized radius r around each tracker coordinate:

$$M_t(x, y) = \max_{p \in P_t} \mathbf{1}\big(\|(x, y) - p\|_2 \le r\big). \tag{1}$$

This mask modulates pixel intensities by an amplification factor α within these regions of interest, while maintaining contextual information elsewhere. The resulting perceptually-enhanced frame exhibits selective luminance amplification at interaction-critical regions. This pre-

processed frame is then passed through a pre-trained vision encoder ϕ (e.g., MAE [33]) which is frozen during policy learning, transforming the flow-highlighted observations into our implemented ViSA-Flow representation z_t:

$$z_t = \phi\big(o_t \odot [\,1 + \alpha\, M_t\,]\big). \tag{2}$$

This implementation aims to focus on tracked semantic entities and modulating features accordingly.

C. Policy Learning through ViSA-Flow Representation

Our learning framework leverages the extracted ViSA-Flow representations z_t within a two-stage pre-training and fine-tuning scheme, implemented using a transformer architecture, denoted g_ψ (parameters ψ), inspired by prior work such as GR-1 [25].

a) Model Architecture.: A transformer g_ψ is designed to process multimodal sequences for both generative prediction and policy inference shown in Fig. 2. Its input is a sequence formed by concatenating tokens representing various modalities and special learnable query tokens. Primary input modalities include language instruction embeddings $\text{Emb}(l)$ (e.g., from CLIP [34]), the sequence of recent ViSA-Flow representations $\{z_{t-h}, ..., z_t\}$ encoding flow-conditioned visual features (Sec. III-B), the sequence of proprioceptive states $\{s_{t-h}, ..., s_t\}$ (processed via linear embeddings), and potentially tokens representing a goal state z_{goal}. Added to these are special query tokens: an `[ACT]` token for action prediction and multiple `[OBS]` tokens for predicting future ViSA-Flow states. Standard positional embeddings are added to this combined sequence to encode temporal order before processing by the transformer blocks. The output embeddings corresponding to the query tokens are then directed to task-specific heads; notably, the `[ACT]` token's output yields the action chunk prediction $\hat{a}_{t+1:t+k}$, while the `[OBS]` tokens' outputs yield predictions $\hat{z}_{t+1:t+n}$ for future states.

b) Stage 1: Pre-training – Learning ViSA-Flow Dynamics Prior.: Using the large-scale human video dataset D_v, we pre-train g_ψ to model the dynamics within the ViSA-Flow space. For each sequence $v_i \in D_v$, we extract $\{z_{i,t}\}$ (Sec. III-B). The model is trained to predict future representations $z_{t+1:t+n}$ based on past context $z_{\le t}$ and l, using the `[OBS]` query tokens. The objective is to minimize the prediction error, typically via Mean Squared Error (MSE):

$$\mathcal{L}_{\text{pretrain}}(\psi) = \mathbb{E}_{v\sim D_v}\left[||g_\psi(z_{\le t}, l)_{\text{[OBS]}} - z_{t+1:t+n}||^2\right]. \tag{3}$$

This stage yields pre-trained parameters ψ_{pre}, encoding a prior over interaction dynamics.

c) Stage 2: Fine-tuning – Policy Adaptation.: Using the small-scale robot demonstration dataset D_τ, we fine-tune the model, initialized with ψ_{pre}, to learn the target policy π_θ (where $\theta \subseteq \psi$). For each robot trajectory $\tau_j \in D_\tau$, we extract ViSA-Flow representations $\{z_{j,t}\}$ using the identical pipeline. The model is trained end-to-end with a multi-task objective combining action prediction and continued dynamics modeling:

$$\begin{aligned}\mathcal{L}_{\text{finetune}}(\psi) = \mathbb{E}_{\tau\sim D_\tau}\Big[\mathcal{L}_{\text{act}}\big(a_{t+1:t+k},\, \hat{a}_{t+1:t+k}\big) \\ + \lambda_{\text{fwd}}\,\mathcal{L}_{\text{obs}}\big(z_{t+1:t+n},\, \hat{z}_{t+1:t+n}\big) + \lambda_{\text{prog}}\,\mathcal{L}_{\text{prog}}\big(p_t,\, \hat{p}_t\big)\Big]\end{aligned} \tag{4}$$

Here, $\hat{a}_t = g_\psi(z_{\le t}, s_{\le t}, l)_{\text{[ACT]}}$ is the predicted action. $\mathcal{L}_{\text{act}}$ is the action loss combining Smooth L1 (joint regression), BCE (gripper command), and KL divergence (distribution regularization). $\hat{z}_{t+1:t+n} = g_\psi(z_{\le t}, s_{\le t}, l)_{\text{[OBS]}}$ are predicted future ViSA-Flow states, and $\mathcal{L}_{\text{obs}}$ is the forward dynamics loss (MSE, identical form to Eq. 3 but on D_τ) weighted by λ_{fwd}. $\hat{p}_t$ is the optional predicted progress, with $\mathcal{L}_{\text{prog}}$ being the progress loss (e.g., MSE) weighted by λ_{prog}. This stage adapts the general dynamics prior to the specific robot and learns the mapping from ViSA-Flow states (and proprioception) to robot actions, yielding the final policy parameters ψ.

IV. Evaluation

We conduct extensive experiments in both simulated and real-world environments to systematically evaluate ViSA-Flow's performance. Our evaluation is designed to answer the following key questions: 1) Can ViSA-Flow generalize across tasks with distractors, different backgrounds, and novel objects? 2) Can ViSA-Flow effectively learn and generalize across diverse tasks when expert demonstration data with language annotations are scarce? 3) Do semantic actions from human videos benefit robot skill learning?

A. Simulation Experiments

Evaluation Setup. We evaluate ViSA-Flow on the CALVIN benchmark [35], a standard testbed for long-horizon, language-conditioned manipulation requiring generalization. We use the ABC→D split, training on environments A, B, C and evaluating zero-shot on the unseen environment D as shown in the lower row of Fig. 3.

Pre-training Data. The ViSA-Flow model undergoes pre-training (Stage 1, Sec. III-C) using the large-scale Something-Something-V2 (SthV2) dataset [36] as the source domain. SthV2 contains approximately 220k short videos of human-object interactions with template text labels (examples visualized in the upper row of Fig. 3). The videos are processed to extract ViSA-Flow representations which are used for the pre-training as described in Secs. III-B and III-C.

Fine-tuning Data. Following pre-training, ViSA-Flow is fine-tuned (Stage 2, Sec. III-C) specifically for the CALVIN environment [35]. To evaluate performance under data scarcity, we utilize only **10%** (1,768 trajectories) of the available language-annotated robot demonstrations from CALVIN's ABC dataset as our target domain dataset. Each trajectory consists of the language instruction and the sequence of robot states, observations, and actions. Robot observations are processed into ViSA-Flow representations using the pipeline described in Sec. III-B.

Baselines. We compare ViSA-Flow against two groups of SOTA methods: *(i) Low-Data Baselines* trained on

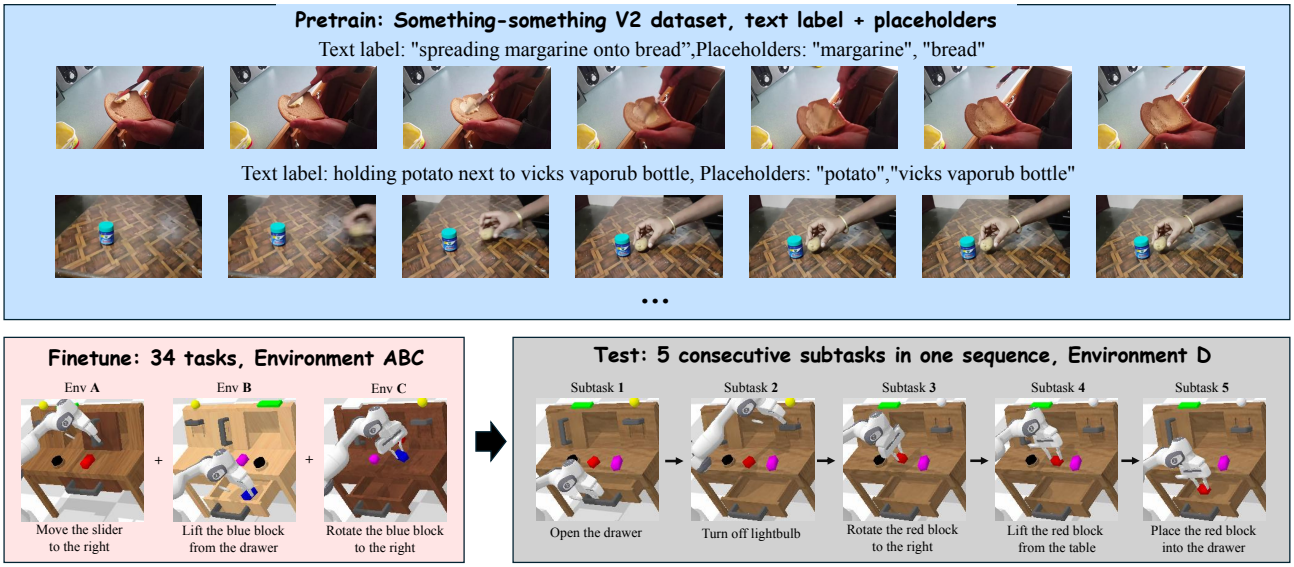

Fig. 3: **Datasets used for pretraining, finetuning, and evaluation.** We pretrain on Something-Something-V2 with text labels and placeholders to extract semantic action flow. We finetune on 34 tasks across CALVIN environments A–C and evaluate zero-shot on environment D [35], where the robot completes 5 consecutive subtasks in one sequence.

TABLE I: Comparative evaluation on CALVIN ABC→ D benchmark. Performance metrics include success rates for completing 1-5 consecutive tasks and average sequence length (Avg. Len). Methods in the top section use 100% of training data, while methods in the bottom section use only 10%. The robot executed 1,000 test sequences with five tasks each. **Bold** indicates best performance.

Method	Fully-Annotated Data (Demo No.)	Partially-Annotated Data	Tasks Completed in A Row					Avg. Len.
			1	2	3	4	5	
Hulc [13]	100% (17870)	✓	41.8%	16.5%	5.7%	1.9%	1.1%	0.67
MDT [20]	100% (17870)	✓	61.7%	40.6%	23.8%	14.7%	8.7%	1.54
Spil [21]	100% (17870)	✓	74.2%	46.3%	27.6%	14.7%	8.0%	1.71
Roboflamingo [22]	100% (17870)	✗	82.4%	61.9%	46.6%	33.1%	23.5%	2.47
SuSIE [14]	100% (17870)	✓	87.0%	69.0%	49.0%	38.0%	26.0%	2.69
ATM [7]	10% (1768)	✗	31.7%	5.1%	1.3%	0.0%	0.0%	0.43
CLOVER [11]	10% (1768)	✗	44.3%	18.0%	5.0%	1.0%	0.0%	0.68
GR-1 [25]	10% (1768)	✗	67.2%	37.1%	19.8%	10.8%	6.9%	1.41
SeeR [10]	10% (1768)	✗	65.5%	38.8%	21.4%	11.7%	6.8%	1.44
GR-MG [6]	10% (1768)	✗	81.8%	59.0%	39.0%	24.0%	16.2%	2.20
ViSA-Flow (Ours)	10% (1768)	✗	**89.0%**	**73.8%**	**56.8%**	**44.8%**	**31.4%**	**2.96**

the same 10% split for a fair data-efficiency comparison—ATM [7], a flow-based interface enabling any-point querying; CLOVER [11], a generative closed-loop visuomotor controller; GR-1 [25], a multimodal transformer pretrained on human videos; SeeR [10], a predictive inverse-dynamics approach; and GR-MG [6], a closely related transformer that augments GR-1 with explicit goal conditioning. We chose these baselines because they span complementary representations (flow-based, generative modeling, inverse dynamics, language/goal-conditioned transformers) and include models both with and without human-video pretraining, allowing us to isolate the effect of cross-domain priors. *(ii) Full-Data Baselines:* Methods trained on 100% of CALVIN annotated robot data (17,870 trajectories), including Hulc [13], MDT [20], Spil [21], Roboflamingo [22] and SuSIE [14]. These represent the performance achievable with substantially more in-domain supervision.

Metrics. Following the standard CALVIN evaluation protocol [35], we measure the success rate to complete 5 consecutive subtasks, evaluated over 1,000 independent sequences. We also report the average successful sequence length (Avg. Len.). These metrics assess single-task proficiency and the ability to maintain performance over long horizons.

Results and Analysis. Table I presents the performance metrics for all methods. The results demonstrate that ViSA-Flow outperforms all baseline methods, achieving highest success rates across all consecutive task completion metrics despite using only 10% of the available annotated robot trajectories. Most impressively, ViSA-Flow maintains strong performance in sequential tasks, completing 5 consecutive

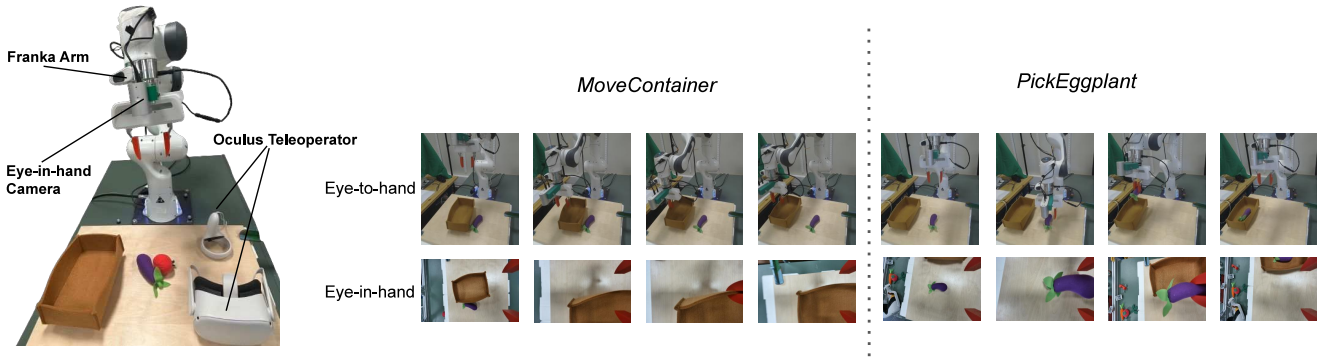

Fig. 4: **The real-world experiment setup.** We evaluate ViSA-Flow on two single-stage manipulation tasks and a two-stage long-horizon manipulation task.

tasks 31.4% of the time, almost twice the rate of the next best method trained with 10% data (GR-MG: 16.2%) and exceeding all methods trained on 100% data, including Susie (26.0%). The average sequence length of 2.96 further demonstrates the effectiveness of ViSA-Flow in handling long-horizon manipulation tasks. Performance degradation from single to sequential tasks (89.0% $\rightarrow$ 31.4%) is notably less severe for ViSA (64.7% reduction) compared to GR-MG (80.2% reduction) and Susie (70.1% reduction). This remarkable performance can probably be attributed to utilization of semantic action representations extracted from human demonstration videos. These results in simulation experiments validate our hypothesis that semantic action representations from human videos can significantly enhance robot skill learning, even when expert demonstrations are scarce and encounter different environments.

Ablation Study of ViSA-Flow Components. Table II summarizes the results when each component within the ViSA-Flow framework is individually removed from the full method. Critically, removing the human-video pre-training stage (**w/o pre.**) leads to a near collapse in performance, indicating that the dynamics prior distilled from large-scale human videos is essential for multi-step task success. Removing semantic entity grounding (**w/o Seg.**, with α=0) and tracking motion over whole images significantly reduces performance across all consecutive-task metrics: success on five-task sequences drops from 31.4% to 9.6%, and average successful length falls from 2.96 to 1.64, which indicates the importance of accurately segmenting and identifying semantic entities to anchor tracking and flow conditioning. Omitting temporal point tracking and instead segmenting each frame independently (**w/o Trace.**) decreases average successful length from 2.96 to 2.78, suggesting that temporal point correspondences help maintain consistent interaction-region emphasis across time . Excluding manipulator grounding (**w/o Hand**) yields a modest drop (2.96 to 2.83) that validates segmentation and tracking are primary drivers while manipulator cues still aid spatial context. Overall, the full ViSA-Flow—integrating segmentation, tracking, and manipulator grounding—achieves the best results, and abtion results confirm that each component contributes to reliable long-horizon, cross-domain execution.

TABLE II: Ablation study evaluating the contribution of key components in ViSA-Flow.

Method	Tasks Completed in A Row					Avg. Len.
	1	2	3	4	5	
ViSA-Flow w/o pre.	16.0%	1.6%	0.0%	0.0%	0.0%	0.18
ViSA-Flow w/o Seg.	71.3%	45.1%	24.5%	14.5%	9.6%	1.64
ViSA-Flow w/o Trace.	87.2%	69.2%	52.0%	39.6%	30.0%	2.78
ViSA-Flow w/o Hand	89.0%	71.8%	54.2%	39.4%	28.4%	2.83
ViSA-Flow (Full)	**89.0%**	**73.8%**	**56.8%**	**44.8%**	**31.4%**	**2.96**

TABLE III: Ablation study evaluating ViSA-Flow data scaling.

Data	Tasks Completed in a Row					Avg. Len.
	1	2	3	4	5	
5% ABC→D	84.6%	59.6%	41.0%	27.6%	18.4%	2.31
10% ABC→D	89.0%	73.8%	56.8%	44.8%	31.4%	2.96
50% ABC→D	**93.8%**	**85.4%**	**76.2%**	**68.8%**	**58.8%**	**3.83**

Data Scaling. In addition, we evaluate ViSA-Flow under different amounts of robot demonstration data (5%, 10%, and 50%). As shown in Table III, performance consistently improves as the amount of data increases, highlighting ViSA-Flow's scalability and data efficiency.

B. Real World Experiments

We evaluate the performance of ViSA-Flow in real-world experiments across diverse settings, focusing on its effectiveness and robustness in solving both single-stage and long-horizon tasks.

Experiment Setup. We evaluate our ViSA-Flow method in two real-world settings: two single-stage manipulation tasks and one long-horizon manipulation task. The demonstrations were collected by teleoperating a 7-DOF Franka Emika Panda arm using the Oculus-based application. We use two cameras (one eye-in-hand, one eye-to-hand) to provide RGB observations. The real-world experiment setup is shown in Fig. 4. For single-stage tasks, we collected 46 and 54 demonstrations for two tasks—*MoveContainer* and *PickEggplant* respectively. We train the ViSA-Flow policy for each single-stage task. For long-horizon tasks, we consider the same two subtasks, *MoveContainer* and *PickEggplant*, requiring the robot to complete the first task before sequentially solving the second. This setup ensures consistency with the testing

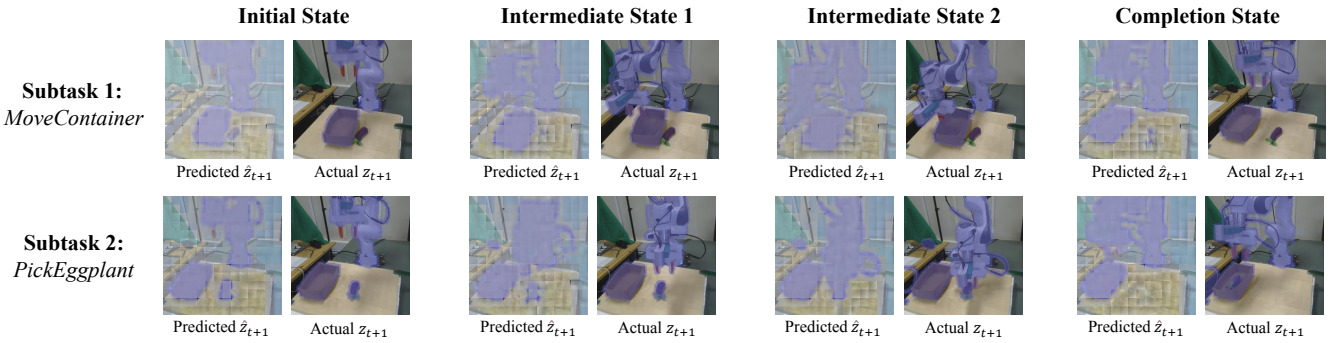

Fig. 5: **Qualitative results on the real world long-horizon task.** We visualize the *decoded* ViSA-Flow prediction at $\hat{z}_{t+1}$ against the *actual* ViSA-Flow z_{t+1} extracted from the next observation for four execution phases. Two rows correspond to the two subtasks that make up the long-horizon evaluation: **(Top)** *Subtask 1 – MoveContainer;* **(Bottom)** *Subtask 2 – PickEggplant.* Qualitatively, the model's one-step predictions closely follow the true motion of the manipulator and task-relevant objects, even as the scene evolves across distinct interaction stages.

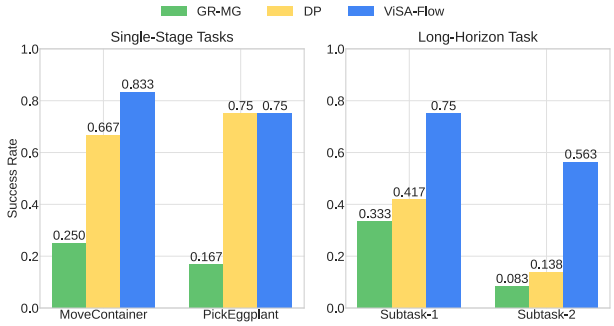

Fig. 6: **Real-world experimental results. Left:** two single-stage tasks; **Right:** a two-stage long-horizon task.

scenario used in our simulation experiments. We pre-train our ViSA-Flow model on an NVIDIA RTX 4090 for 30 epochs, followed by fine-tuning for 30 epochs on single-stage tasks and 50 epochs for long-horizon tasks, respectively. We evaluate each policy across 12 different initial positions.

Baselines. We compare our ViSA-Flow method with GR-MG [6] and the visuomotor Diffusion Policy (DP) [37], both of which leverage RGB and proprioceptive inputs. To ensure fair comparison, all baseline models are trained on the same real-world demonstration datasets for the two single-stage tasks and the long-horizon task.

Quantitative Results and Analysis. The real-world experimental results are presented in Fig. 6. For the single-stage tasks *MoveContainer* and *PickEggplant*, ViSA-Flow significantly outperforms the GR-MG model across 12 trials. Meanwhile, DP achieves a comparable success rate of 75.0% on the *PickEggplant* task. In contrast, for the long-horizon task—which sequentially combines *MoveContainer* and *PickEggplant*—our method demonstrates superior performance, achieving 9/12 successful trials for each subtask and yielding an overall success rate of 56.3% for the full sequence. By comparison, GR-MG and DP attain success rates of only 8.3% and 13.8%, respectively. Notably, DP experiences a significant performance drop when transitioning from single-stage to long-horizon tasks, whereas ViSA-Flow maintains robust and consistent performance.

Qualitative Results and Analysis. Fig. 5 qualitatively demonstrates that the decoded ViSA-Flow one-step prediction $\hat{z}_{t+1}$ remains tightly aligned with the ground-truth flow throughout the entire long-horizon execution: the model persistently focuses on the robot gripper and the task-relevant objects while suppressing background clutter, its spatial support evolves smoothly and coherently as the scene transitions from the initial approach, through two intermediate contact phases, to the completion state, and the same level of accuracy is observed across the two sequential subtasks. This close match between prediction and observation confirms that the cross-domain dynamics prior learned during pretraining effectively captures task-critical interaction structure and generalizes to novel real-world embodiments.

V. Limitations and Future Work

While ViSA-Flow demonstrates strong performance in observational robot learning, it currently lacks explicit modeling of 3D geometry and contact dynamics, which may limit its generalization to tasks requiring fine-grained physical interactions. The framework also relies on pretrained VLM components, potentially restricting adaptability to novel domains or unseen objects. Moreover, ViSA-Flow currently transfers only object–manipulator interactions from human manipulation videos to robots, ignoring more nuanced dexterous movements of human fingers. This simplification leads to some loss of manipulation knowledge, which may limit performance in tasks requiring fine or dexterous control.

Future work aims to address these limitations by enriching ViSA-Flow representations with contact physics and 3D reasoning, reducing reliance on pretrained models through joint or end-to-end training with VLMs, and integrating its learned priors with reinforcement learning to enhance policy

learning. Additionally, we plan to investigate methods to capture and transfer finer-grained human manipulation skills, preserving dexterous finger-level knowledge for robotic use. Scaling pretraining to large-scale video corpora and further analyzing ViSA-Flow's invariance properties and sample efficiency also represent promising directions for advancing robust and generalizable robot learning from observation.

REFERENCES

[1] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation learning: A survey of learning methods," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–35, 2017.

[2] T. Gao, S. Nasiriany, H. Liu, Q. Yang, and Y. Zhu, "Prime: Scaffolding manipulation tasks with behavior primitives for data-efficient imitation learning," *arXiv preprint arXiv:2403.00929*, 2024.

[3] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[4] S. Bahl, A. Gupta, and D. Pathak, "Human-to-robot imitation in the wild," *arXiv preprint arXiv:2207.09450*, 2022.

[5] J. Zeng, Q. Bu, B. Wang, W. Xia, L. Chen, H. Dong, H. Song, D. Wang, D. Hu, P. Luo *et al.*, "Learning manipulation by predicting interaction," *arXiv preprint arXiv:2406.00439*, 2024.

[6] P. Li, H. Wu, Y. Huang, C. Cheang, L. Wang, and T. Kong, "Gr-mg: Leveraging partially-annotated data via multi-modal goal-conditioned policy," *IEEE Robotics and Automation Letters*, 2024.

[7] C. Wen, X. Lin, J. So, K. Chen, Q. Dou, Y. Gao, and P. Abbeel, "Any-point trajectory modeling for policy learning," 2024. [Online]. Available: https://arxiv.org/abs/2401.00025

[8] M. Xu, Z. Xu, Y. Xu, C. Chi, G. Wetzstein, M. Veloso, and S. Song, "Flow as the cross-domain manipulation interface," 2024. [Online]. Available: https://arxiv.org/abs/2407.15208

[9] C. Yuan, C. Wen, T. Zhang, and Y. Gao, "General flow as foundation affordance for scalable robot learning," 2024. [Online]. Available: https://arxiv.org/abs/2401.11439

[10] Y. Tian, S. Yang, J. Zeng, P. Wang, D. Lin, H. Dong, and J. Pang, "Predictive inverse dynamics models are scalable learners for robotic manipulation," 2024. [Online]. Available: https://arxiv.org/abs/2412.15109

[11] Q. Bu, J. Zeng, L. Chen, Y. Yang, G. Zhou, J. Yan, P. Luo, H. Cui, Y. Ma, and H. Li, "Closed-loop visuomotor control with generative expectation for robotic manipulation," 2024. [Online]. Available: https://arxiv.org/abs/2409.09016

[12] Q. Yang, M. C. Welle, D. Kragic, and O. Andersson, "S^2-diffusion: Generalizing from instance-level to category-level skills in robot manipulation," *arXiv preprint arXiv:2502.09389*, 2025.

[13] O. Mees, L. Hermann, and W. Burgard, "What matters in language conditioned robotic imitation learning over unstructured data," 2022. [Online]. Available: https://arxiv.org/abs/2204.06252

[14] K. Black *et al.*, "Zero-shot robotic manipulation with pretrained image-editing diffusion models," *arXiv preprint arXiv:2310.10639*, 2023.

[15] A. Jonnavittula, S. Parekh, and D. P. Losey, "View: Visual imitation learning with waypoints," 2025. [Online]. Available: https://arxiv.org/abs/2404.17906

[16] J. Gao, Z. Tao, N. Jaquier, and T. Asfour, "K-vil: Keypoints-based visual imitation learning," *IEEE Transactions on Robotics*, vol. 39, no. 5, p. 3888–3908, Oct. 2023. [Online]. Available: http://dx.doi.org/10.1109/TRO.2023.3286074

[17] K. Hu, Z. Rui, Y. He, Y. Liu, P. Hua, and H. Xu, "Stem-ob: Generalizable visual imitation learning with stem-like convergent observation through diffusion inversion," 2024. [Online]. Available: https://arxiv.org/abs/2411.04919

[18] K. Kim, M. Lee, M. Whoo Lee, K. Shin, M. Lee, and B.-T. Zhang, "Visual hindsight self-imitation learning for interactive navigation," *IEEE Access*, vol. 12, p. 83796–83809, 2024. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2024.3413864

[19] A. Brohan *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[20] M. Reuss, Ömer Erdinç Yağmurlu, F. Wenzel, and R. Lioutikov, "Multimodal diffusion transformer: Learning versatile behavior from multimodal goals," 2024. [Online]. Available: https://arxiv.org/abs/2407.05996

[21] H. Zhou, Z. Bing, X. Yao, X. Su, C. Yang, K. Huang, and A. Knoll, "Language-conditioned imitation learning with base skill priors under unstructured data," 2024. [Online]. Available: https://arxiv.org/abs/2305.19075

[22] X. Li, M. Liu, H. Zhang, C. Yu, J. Xu, H. Wu, C. Cheang, Y. Jing, W. Zhang, H. Liu, H. Li, and T. Kong, "Vision-language foundation models as effective robot imitators," 2024. [Online]. Available: https://arxiv.org/abs/2311.01378

[23] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, and S. Levine, "Time-contrastive networks: Self-supervised learning from video," 2018. [Online]. Available: https://arxiv.org/abs/1704.06888

[24] B. Baker, I. Akkaya, P. Zhokhov, J. Huizinga, J. Tang, A. Ecoffet, B. Houghton, R. Sampedro, and J. Clune, "Video pretraining (vpt): Learning to act by watching unlabeled online videos," 2022. [Online]. Available: https://arxiv.org/abs/2206.11795

[25] H. Wu, Y. Jing, C. Cheang, G. Chen, J. Xu, X. Li, M. Liu, H. Li, and T. Kong, "Unleashing large-scale video generative pre-training for visual robot manipulation," 2023. [Online]. Available: https://arxiv.org/abs/2312.13139

[26] V. Jain, M. Attarian, N. J. Joshi, A. Wahid, D. Driess, Q. Vuong, P. R. Sanketi, P. Sermanet, S. Welker, C. Chan, I. Gilitschenski, Y. Bisk, and D. Dwibedi, "Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers," 2024. [Online]. Available: https://arxiv.org/abs/2403.12943

[27] M. Lepert, J. Fang, and J. Bohg, "Phantom: Training robots without robots using only human videos, 2025," *URL https://arxiv.org/abs/2503.00779*.

[28] W. Ye, F. Liu, Z. Ding, Y. Gao, O. Rybkin, and P. Abbeel, "Video2policy: Scaling up manipulation tasks in simulation through internet videos," 2025. [Online]. Available: https://arxiv.org/abs/2502.09886

[29] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani, "Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation," 2024. [Online]. Available: https://arxiv.org/abs/2405.01527

[30] J. Ren, P. Sundaresan, D. Sadigh, S. Choudhury, and J. Bohg, "Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning," 2025. [Online]. Available: https://arxiv.org/abs/2501.06994

[31] A. Kirillov *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[32] N. Karaev, I. Makarov, J. Wang, N. Neverova, A. Vedaldi, and C. Rupprecht, "Cotracker3: Simpler and better point tracking by pseudo-labelling real videos," 2024. [Online]. Available: https://arxiv.org/abs/2410.11831

[33] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," 2021. [Online]. Available: https://arxiv.org/abs/2111.06377

[34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

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

[36] R. Goyal, S. E. Kahou, V. Michalski, J. Materzyńska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The "something something" video database for learning and evaluating visual common sense," 2017. [Online]. Available: https://arxiv.org/abs/1706.04261

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