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

Author Names Omitted for Anonymous Review. Paper-ID 6

***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[13, 9] to intricate assembly procedures[7]. However, the scalability of traditional 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. Whether it be in person, instructional videos or even from sports broadcasts, humans instinctively focus on the 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.

Recent research[1, 32, 19] has explored enabling robots to acquire skills by directly observing unstructured human videos. These approaches have demonstrated strong generalizability, allowing robots to adapt to new tasks effectively. In most real-world scenarios, when humans learn a skill, we primarily focus on the interaction between the human hand (or arm) and the manipulated object, while disregarding irrelevant background elements or distractions. Mimicking this selective attention mechanism could enhance the efficiency and effectiveness of robot learning from videos.

Drawing inspiration from this, we propose a novel approach that enables robots to learn skills by extracting and leveraging semantic representations from large-scale video collections. Our framework outlined in Fig. 1 focuses on identifying the key semantic elements relevant to skill acquisition, much like how humans naturally attend to meaningful components while learning from visual demonstrations. By concentrating on these semantic features - such as object interactions, body poses, and motion patterns - rather than processing entire scenes indiscriminately, our approach aims to make video-based skill learning more efficient and generalizable. 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 Internet-scale human video data to robotic manipulation policies.
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. Our method boosts task success rates, highlighting the effectiveness of video-driven robot skill learning.

II. Related Work

Visual-Feature-Based Imitation Learning. Recent advancements[28, 5, 31, 21, 3] in visual feature-based imitation learning have significantly improved the efficiency, generalization, and robustness 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

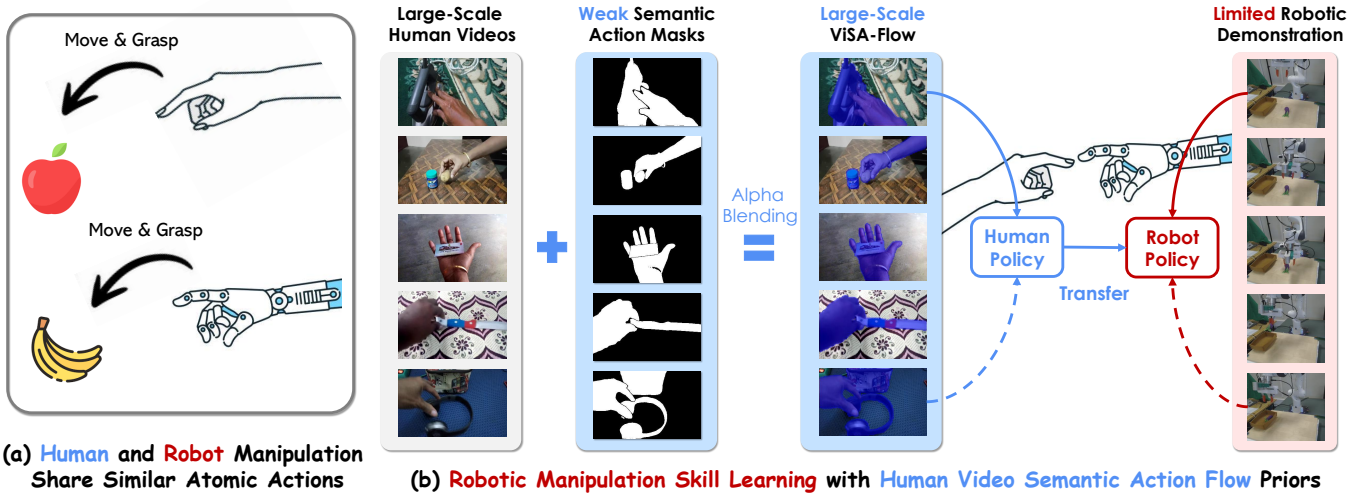

Fig. 1: **Learning Robot Manipulation Skills from Human Videos via Semantic Action Transfer.** (**a**) Humans and robots often share underlying atomic actions for similar tasks (e.g., Move & Grasp). (**b**) Our framework leverages large-scale, unlabeled human videos by extracting weakly supervised semantic action flow priors (ViSA-Flow). This knowledge is distilled into a human policy and efficiently transferred to learn a corresponding robot policy.

[8] 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 [12] 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 [17] introduces hindsight goal re-labeling and prototypical goal embedding, enhancing sample efficiency in vision-based tasks.

Video-Based Robot Learning. Recent advancements[4, 24, 33, 20] 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) [26] have pioneered the extraction of temporally consistent features to align human demonstrations with robot actions. Building on this foundation, video pretraining [2] has shown that large-scale video data can be used to pretrain robust visual representations for downstream manipulation tasks. More recent works[30] have further leveraged large-scale video datasets to enhance manipulation performance. Similarly, Vid2Robot [14] presents an end-to-end framework that directly translates video demonstrations and real-time observations into robot actions, leveraging cross-attention mechanisms for improved alignment. [19] highlights the potential of leveraging partially-annotated data to enhance robot policy learning by integrating multi-modal information.

Vision-Language Models for Robotics. Multi-modal models such as CLIP[23] and SAM[18] have shown considerable promise in grounding visual tasks via language. Building upon this foundation, recent studies have further enriched the interplay between vision and language in robotics. [27] proposed to couple vision-language embeddings with reinforcement learning to enhance adaptability in unfamiliar environments. In a similar vein, Wang et al. [29] introduced a multi-modal transformer architecture that fuses language cues with visual inputs, demonstrating improvements in object localization and manipulation planning. Rodriguez et al.[25] showed that fine-tuning pre-trained vision-language models on domain-specific datasets can markedly boost action segmentation and affordance detection accuracy.

III. Method

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

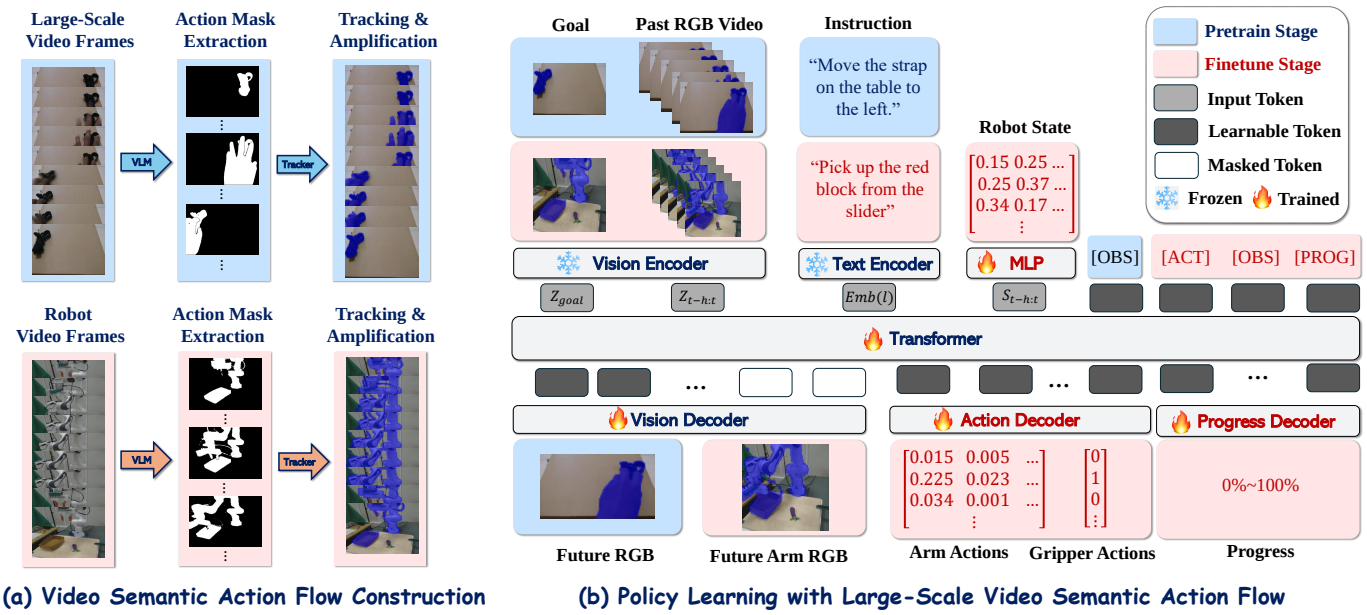

Fig. 2: **ViSA-Flow Architecture and Policy Learning Framework.** (**a**) During pretraining, hand-object interaction masks are extracted from large-scale video frames and amplified via tracking to generate semantic flow representations. (**b**) In the finetuning stage, a multi-modal Transformer architecture 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.

$s_{t-h:t} \in S$. 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 $f : O \times L \to Z_{\text{ViSA-Flow}}$. The motivation is to define a representation space $Z_{\text{ViSA-Flow}}$ where the manipulation interaction relevant to the task is preserved while the domain-specific nuisance factors are mitigated, facilitating skill transfer from O^S to O^T.

a) 1) 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[18]) then generates initial segmentation masks for these grounded entities, including manipulators and objects, *i.e.*, $\{m_{M,0}, m_{O_k,0}\}$.

b) 2) 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[16]) with points densely sampled within the initial masks. The tracker estimates the 2D image trajectories $P_t = \{p_{j,t}\}_{j=0}^{N}$ for these points across the sequence $\{o_t\}_{t=0}^{T}$. These trajectories P_t represent the extracted raw flow information, capturing the motion of key interaction points.

c) 3) 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 vision encoder ϕ (e.g., MAE[11]), 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[30].

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 Emb(l) (e.g., from CLIP[23]), 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 $z_{\leq t}$ and l, using the `[OBS]` query tokens. The objective is to minimize the prediction error, typically via Mean Squared Error (MSE):

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

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:

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

Here, $\hat{a}_t = g_\psi(z_{\leq t}, s_{\leq t}, l)_{\text{[ACT]}}$ is the predicted action. $\mathcal{L}_{\text{act}}$ is the action loss (e.g., a weighted combination of Smooth L1, BCE, KL divergence terms appropriate for the action space). $\hat{z}_{t+1:t+n} = g_\psi(z_{\leq t}, s_{\leq t}, l)_{\text{[OBS]}}$ 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 effectively learn and generalize across multiple tasks, particularly in challenging scenarios involving distractors, different backgrounds, and new objects? 2) Can ViSA-Flow effectively learn and generalize across diverse tasks using minimal expert demonstration data, particularly in scenarios where expert demonstration data with language annotations are scarce? 3) Do semantic actions extracted from human demonstrations benefit robot skill learning?

A. Simulation Experiments

Evaluation Setup. We evaluate ViSA-Flow on the CALVIN benchmark[22], 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[10] as the source domain. SthV2 contains approximately 220,000 short videos depicting diverse human-object interactions (examples visualized in the upper row of Fig. 3). Each video is associated with a template-based textual description indicating the action performed (e.g., 'Pushing [something] from left to right') and includes placeholder labels identifying key objects within frames. 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. 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.

Baselines. We compare ViSA-Flow against two groups of SOTA methods: *(i) Low-Data Baselines:* Strong contemporary methods trained under the identical 10% data condition as ViSA-Flow for direct comparison of data efficiency. This includes CLOVER[5], GR-1[30], SeeR[28] and GR-MG[19]. *(ii) Full-Data Baselines:* Methods trained on 100% of CALVIN annotated robot data (17,870 trajectories), including Hulc[21], MDT[24], Spil[33], Roboflamingo[20] and SuSIE[3]. These represent the performance achievable with substantially more in-domain supervision.

Metrics. Following the standard CALVIN evaluation protocol, we measure the success rate to complete 5 consecutive subtasks within a longer instruction sequence, evaluated over 1,000 independent sequences. We also report the average successful sequence length (Avg. Len.). These metrics assess

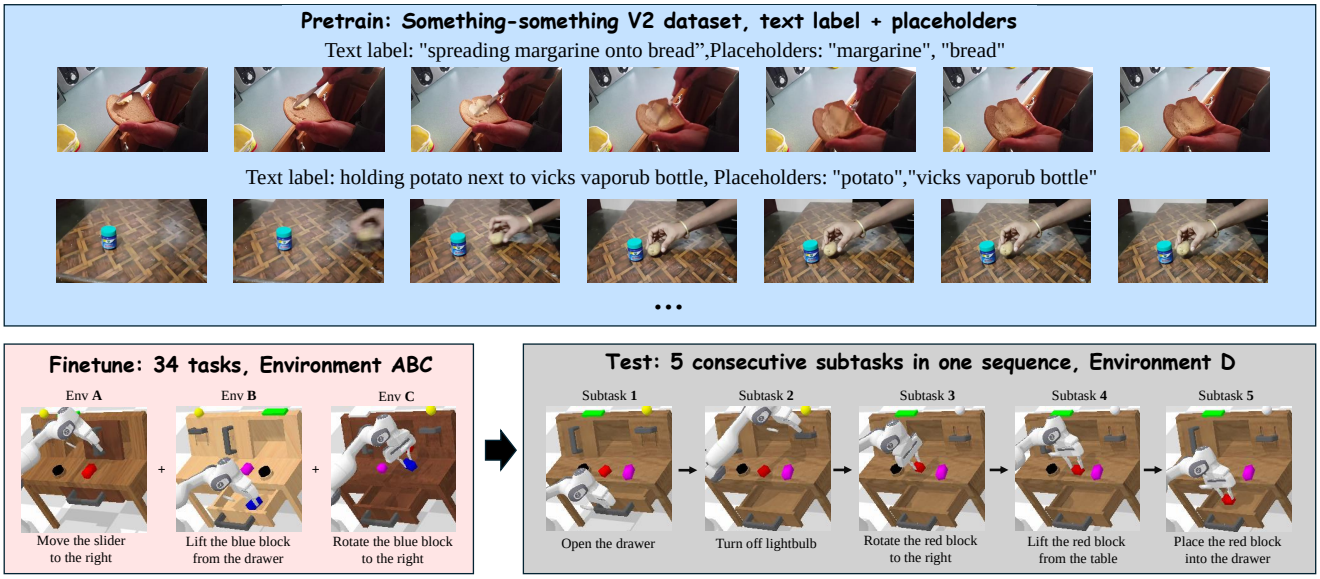

Fig. 3: **Datasets used for pretraining, finetuning, and evaluation.** A model is trained on the Something-Something-V2 dataset with text labels. Placeholders are used to extract underlying semantic action flow. The finetuning stage involves 34 manipulation tasks across three simulated environments (Env A, B, and C) in CALVIN benchmark [22]. The evaluation is on Environment D, where the robot complete 5 consecutive subtasks within one continuous 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 [21]	100% (17870)	✓	41.8%	16.5%	5.7%	1.9%	1.1%	0.67
MDT [24]	100% (17870)	✓	61.7%	40.6%	23.8%	14.7%	8.7%	1.54
Spil [33]	100% (17870)	✓	74.2%	46.3%	27.6%	14.7%	8.0%	1.71
Roboflamingo [20]	100% (17870)	✗	82.4%	61.9%	46.6%	33.1%	23.5%	2.47
SuSIE [3]	100% (17870)	✓	87.0%	69.0%	49.0%	38.0%	26.0%	2.69
CLOVER [5]	10% (1768)	✗	44.3%	18.0%	5.0%	1.0%	0.0%	0.68
GR-1 [30]	10% (1768)	✗	67.2%	37.1%	19.8%	10.8%	6.9%	1.41
SeeR [28]	10% (1768)	✗	65.5%	38.8%	21.4%	11.7%	6.8%	1.44
GR-MG [19]	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**

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 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% → 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. Removing the semantic entity grounding stage and tracking the

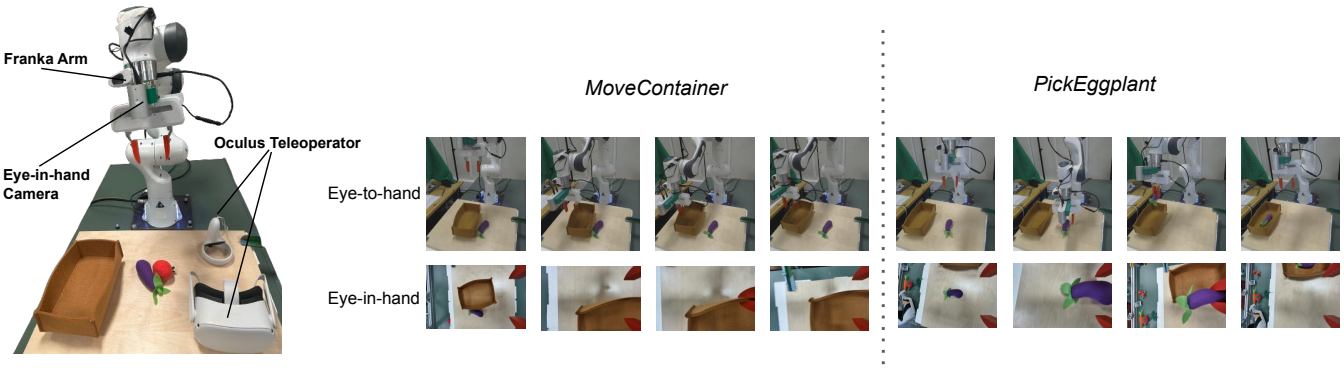

Fig. 4: **The real-world experiment setup.** We evaluate ViSA-Flow on two single-stage manipulation tasks and a two-stage long-horizon manipulation task.

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 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**

motion of points across whole observation images significantly reduces performance across all consecutive-task metrics. Success rate on five-task sequences drops from 31.4% to just 9.6% with the average successful length falling from 2.96 to 1.64, which indicates the importance of accurately segmenting and identifying semantic entities to anchor tracking and flow conditioning. Omitting the robust temporal tracking stage decreases the average successful length over five-task sequences from 2.96 to 2.78, highlighting that consistent point correspondences are essential for preserving temporal dynamics across multi-step interactions. Excluding explicit manipulator grounding results in a modest drop in average sequence length, from 2.96 to 2.83, indicating that while segmentation and tracking are primary drivers of performance, manipulator cues still play a meaningful role in providing spatial context for action understanding. Overall, the full ViSA-Flow configuration—integrating segmentation, tracking, and manipulator grounding—achieves the best results across all metrics, confirming that each component contributes to capturing semantic action flow and enabling reliable long-horizon, cross-domain task execution.

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 scenario used in our simulation experiments. We evaluate each policy across 12 different initial positions.

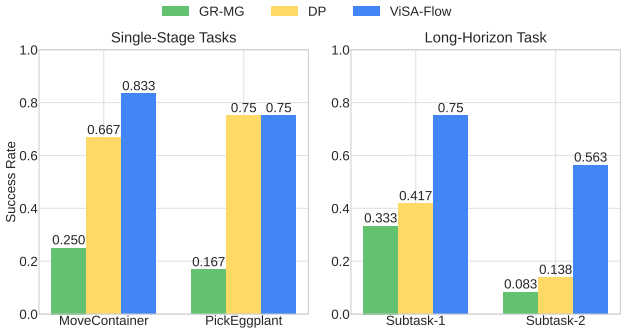

Fig. 5: **Real-world experimental results. Left:** two single-stage tasks; **Right:** a two-stage long-horizon task.

Baselines. We compare our ViSA-Flow method with GR-MG [19] and the visuomotor Diffusion Policy (DP) [6], 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. 5. 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

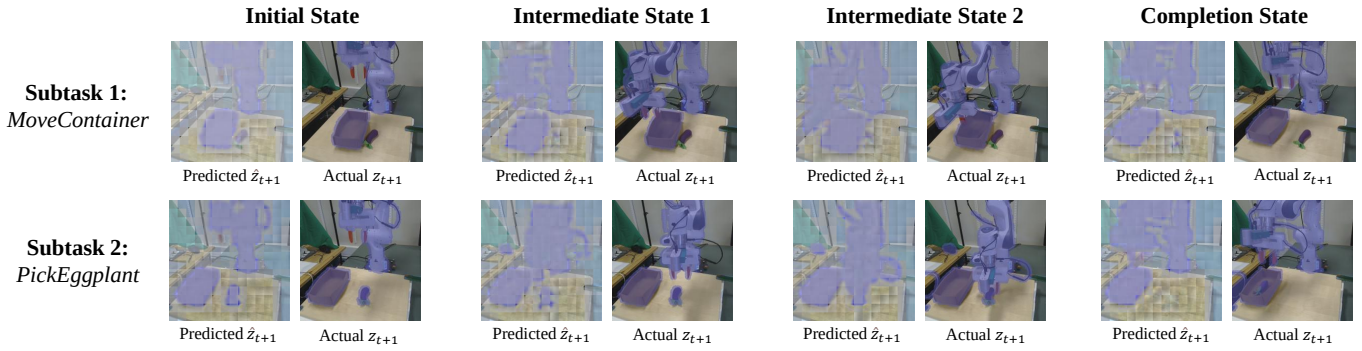

Fig. 6: **Qualitative results on the real world long-horizon task.** We visualize the *decoded* ViSA-Flow prediction at $\hat{z}_{t+1}$ against the *actual* ViSA-Flow z_{t+1} extracted from the next observation for four execution phases. Two rows correspond to the two subtasks that make up the long-horizon 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.

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. 6 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 involving fine-grained physical interactions. The current framework also relies on pretrained VLM components that potentially restrict adaptability to novel domains. Future work includes enriching ViSA-Flow representations with contact physics and reducing reliance on pretrained components by jointly training ViSA-Flow with VLMs. Additionally, integrating ViSA-Flow's priors with reinforcement learning algorithms and scaling pretraining to web-scale video corpora offer promising directions for advancing generalizable robot learning.

References

[1] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022.

[2] Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos, 2022. URL https://arxiv.org/abs/2206.11795.

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

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

[5] Qingwen Bu, Jia Zeng, Li Chen, Yanchao Yang, Guyue Zhou, Junchi Yan, Ping Luo, Heming Cui, Yi Ma, and Hongyang Li. Closed-loop visuomotor control with generative expectation for robotic manipulation, 2024. URL https://arxiv.org/abs/2409.09016.

[6] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[7] 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*, page 02783649241273668, 2023.

[8] Jianfeng Gao, Zhi Tao, Noémie Jaquier, and Tamim Asfour. K-vil: Keypoints-based visual imitation learning. *IEEE Transactions on Robotics*, 39(5):3888–3908, October 2023. ISSN 1941-0468. doi: 10.1109/tro.2023.3286074. URL http://dx.doi.org/10.1109/TRO.2023.3286074.

[9] Tian Gao, Soroush Nasiriany, Huihan Liu, Quantao Yang, and Yuke Zhu. Prime: Scaffolding manipulation tasks with behavior primitives for data-efficient imitation learning. *arXiv preprint arXiv:2403.00929*, 2024.

[10] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michal-

ski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense, 2017. URL https://arxiv.org/abs/1706.04261.

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. URL https://arxiv.org/abs/2111.06377.

[12] Kaizhe Hu, Zihang Rui, Yao He, Yuyao Liu, Pu Hua, and Huazhe Xu. Stem-ob: Generalizable visual imitation learning with stem-like convergent observation through diffusion inversion, 2024. URL https://arxiv.org/abs/2411.04919.

[13] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50 (2):1–35, 2017.

[14] Vidhi Jain, Maria Attarian, Nikhil J Joshi, Ayzaan Wahid, Danny Driess, Quan Vuong, Pannag R Sanketi, Pierre Sermanet, Stefan Welker, Christine Chan, Igor Gilitschenski, Yonatan Bisk, and Debidatta Dwibedi. Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers, 2024. URL https://arxiv.org/abs/2403.12943.

[15] Ananth Jonnavittula, Sagar Parekh, and Dylan P. Losey. View: Visual imitation learning with waypoints, 2025. URL https://arxiv.org/abs/2404.17906.

[16] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos, 2024. URL https://arxiv.org/abs/2410.11831.

[17] Kibeom Kim, Moonhoen Lee, Min Whoo Lee, Kisung Shin, Minsu Lee, and Byoung-Tak Zhang. Visual hindsight self-imitation learning for interactive navigation. *IEEE Access*, 12:83796–83809, 2024. ISSN 2169-3536. doi: 10.1109/access.2024.3413864. URL http://dx.doi.org/10.1109/ACCESS.2024.3413864.

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

[19] Peiyan Li, Hongtao Wu, Yan Huang, Chilam Cheang, Liang Wang, and Tao Kong. Gr-mg: Leveraging partially-annotated data via multi-modal goal-conditioned policy. *IEEE Robotics and Automation Letters*, 2024.

[20] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. Vision-language foundation models as effective robot imitators, 2024. URL https://arxiv.org/abs/2311.01378.

[21] Oier Mees, Lukas Hermann, and Wolfram Burgard. What matters in language conditioned robotic imitation learning over unstructured data, 2022. URL https://arxiv.org/abs/2204.06252.

[22] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):7327–7334, 2022.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

[24] Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals, 2024. URL https://arxiv.org/abs/2407.05996.

[25] Maria Rodriguez, Carlos Hernandez, and Sung Lee. Domain-specific fine-tuning of vision-language models for robotic action segmentation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2000–2007, 2024.

[26] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video, 2018. URL https://arxiv.org/abs/1704.06888.

[27] Fang Shi, Li Zhang, and Yong Wu. Vision-language embeddings for adaptive robot control. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1500–1507, 2022.

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

[29] Mei Wang, Chen Zhao, and Jun Li. Multimodal transformer for enhanced object localization in robotic manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1020–1028, 2023.

[30] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation, 2023. URL https://arxiv.org/abs/2312.13139.

[31] Quantao Yang, Michael C Welle, Danica Kragic, and Olov Andersson. S^2-diffusion: Generalizing from instance-level to category-level skills in robot manipulation. *arXiv preprint arXiv:2502.09389*, 2025.

[32] Jia Zeng, Qingwen Bu, Bangjun Wang, Wenke Xia, Li Chen, Hao Dong, Haoming Song, Dong Wang, Di Hu, Ping Luo, et al. Learning manipulation by predicting interaction. *arXiv preprint arXiv:2406.00439*, 2024.

[33] Hongkuan Zhou, Zhenshan Bing, Xiangtong Yao, Xiaojie Su, Chenguang Yang, Kai Huang, and Alois Knoll. Language-conditioned imitation learning with base skill priors under unstructured data, 2024. URL https://arxiv.org/abs/2305.19075.