Appendix

A. Additional Experimental Results

A.1. Visualization of Optical Flow

As illustrated in Fig. 6, we visualize the optical flow for a representative complete workflow of a task in LIBERO. SPATIAL tests reasoning about spatial relationships for accurate bowl placement. OBJECT evaluates generalization across varying object types within the same layouts. GOAL examines adaptive, goal-oriented behavior under varying objectives. LONG focuses on long-horizon tasks with multiple subgoals, assessing multi-step planning with heterogeneous objects and layouts.

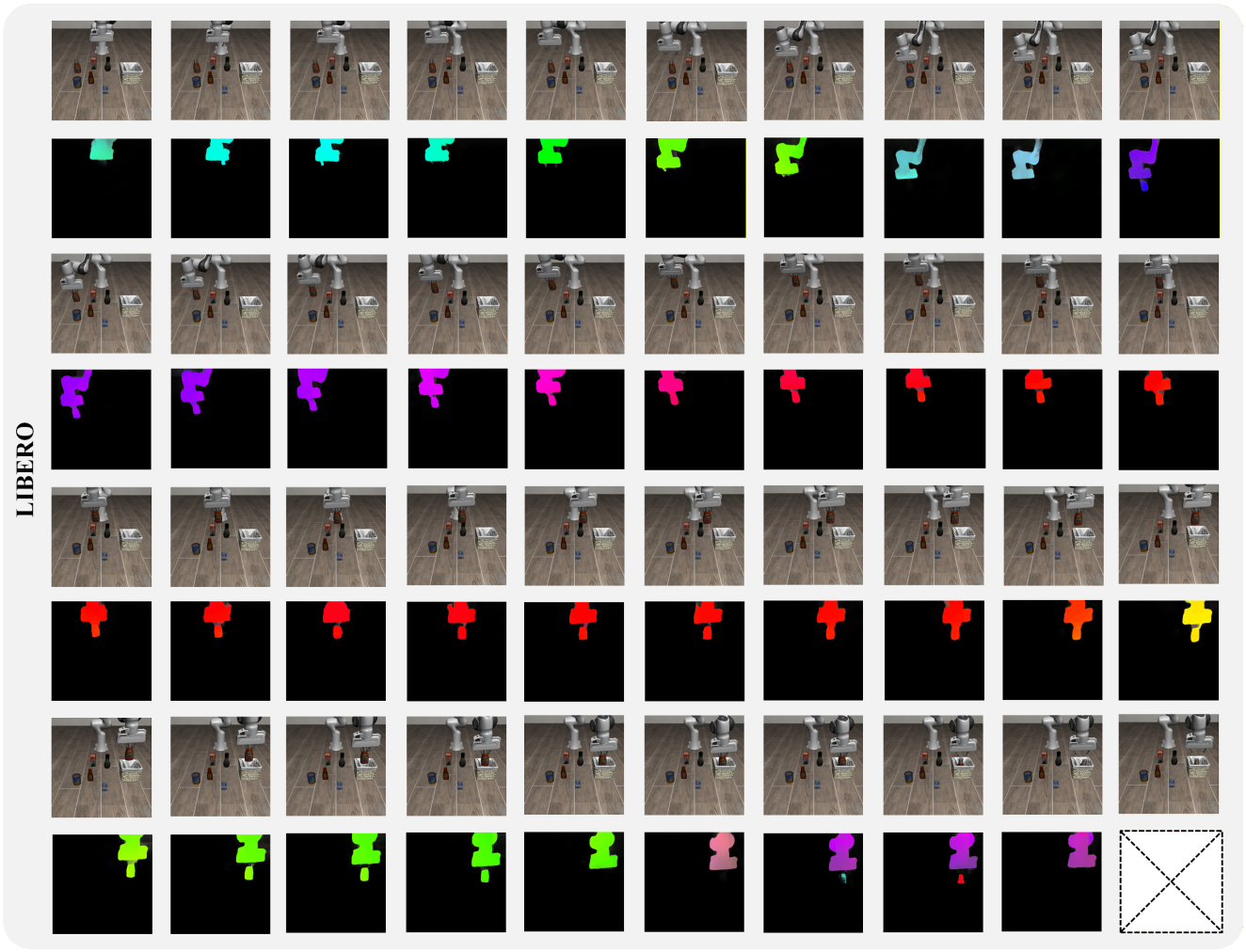


Figure 6. Visualization of optical flow for the LIBERO task: *"pickup the ketchup and place it in the basket"*. To illustrate the workflow, we sampled 40 frames from the full 147 frames sequence at intervals of 3 frames, preserving the temporal order.

Fig. 7 shows the workflows of four tasks on the PROCGEN benchmark, along with their corresponding raw and object-centric optical flow. The tasks are: BIGFISH, where the player grows by eating smaller fish while avoiding larger ones, receiving small rewards for eating and a large reward for becoming the biggest; CHASER, a maze navigation task in which the player collects green orbs while avoiding enemies, with power-ups that temporarily make enemies vulnerable; HEIST, where the player must steal a gem by collecting color-coded keys to unlock corresponding locks in a maze, with keys in possession displayed on-screen; and LEAPER, inspired by Frogger, in which the player crosses lanes of cars and hops on logs to traverse rivers, earning rewards for reaching the finish line.

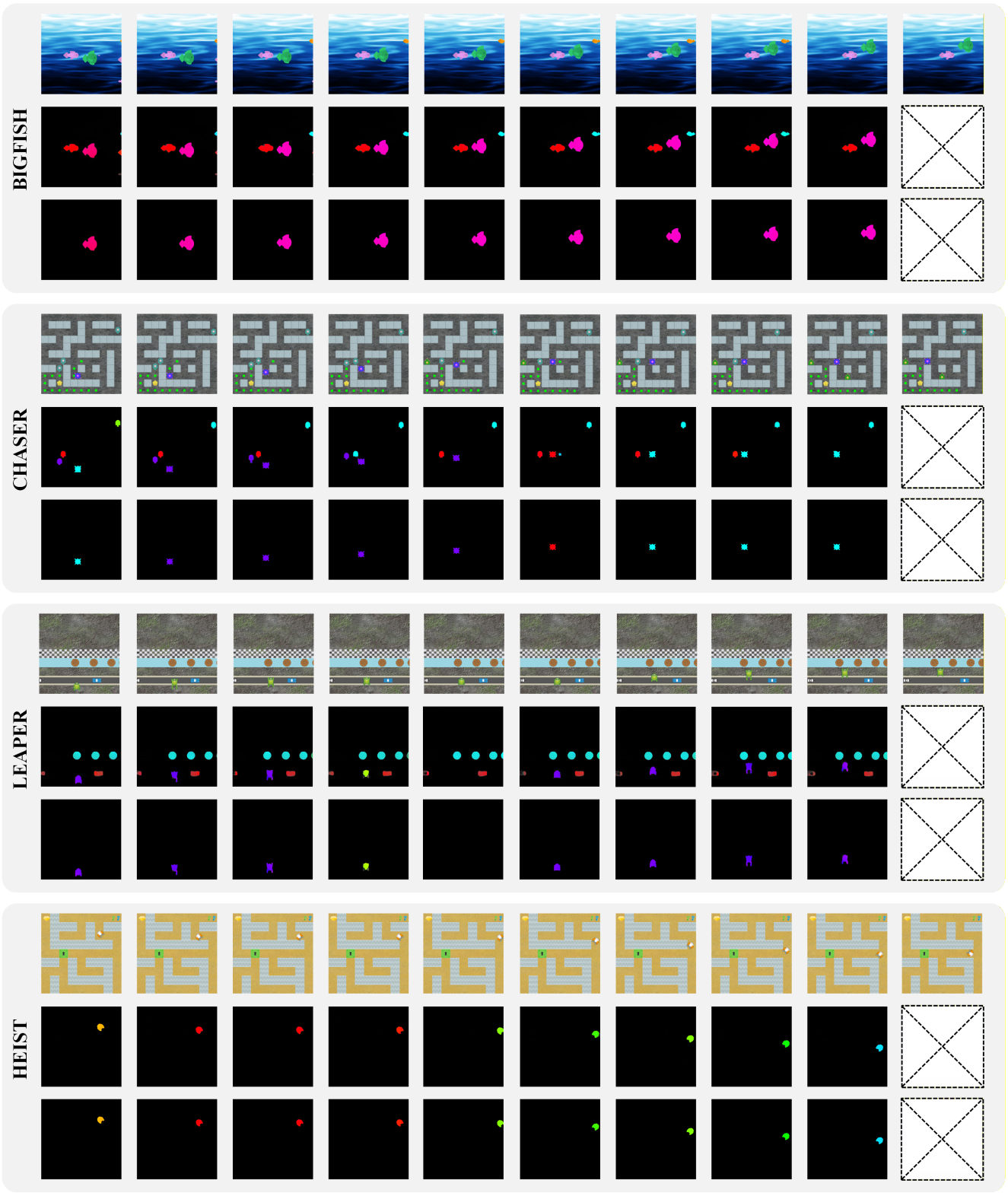


Figure 7. Visualization of optical flow for four tasks on the PROCGEN benchmark. For each task, 10 frames were sampled to illustrate the workflow, with each column representing a frame in temporal order. The first row shows the original images, while the second and third rows display the raw optical flow and the object-centric optical flow, respectively.

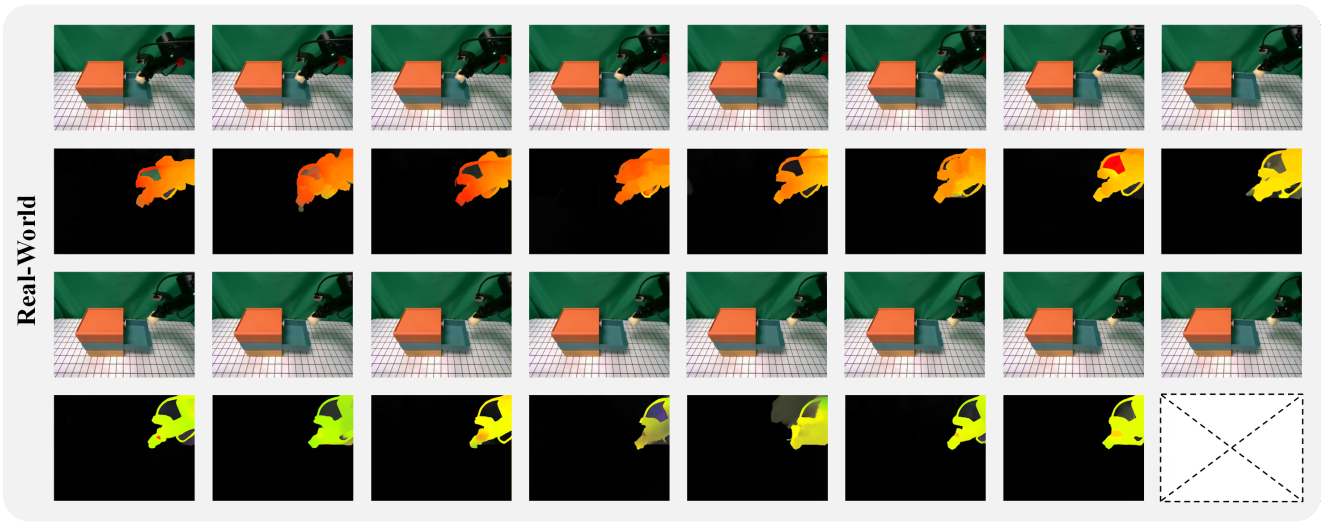


Figure 8. Real-World experiments were conducted on an ARX-R5 robot for a drawer-opening task, which involves opening the drawer, retrieving a sponge, and placing it on the table. Data were collected at 20 Hz over 100 trajectories (47,989 samples) across five different background settings, with each evaluation repeated 20 times.

Table 4. Canonical Correlation Analysis between optical flow representations and physical actions across different models. Optical flow features were first extracted using DINOv2 (768-dim) and then reduced to 50 dimensions via Principal Component Analysis (PCA) before computing CCA. Each benchmark reports results over 200K randomly drawn samples. Here, ΔDINOv2 denotes features obtained by differencing consecutive observations extracted using DINOv2. "Noise" denotes Gaussian noise, and "Action" denotes physical actions.

Method	LIBERO-Plus		PROCGEN		Real-World	
	Action	Noise	Action	Noise	Action	Noise
RAFT	0.67	0.03	**0.78**	0.03	0.61	0.06
SEA-RAFT	**0.68**	0.03	0.76	0.03	**0.65**	0.06
ΔDINOv2	**0.43**	0.03	**0.42**	0.03	**0.45**	0.06

Table 5. Experimental results of different optical flow models on LIBERO-Plus and Real-World. LIBERO-Plus [15] extends LIBERO by including robustness evaluations under distractors. The models were trained on mixture data (2M samples) and evaluated under Light conditions (variations in intensity, direction, color, and shadow) and Background conditions (changes in scene and surface appearance). Here, LAOF-R denotes optical flow extracted using RAFT, LAOF-S denotes SEA-RAFT, and LAOF-Δ denotes features obtained by differencing consecutive observations extracted using DINOv2

Method	LIBERO	LIBERO-Plus	PROCGEN	Real-World
LAPO	**72.6**	**43.4**	**0.535**	**14/20**
LAOF-R	76.8	63.5	**0.693**	**18/20**
LAOF-S	**78.4**	**68.3**	0.651	**18/20**
LAOF-Δ	73.7	57.7	0.544	16/20

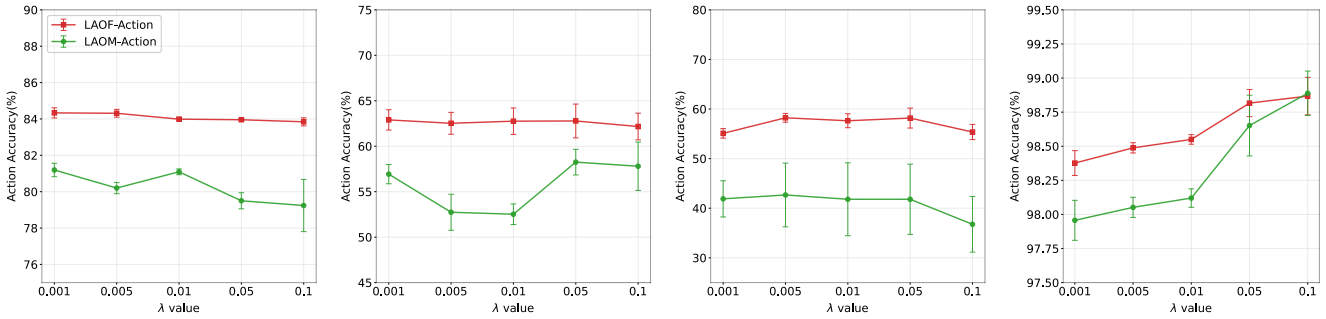
A.2. LIBERO-Plus and Real-World Experiments

As shown in Table 5, LAOF consistently improves robustness over LAPO, regardless of the specific optical flow model employed. This improvement holds across different environments and evaluation settings, indicating that incorporating optical flow constraints into latent action learning provides a systematic advantage under distribution shifts and visual perturbations. A comparison with Table 4 further reveals a clear relationship between motion–action alignment and downstream performance. On PROCGEN, RAFT achieves higher Canonical Correlation Analysis (CCA) correlations between optical flow representations and physical actions, which coincide with superior task performance. In contrast, on LIBERO-Plus, SEA-RAFT attains higher CCA correlations and correspondingly better overall results. Notably, even LAOF-Δ, which replaces explicit optical flow estimation with simple differencing of consecutive DINOv2 features, achieves competitive performance.

These results indicate that when motion representations more effectively capture action-relevant dynamics, the resulting latent action representations become more informative and inherently more robust to visual distractors.

A.3. Ablation Study on λ

The coefficient λ balances the contributions of action supervision and optical flow constraints. In our experiments, we adopt the default setting $\lambda = \frac{M}{N+M}$, reflecting the proportion of action-labeled data available during training. To eliminate the confounding influence of λ on assessing training stability, we conduct an ablation study under a fixed action ratio of 1%, varying the coefficient across roughly an order of magnitude, i.e., [0.001, 0.1]. We report the mean and standard deviation of action accuracy in Fig. 9 to assess its influence on the learning process. The results indicate that LAOF-Action remains highly stable across the entire range, exhibiting consistently lower variance and smoother performance changes. In contrast, LAOM-Action demonstrates substantially higher sensitivity, with notable performance fluctuations and clear signs of training instability. These findings show that LAOF-Action is significantly more robust to changes in λ and provides more reliable learning behavior under sparse action supervision, eliminating the need to learn λ. Future work could explore making λ a learnable parameter to potentially achieve optimal performance.



(a) BIGFISH (b) CHASER (c) LEAPER (d) HEIST

Figure 9. Effect of the coefficient λ on training stability. A shared legend for all four subfigures is shown in subfigure (a).

Table 6. Hyperparameters, dataset sizes, and learning rates for tasks (action ratio = 1%)

Task	σ	λ	Training Samples	Testing Samples	LangSAM Prompts	Learning Rate
SPATIAL	0.05	0.01	47,056	6,173	-	Stage1: 1e-4
OBJECT	0.05	0.01	59,860	7,449	-	Stage2: 3.5e-4
GOAL	0.05	0.01	46,444	6,451	-	Stage3: 3.5e-4
LONG	0.05	0.01	90,695	13,585	-	
BIGFISH	0.01	0.01	574,566	58,073	green	Stage1: 3e-4
CHASER	0.01	0.01	820,762	68,605	purple	Stage2: 2e-4
LEAPER	0.005	0.01	356,685	39,063	orange-white hair	Stage3: 3e-5
HEIST	0.01	0.01	342,590	36,741	green-frog	

B. Implementation Details

B.1. Task Setup

For each task, we adopt task-specific environment hyperparameters, as detailed in Table 6. Instead of using the full resolution normalization term $\sqrt{H^2 + W^2}$, we introduce a coefficient σ to avoid excessively attenuating subtle motions and to ensure sensitivity to small but meaningful movements. Since the magnitude of agent motions controlled by ground-truth actions differs across environments, σ is set accordingly. For LIBERO, the robotic arm control magnitudes are consistent across tasks, so the same σ is used. In contrast, for PROCGEN tasks such as LEAPER, the control dynamics differ from other games. The dataset is divided into separate training and testing splits. The model is trained solely on the training set, while the testing set is used exclusively for evaluation, where we report either the accuracy or the mean squared error of decoding

latent actions back into physical actions. Furthermore, to obtain high-quality optical flow, the initial data are collected at a resolution of 512×512, and the corresponding optical flow is extracted using RAFT [40]. For data from the PROCGEN benchmark, additional processing with LangSAM [19] is applied to obtain object-centric optical flow. Finally, for LIBERO and PROCGEN datasets, both the original images and the generated optical flow are resized to 256×256 for storage.

B.2. Model Setup

We use a frozen DINOv2 [33] encoder (ViT-B/14) to extract images features from normalized input frames. Task instructions are encoded using a frozen T5 [35] text encoder, which provides language embeddings that condition the downstream model components. The action decoder is implemented as a lightweight multilayer perceptron. For LIBERO tasks, the IDM is implemented as a spatio-temporal transformer [44], while both the FDM and the flow decoder adopt spatial transformers. For PROCGEN tasks, since we do not employ a multi-task model and each task must be trained independently, we use a simpler architecture to accelerate the extensive ablation studies: the IDM is implemented as a lightweight CNN encoder, and both the FDM and the flow decoder are directly implemented using a U-Net [36] architecture. To investigate different forms of latent action representations, we use a VQ-VAE [41] for the discrete setting, and a standard VAE for the continuous setting, with the KL-divergence term weighted by 1e-6.