Supplement

1. Additional Experimental Results

1.1. Added computational cost

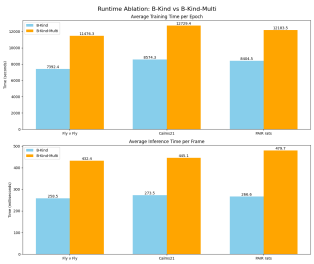

Figure 1. Computational Overhead comparison of B-Kind and B-Kind Multi on different datasets. The training time for B-Kind-Multi is only about 40% longer in both training and inference times.

1.2. Additional Ablations

We conducted several additional ablation studies to systematically evaluate the performance and sensitivity of our proposed method under various experimental conditions.

Video Segmentation Ablation. The first set of ablation studies focused on video segmentation techniques, comparing different approaches to extracting features from video sequences (Table 5).

Loss Function Ablation. We also performed a comprehensive loss function ablation to understand the contribution of different loss components to the model's learning process (Table 1).

Resolution Ablation. The final ablation explored the impact of image resolution on the model's capabilities (Table

2. Additional Implementation Details

Architecture Details. Our method uses ResNet-50 [**?**] as an encoder Φ, GlobalNet [**?**] as a pose decoder Ψ, and a series of convolution blocks as a reconstruction decoder ψ, following the unsupervised keypoint discovery model from [**?**]. Architecture details about reconstruction decoder is shown in Table 3.

Hyperparameters. The hyperparameters for the keypoint discovery model are listed in Table 4. Unless otherwise specified, all models use the SSIM image as the reconstruction target. Each keypoint discovery model is trained on an NVIDIA A100 Tensor Core GPU until the training loss converges.

Segmentation Parameters. Our method utilizes the DEVA framework [**?**] for video object segmentation, guided by text prompts. The parameters for DEVA segmentation are detailed in Table 6. In text-prompted mode, we use the following key parameters:

Recon.	Rot. Eq.	Sep.	%-MSE
		Fly v. Fly	
✓			$.298 \pm .067$
✓	✓		$.173 \pm .012$
✓		✓	$.221 \pm .059$
✓	✓	✓	$.115 \pm .032$
		CalMS21	
✓			$.896 \pm .021$
✓	✓		$.722 \pm .031$
✓		✓	$.764 \pm .055$
✓	✓	✓	$.693 \pm .061$
		PAIR-R24M	
✓			$.248 \pm .015$
✓	✓		$.192 \pm .037$
✓		✓	$.247 \pm .065$
✓	✓	✓	$.145 \pm .056$

Table 1. **Loss Ablation Results on Fly v. Fly, CalMS21, and PAIR-R24M**. "Recon." represents reconstruction loss, "Rot. Eq." represents rotation equivariance loss, and "Sep." represents separation loss. %-MSE error is reported for different combinations of loss functions. Results are from 5 runs.

Image Resolution	%-MSE
Fly v. Fly	
128x128	**.129** $\pm$ **.023**
256x256	**.115** $\pm$ **.032**
512x512	**.110** $\pm$ **.038**
CalMS21	
128x128	**.747** $\pm$ **.059**
256x256	**.693** $\pm$ **.061**
512x512	**.603** $\pm$ **.065**
PAIR-R24M	
128x128	**.180** $\pm$ **.031**
256x256	**.168** $\pm$ **.027**
512x512	**.145** $\pm$ **.056**

Table 2. **Image resolution ablation on Fly v. Fly, CalMS21, and PAIR-R24M**. %-MSE error is reported for different image resolutions. Results are from 5 B-KinD runs.

- **DINO_THRESHOLD**: This is the threshold for DINO to consider a detection as valid. A higher threshold would result in fewer but more confident detections, while a lower threshold might include more detections but with a risk of false positives.
- **prompt**: This is the text prompt used to guide the segmentation. The prompts are separated by a full stop; for example, "rat.mouse". The wording of the prompt and minor details like pluralization might affect the results. For instance, "cats" might yield different results from "cat".
- **size**: This is the internal processing resolution for the propagation module, which defaults to 480. A higher resolution might lead to more detailed segmentations but would require more computational resources.

All other parameters were kept at their default values.

Type	Input dimension	Output dimension	Output size
Upsampling	-	-	16x16
Conv_block	2048 + # keypoints $\times$ 2 $\times$ agents	1024	16x16
Upsampling	-	-	32x32
Conv_block	1024 + # keypoints $\times$ 2 $\times$ agents	512	32x32
Upsampling	-	-	64x64
Conv_block	512 + # keypoints $\times$ 2 $\times$ agents	256	64x64
Upsampling	-	-	128x128
Conv_block	256 + # keypoints $\times$ 2 $\times$ agents	128	128x128
Upsampling	-	-	256x256
Conv_block	128 + # keypoints $\times$ 2 $\times$ agents	64	256x256
Convolution	64	3	256x256

Table 3. **Architecture details of the reconstruction decoder.** "Conv_block" refers to a basic convolution block which consists of 3$\times$3 convolution, batch normalization, and ReLU activation.

Dataset	# Keypoints	Batch size	Resolution	Frame Gap	Learning Rate	Agents
CalMS21	10	5	256	6	0.001	2
Fly vs. Fly	10	5	256	3	0.001	2
PAIR-R24M	12	5	256	6	0.001	2

Table 4. **Hyperparameters for Keypoint Discovery.**

Video Segmentation Type	%-MSE
CalMS21	
Frame by Frame	**3.454** $\pm$ **.167**
DEVA	**.693** $\pm$ **.061**
Fly v. Fly	
Frame by Frame	**2.531** $\pm$ **.097**
DEVA	**.115** $\pm$ **.032**
PAIR-R24M	
Frame by Frame	**2.874** $\pm$ **.121**
DEVA	**.168** $\pm$ **.027**

Table 5. **Video segmentation ablation on CalMS21, Fly, and PAIRS**. %-MSE error is reported by changing the segmentation type. Extracted features correspond to keypoint locations, confidence, and covariance. Results are from 5 B-KinD runs.

Dataset	Grounded Segmentation Prompt	DINO Threshold	Size
CalMS21	rat.mouse	0.45	480
Fly vs. Fly	fly	0.5	480
PAIR-R24M	rat.mouse	0.45	480

Table 6. **Parameters for Segmentation.**