Future Optical Flow Prediction Improves Robot Control & Video Generation

Kanchana Ranasinghe[⋆,1,2], Honglu Zhou[1], Yu Fang[⋆,1], Luyu Yang[1], Le Xue[1],
Ran Xu[1], Caiming Xiong[1], Silvio Savarese[1], Michael S Ryoo[1,2], Juan Carlos Niebles[1]

[1]Salesforce AI Research [2]Stony Brook University

FOFPred.github.io

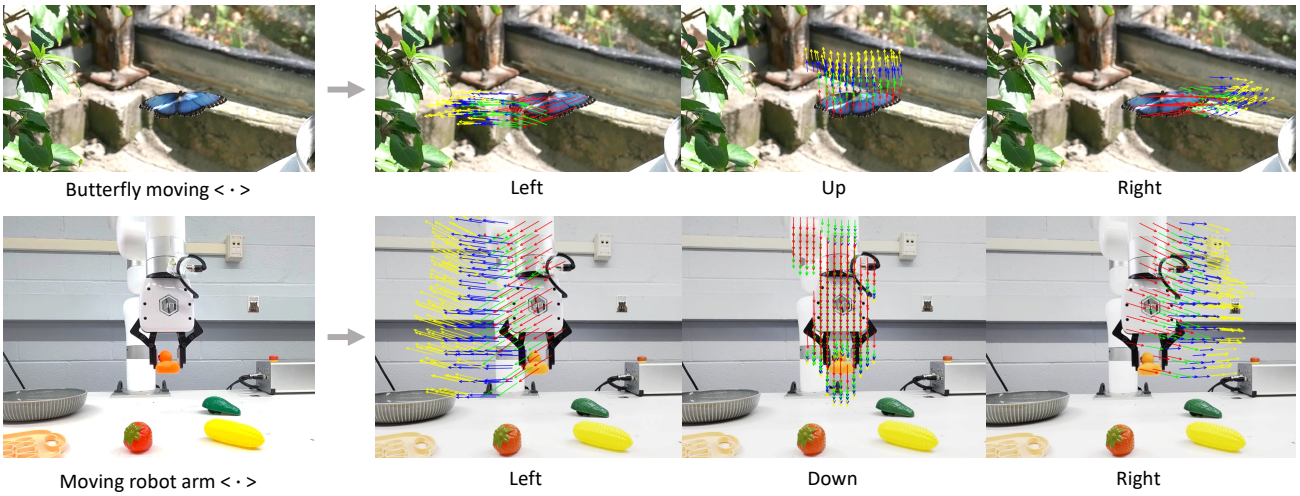

Figure 1. **Our proposed model, FOFPred, learns generalizable *language-driven future optical flow prediction* that supports robot control and video generation.** $< \cdot >$ is a placeholder for text (e.g., Left/Up/Right). For visualization purposes, we sample points from the predicted future optical flow and show their predicted trajectories in this figure.

Abstract

Future motion representations, such as optical flow, offer immense value for control and generative tasks. However, forecasting generalizable spatially dense motion representations remains a key challenge, and learning such forecasting from noisy, real-world data remains relatively unexplored. We introduce FOFPred, a novel language-conditioned optical flow forecasting model featuring a unified Vision-Language Model (VLM) and Diffusion architecture. This unique combination enables strong multimodal reasoning with pixel-level generative fidelity for future motion prediction. Our model is trained on web-scale human activity data—a highly scalable but unstructured source. To extract meaningful signals from this noisy video-caption data, we employ crucial data preprocessing techniques and our unified architecture with strong image pretraining. The resulting trained model is then extended to tackle two distinct downstream tasks in control and generation. Evaluations across robotic manipulation and video generation under language-driven settings establish the cross-domain versatility of FOFPred, confirming the value of a unified VLM-Diffusion architecture and scalable learning from diverse web data for future optical flow prediction.

1. Introduction

Recent studies have investigated the use of motion representations such as optical flow [69, 99], sparse motion trajectories [8, 92, 105], and direction commands [13, 91, 103] for multiple control and generation tasks. The main idea behind such motion representations, often in the form of *future* pixel movements like optical flows and trajectories, is to explicitly capture desired dynamics necessary for downstream tasks such as robot control and video generation.

In the case of robot control, models with such future motion representations have the capability to explicitly con-

sider future pixel displacement information to infer robot actions [8, 69, 92, 99, 105]. This is in contrast to the conventional Vision-Language-Action (VLA) models taking only RGB-frames and language instructions as input. Similarly, video diffusion models utilizing motion representations enables generation of videos with consistent and detailed movements, and it does it better than the counterparts utilizing language-only instructions [15, 24, 42, 53].

The key is in reliable and generalizable computation of such motion representation, which often involves *future forecasting*. What has to be provided to the models for many (robot) control and (video) generation tasks is how things should move in the future, not what motion has been observed in the past. This inherently is a challenging task: it requires a clear formulation of what type of 'future motion' we will be computing, curating large enough training data with strong supervisory signals, and an elegant model capable of fully digesting such (potentially cross-domain) training data while generalizing across domains.

In this paper, we formulate the problem as the forecasting of future *optical flows* conditioned on language instructions, and present a new model architecture to forecast them. We also discuss how we can train our model with videos from various sources, ranging from human videos to robot videos. Optical flows are spatially dense motion representations, capturing movements of every pixel in the scene. This better preserves motion details compared to sparse representation, and enables easy use of diffusion models to learn their distributions in the form of images. In contrast to predicting RGB frame sequences, which also capture future motion [35], optical flow eliminates static information irrelevant to motion, leading to more motion-centric representations [69].

We introduce an optical flow forecasting model unifying both Vision-Language Models (VLM) and Diffusion. Fully taking advantage of VLM reasoning capability and pre-trained diffusion model image generation ability, we train our VLM-Diffusion model to generate future optical flow images. Our model is trained on web-scale human activity videos with paired captions allowing highly-scalable training. The resulting model, named FOFPred, enables highly generalizable language-driven future optical flow prediction as illustrated in Figure 1.

While learning from web videos provides this strong generality, such data is highly unstructured with noisy videos and captions. Most prior work learning future motion representations either avoid training with such data [8, 69, 92, 99, 105] or utilize RGB frame prediction [35, 40]. In [100] where such training data is used, explicit camera motion correction is used to handle noisy videos and a VLM backbone enables learning from diverse noisy captions. However, the use of only a VLM restricts their future motion prediction to sparse pixel trajectories. In our work, we calculate dense relative optical flow compensating for the camera motion and model this with diffusion, which improves our learning from noisy web videos.

Finally, we validate the effectiveness and generality of FOFPred through evaluation on two challenging downstream tasks: language-driven robotic manipulation and language-guided motion video generation. For each task, we attach a diffusion policy head or a video diffusion head on top of FOFPred and fine-tune the two heads separately for the respective downstream domains.

We summarize our key contributions as follows:

- **Unified VLM-Diffusion:** Adopting this recent architecture for generalizable future optical flow prediction.
- **Scalable Learning:** Establish framework for learning future optical flow prediction from web-scale human activity videos.
- **Cross-Domain Versatility:** To the best of our knowledge, the first to utilize VLM–diffusion backbone for robotic manipulation and controlled video generation under language-driven settings.

2. Related Work

Unified Model for Control Generation. Recent progress in foundation models has demonstrated a powerful trend toward unified model architectures that integrate diverse modalities and achieve highly controlled visual input manipulation. These generalist models [6, 17, 30, 37, 63, 64, 94, 97, 100], often building on transformers and large-scale pre-training (e.g., Gato [70]), establish single frameworks capable of handling hundreds of tasks. In the generative domain, this unification enables advanced control: models like UniDiffuser [6] tackle various image generation and editing tasks within a single conditional framework, while BLIP-3o [17], OmniGen-2 [94], and MetaQuery [64] demonstrate sophisticated language-conditioned iterative modification and transfer across modalities. Crucially, the development of plug-and-play architectures like ControlNet [108] established the blueprint for adding flexible, low-level spatial controls to pretrained diffusion models. This pioneering work inspired recent unified generative backbones that accept diverse conditioning signals, enabling versatile controllable generation [21, 30, 36–38, 43, 55, 81]. This foundation of multi-modal input, unified architecture, and flexible control is a key enabler for tackling complex tasks like controlled video [66, 67, 84, 88] and robotic action [51, 57, 68, 87] generation.

Motion Representations in Robot Control. Learning motion related information from videos has been widely explored in robot learning [3, 16, 22, 23, 35, 40, 44, 48, 61, 65, 71, 74, 75, 78, 80, 82]. Building on this, the field has increasingly explored vision-language-action (VLA)

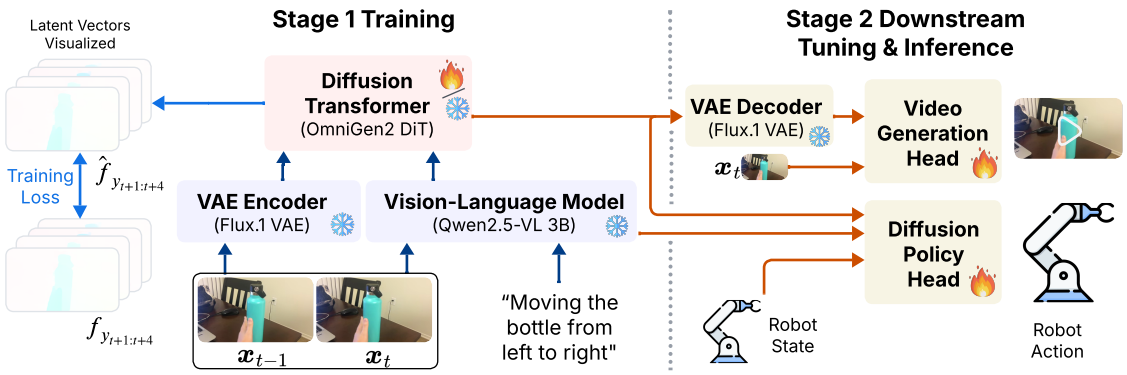

Figure 2. **Overview of Proposed FOFPred:** (Left & Center) We present the unified VLM-Diffusion architecture used in FOFPred. Only the DiT module is trained while the VAE and VLM remain frozen. (Right) We illustrate two distinct pipelines constructed with FOFPred for two orthogonal tasks in control and generation. Each task specific head is first finetuned prior to inference on the downstream task.

models to learn policies conditioned on natural language [22, 35, 40, 80], although many of these still rely on extensive action-trajectory annotations or task-specific heuristics. Recent foundation models aim to overcome this by better integrating spatio-temporal intelligence: MAGMA [100] utilizes "Trace-of-Mark" annotations on video movements for action grounding and planning across digital and physical tasks. Other approaches focus on improving world model reasoning before action generation: FlowVLA [112] employs a "Visual Chain of Thought" to explicitly predict optical flow before the next visual frame, disentangling motion dynamics from static appearance for more efficient policy learning, while DreamVLA [109] forecasts a compact set of crucial "world knowledge", including dynamic regions, depth, and semantics, to guide action planning and establish a robust perception-prediction-action (PPA) loop. Independently, optical flow has served as a primary representation in computer vision for motion learning [26, 31, 59, 76, 90], and its extension to trajectories of pixel subsets [8, 92, 99, 105] has found utility in robot control. However, methods relying on localized trajectories [14, 45, 79, 104] often limit their focus to specific object movements, overlooking crucial global information, such as the overall movement of a manipulator. In contrast, our FOFPred predicts spatially dense future optical flow leveraging powerful unified model architectures.

Motion Control in Video Generation. Controlling motion in video generation using spatial conditioning signals has been widely explored [13, 19, 27, 32, 49, 52, 62, 77, 89, 91, 96, 103, 107, 110, 113]. Some methods define object motion through sparse user-provided trajectories and points, including DragNUWA [103], DragAnything [96], Tora [107], TrackGo [113], and the foundational Controllable Video Generation [32]. Others employ explicit motion models, such as MotionCtrl [91], Motion-I2V [77], and Motion-Conditioned Diffusion Models [19]. Additional techniques achieve control through motion fields (MOFA-Video [62]), image-based precision guidance (Image Conductor [52]), compositional video synthesis (VideoComposer [89]), or generalized animation (AnimateAnything [49]). However, language-based explicit motion control remains underexplored, which we focus on in this work.

3. Method

Optical Flow, the apparent displacement of pixels between frame pairs, has a long history in computer vision [7, 34]. Estimating the true optical flow given a frame pair is a well-established problem [83]. On the other hand, several recent works explore a different task of estimating future optical flow from a *single frame* [2, 25, 85], with some conditioned on language [69, 99]. To distinguish this task of future pixel displacement estimation from a single frame, we use the term future optical flow in the following sections.

First, we introduce FOFPred, our framework that generates future optical flow, given a language instruction and one or more images. This future optical flow generation is learned from video-caption pairs without any pixel-level human annotation. We train a unified LLM-diffusion architecture based model to generate sequences of future optical flow frames. In the following sections, we first detail the architecture of our framework, followed by training and inference procedures. Next we describe our extension for robot control, a task requiring strong awareness of object and ego motions. Finally we present our second extension for motion control in text-to-video generation.

3.1. Architecture

Motivated by recent success of unified model architectures in language conditioned visual generation [17, 64, 94, 97], we construct FOFPred that maps frame sequences and language into future optical flow sequences. We illustrate an architectural overview of FOFPred in Figure 2.

Consider observations (images) $\boldsymbol{x}_t \in \mathbb{R}^{h \times w \times c}$ corresponding to time step t and paired natural language cap-

tion c. We first utilize an autoregressive transformer based vision-language model as a textual feature encoder that encodes both language caption c & visual input $\boldsymbol{x}_{t-1}, \boldsymbol{x}_t$ to obtain $\boldsymbol{f}_c$. We next encode visual inputs, $\boldsymbol{x}_{t-1}, \boldsymbol{x}_t$ with a VAE encoder to obtain $\boldsymbol{f}_v$. The textual ($\boldsymbol{f}_c$) & visual ($\boldsymbol{f}_v$) features are passed through MLP layers and input to our diffusion transformer (DiT). Outputs of the DiT, $\hat{\boldsymbol{f}}_y$ are passed through a VAE decoder to obtain $\hat{\boldsymbol{y}}$ sequences which correspond to future optical flow sequences.

For our DiT architecture, we utilize the OmniGen transformer [94], introducing several modifications to enable temporal modeling, given the time dimension of our framework's inputs and outputs. We first modify the DiT 2D RoPE encoding to handle both input and output frame sequences. Next we update the DiT transformer blocks to perform full spatio-temporal attention in order to model the temporal axes in frame sequences. This design introduces no additional learnable parameters, enabling our framework to directly benefit from image domain pretraining used in [94]. In terms of other components, we use Qwen2.5-VL [4] as our VLM and Flux.1 VAE [46] as our visual encoder-decoder. The two MLP layers ensure channel dimension of both textual and visual features are common. These conditional inputs are simply appended to the DiT input sequence. Further details in Appendix A.

3.2. Optical Flow Representation

In contrast to prior future optical flow generation works [69, 99], we adopt an RGB space representation for optical flow (OF). This allows us to directly benefit from existing powerful VAE models to encode our OF, in contrast to training or finetuning the VAE model [69, 99]. Motivated by [15, 112], we explore representing OF in RGB formats by mapping the polar coordinate values of OF to the HSV color space. We map magnitude, rotation, and a linear combination of these two from the polar representation into the three H-S-V channels respectively. We tune the scaling values for these mappings to maximize visual continuity across frames (e.g. no flickering) and to resemble RGB images (e.g. animated graphics). See Appendix B for further details.

3.3. Training

Our FOFPred framework is trained end-to-end to predict future optical flow sequences conditioned on image sequences and paired textual captions. During training, the target optical flow sequence, $\boldsymbol{y}$, is calculated using a suitable optical flow calculation algorithm, $\mathcal{F}$, which can access the future frames for the calculation. We use $\boldsymbol{y}[i] = \mathcal{F}(\boldsymbol{x}_i, \boldsymbol{x}_{i+1})$ where $\cdot[i]$ indexes the i-th frame of each sequence. These optical flow targets are encoded with the VAE encoder to obtain $\boldsymbol{f}_y$ which is in turn used with a flow matching diffusion loss as the training objective.

Let us define the learnable components of our framework, i.e. the two MLPs and DiT, as $\mathcal{D}$ such that $\mathcal{D}(\boldsymbol{f}_c, \boldsymbol{f}_v, \boldsymbol{f}_0) \rightarrow \hat{\boldsymbol{f}}_y$ where $\boldsymbol{f}_0$ is the noise sample and $\hat{\boldsymbol{f}}_y$ is predicted tensor output of the DiT module. Given targets $\boldsymbol{f}_y$ and interpolated feature $\boldsymbol{f}_y^t = (1-t)\boldsymbol{f}_0 + t \cdot \boldsymbol{f}_y$ for $t \sim \mathcal{U}(0,1)$, we obtain our training objective as,

$$\mathcal{L}_{\text{FM}}(\theta) = \|\{\mathcal{D}(\boldsymbol{f}_c, \boldsymbol{f}_v, \boldsymbol{f}_y^t) - (\boldsymbol{f}_y - \boldsymbol{f}_0)\|_2^2 \qquad (1)$$

where θ are our learnable parameters. During training, we perform classifier-free guidance dropping the textual condition ($\boldsymbol{f}_c$) or the visual condition ($\boldsymbol{f}_v$). We also partially mask our visual condition ($\boldsymbol{f}_v$) along time (e.g. $\boldsymbol{x}_{t-1}$ is masked) and viewpoint axes. Note that our visual observation, $\boldsymbol{x}_t$, may contain images corresponding to multiple viewpoints, hence the viewpoint masking.

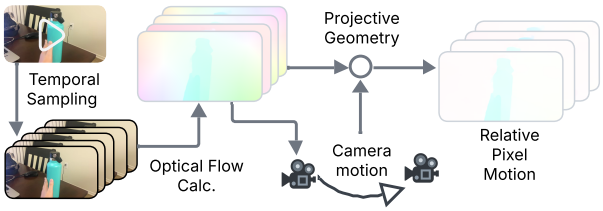

Figure 3. **Relative Optical Flow Calculation:** We illustrate the key stages of the algorithm for calculating optical flow targets for our training.

Relative Optical Flow Calculation: We construct our framework to train from internet videos of human activities (e.g. taken from a mobile phone), where each video may contain both camera (i.e. ego) motion and object motions. We construct our optical flow calculation algorithm, $\mathcal{F}$, accounting for these possible structures of training data. Given a current-future frame pair, we first use an off-the-shelf optical flow calculation algorithm [83] to calculate flow vectors. Next, motivated by prior work [100], we use homography techniques with deep features [47] to estimate the camera motion. Then we utilize the camera motion estimates with projective geometry to calculate flow vectors relative to the camera. Object motions are disentangled from camera motions in these relative optical flow vectors. Further details on this calculation in Appendix C. We also note how motion in natural videos is not uniformly distributed. To tackle this, given optical flow between frame pairs, we take the top-k percent of flow values as a proxy to motion between those frame pairs. We select frame sequences such that motion between consecutive pairs exceeds a specific threshold. The final optical flow targets used to train our framework are calculated for these filtered frame sequences. Further details of our frame selection in Appendix D. Given the computational cost of our data preprocessing, initial frame selection utilizes a fast optical calculation [58] on low spatial resolution frames. Relative optical flow calculation is performed on original resolution images. Both steps are performed offline as one time processes, resulting in similar

sampling costs as uniform frame sampling during training.

3.4. Inference

Our VLM and VAE encoders act as feature extractors, followed by the MLP layers projecting these features to a common dimension. The DiT module uses these features as conditional inputs to perform k iterations of reverse diffusion. The value k is varied based on the downstream application. For the DiT, we perform classifier-free guidance on the text and visual features.

3.5. Downstream Use Case I: Robot Control

We first validate the usefulness of our FOFPred framework through language driven robot control or vision-language-action (VLA). These tasks require robust motion awareness and precise language-based control. Motivated by prior work [35, 69], we integrate the future optical flow predicted by FOFPred into a diffusion policy network (DPN) [72]. This DPN processes the predicted future optical flow alongside textual inputs and current state information, mapping these multimodal inputs directly into action vectors that control a robotic manipulator. We provide an overview of our robot control pipeline in Figure 2.

Given the embodiment awareness necessary for robotic manipulation, we finetune our framework for downstream robotic tasks. Aligning with established methodologies [35, 69], this process comprises two stages: (1) future optical flow prediction finetuning on robotic video-caption data, and (2) action head training on robotic demonstration data. A key distinguishing feature of our work is the future optical flow finetuning phase, which explicitly accounts for the two types of viewpoints, fixed external and moving wrist cameras, inherent in robotic data. We achieve this by ensuring cross-view conditioning within the DiT backbone and augmenting our prediction targets ($\mathbf{f}_y$) to encompass optical flow from both viewpoint types. These architectural and data handling innovations differentiate our framework from previous robot control approaches [35, 69]. For final robot control, the predicted motions are mapped to robotic actions with the diffusion policy network from [35] modified to condition on our future optical flow predictions.

3.6. Downstream Use Case II: Video Generation

We further demonstrate the versatility of FOFPred framework by exploring Text-to-Video (T2V) generation focused on purely natural language-based motion control. While prior motion-guided video generation methods typically rely on low-level visual inputs like pixel-level trajectories or motion masks generated by humans through suitable software interfaces [13, 96, 103], we investigate the use of textual descriptions as a direct, human-centric motion control signal. As illustrated in Figure 2, we establish a two-stage pipeline connecting FOFPred with video generation head, Go-with-the-Flow (GWTF) [13]. First, FOFPred takes a static visual observation (the initial frame) and a textual description of the desired future motion as input. Conditioned on these inputs, FOFPred generates a sequence of future optical flow frames. This sequence is subsequently interpolated to create a dense motion signal. In the second stage, this dense motion and the initial frame are passed to the video generation head (GWTF), which synthesizes the final video sequence, ensuring the output realizes the motion pattern defined in the original text. Compared to single-stage T2V methods [41, 86, 101], our approach is computationally more intensive but offers two distinct advantages. It demonstrates superior fidelity to complex textual motion instructions, and the interpretable future optical flow sequence provides a crucial, intermediate layer of transparency to the generation process. We also highlight this as an experiment to validate the strength of language-driven future optical flow forecasting ability learned by our framework.

4. Experiments

In this section, we first detail our experimental setup followed by evaluations on language-driven robotics manipulation and motion focused text-to-video generation. Finally we provide ablations of our FOFPred framework to validate our design choices.

Implementation Details: Our FOFPred framework contains three key components: VLM, VAE, and DiT. We use the Qwen2.5-VL [4] architecture for our VLM and the Flux.1 [46] architecture for our VAE. For both models, we initialize with their official pretrained weights. Both the VLM and VAE weights are kept frozen. Our DiT uses a standard diffusion transformer architecture following OmniGen [94] with modifications for video generation as described in Section 3.1. We initialize the DiT (including projector layer MLP) with weights from a model pretrained on image editing tasks from [94].

Training Details: We train our FOFPred framework on the human activity data from the train split of Something-Something-V2 dataset [28] and the EgoDex dataset [33]. Our training data contains roughly 500,000 video-caption pairs. We train our model on 8xH200 GPUs with a global batch size of 4096 for 100 epochs. We note how different frame subsequences are sampled from the same video across epochs based on motion based sampling as explained in Section 3.3. For robotic manipulation, we additionally finetune our model on all available video data (without access to action trajectory labels) on the downstream tasks following prior robotic manipulation works [35, 69, 109].

Method	Training Data	i^{th} Task Success Rate ↑					Avg. Len ↑
		1	2	3	4	5	
RT-1 [12]	100% ABC	53.3	22.2	9.4	3.8	1.3	0.90
Diffusion Policy [20]	100% ABC	40.2	12.3	2.6	0.8	0.0	0.56
Robo-Flamingo [50]	100% ABC	82.4	61.9	46.6	33.1	23.5	2.47
Uni-Pi [22]	100% ABC	56.0	16.0	8.0	8.0	4.0	0.92
MDT [72]	100% ABC	63.1	42.9	24.7	15.1	9.1	1.55
Susie [9]	100% ABC	87.0	69.0	49.0	38.0	26.0	2.69
GR-1 [95]	100% ABC	85.4	71.2	59.6	49.7	40.1	3.06
Vidman [93]	100% ABC	91.5	76.4	68.2	59.2	46.7	3.42
RoboUniview [54]	100% ABC	94.2	84.2	73.4	62.2	50.7	3.65
LTM [69]	100% ABC	97.1	82.4	72.8	67.2	60.6	3.81
DreamVLA [109]	100% ABC	98.2	94.6	89.5	83.4	78.1	4.44
VPP [35]	100% ABC	96.5	90.9	86.6	82.0	76.9	4.33
FOFPred	100% ABC	98.8 (+2.3)	95.0 (+4.1)	90.4 (+3.8)	84.6 (+2.6)	78.7 (+1.8)	4.48 (+0.15)
GR-1 [95]	10% ABC	67.2	37.1	19.8	10.8	06.9	1.41
VPP [35]	10% ABC	87.8	74.6	63.2	54.0	45.3	3.25
FOFPred	10% ABC	90.4 (+2.6)	77.4 (+2.8)	65.9 (+2.7)	62.8 (+8.8)	46.9 (+1.6)	3.43 (+0.18)

Table 1. **CALVIN Evaluation:** Zero-shot long-horizon evaluation on the Calvin ABC→D benchmark where agent is asked to complete five chained tasks sequentially based on instructions. We report success rate (%) for each sequential task and the average length of task completion following standard protocol. Improvements over VPP [35] baseline reported for our method.

Method	Handover Block	Handover Mic	Pick Diverse Bottles	Pick Dual Bottles	Place Dual Shoes	Average
RDT [56]	45	90	2	42	4	36.6
ACT [111]	42	85	7	31	9	34.8
DP [20]	10	53	6	24	8	20.2
DP3 [106]	70	100	52	60	13	59.0
π_0 [10]	45	98	27	57	15	48.4
VPP [35]	54	80	60	63	52	61.8
FOFPred (ours)	61 (+7)	87 (+7)	67 (+7)	68 (+5)	60 (+8)	68.6 (+6.8)

Table 2. **RoboTwin Evaluation:** We report success rates on the RoboTwin 2.0 benchmark following their official protocol on the easy setting. This environment contains a bimanual robot. The five tasks are selected based on necessity of both arms to complete the task. Our proposed FOFPred achieves consistent improvements over the VPP baseline implemented under identical settings.

4.1. Language-Driven Robot Manipulation

We evaluate on two robot manipulation benchmarks, CALVIN [60] and RoboTwin 2.0 [18]. The CALVIN benchmark is an open-source, simulated environment designed for long-horizon, language-conditioned tasks. It challenges agents to perform complex, multi-step manipulation sequences using continuous control, with task goal signals specified only via unconstrained natural language instructions. This dataset focuses on developing policies that can generalize to novel language instructions and unseen environments, making it an ideal evaluation task for our framework. For this task, we utilize the CALVIN ABC→D setting, where training data is available only for ABC environments, while evaluation is performed zero-shot on the novel D environment. Given the multi-step sequential manipulation tasks within this benchmark, the standard evaluation metrics are i^{th} task success rate and the average length of task stages completed. We report this standard metrics from the benchmark used across prior work [35, 69] for our evaluations. The RoboTwin 2.0 framework provides a scalable approach for bimanual (dual-arm) robotic manipulation, specifically addressing the limitations of oversimplified simulations for two-arm tasks. We select 5 tasks that specifically require utilizing both arms for manipulation in order to successfully complete the task. This benchmark allows evaluating the generality of our framework across both single and dual arm robotic environments. The standard evaluation metric for this benchmark is success rate which we use in our evaluations, along with the average success rate across all tasks.

CALVIN Results: Our results reported in Table 1 demonstrate state-of-the-art performance of our FOFPred framework in long-horizon, zero-shot robot manipulation. All evaluations follow standard protocol [60]. Results for baselines are reported directly from prior work [35, 109]. Training on the full split for the ABC→D setting, our FOF-

Method	SSIM ↑	PSNR ↑	LPIPS ↓	FVD↓	KVD↓	MF ↑ [102]
Seer [29]	41.8	10.71	58.8	287.46	81.31	—
Dynamicrafter [98]	—	—	—	204.11	31.81	—
CosHand-I [80]	61.5	16.87	31.3	91.18	19.24	0.432
CosHand-A [80]	53.1	14.92	40.8	90.30	13.68	0.570
InterDyn [1]	66.4	18.60	26.0	19.27	1.99	0.633
InterDyn-R [1]	68.0	19.04	25.2	22.22	2.09	0.641
CogVideoX [101]	67.2	21.51	30.3	78.47	12.46	0.594
FOFPred (ours)	$68.4_{(+1.2)}$	$22.26_{(+0.75)}$	$28.5_{(+1.8)}$	$75.39_{(+3.08)}$	$11.38_{(+1.08)}$	$0.662_{(+0.068)}$

Table 3. **Evaluation of Language-Driven Motion Control in T2V:** We evaluate the ability of our T2V pipeline for language-driven motion control using the motion heavy SSv2 dataset following prior work [1]. Over the CogVideoX baseline, our FOFPred framework shows consistent improvements in generation quality.

Pred framework registers the highest success rate across all five chained tasks, culminating in a Task 5 Success Rate of 0.787 and the highest Average Length of task completion at 4.48, marginally surpassing the prior best model, DreamVLA (4.44). This strong performance confirms our method's superior ability to generalize to novel language instructions for sequential, multi-step manipulation. Furthermore, in the data-limited regime (using only 10% of the ABC dataset), our method continues to excel, achieving an average length of 3.43, which is noticeably higher than other models in this setting, highlighting its data efficiency. Similar trends of data efficiency have been observed in prior works utilizing motion representations [25, 69], another benefit of our framework.

RoboTwin Results: We report RoboTwin results in Table 2, demonstrating the effectiveness of our FOFPred framework in complex bimanual manipulation tasks. Results for baselines are directly from the official leaderboard [18], with attempted replication following the official codebase. We note that we were unable to replicate the baseline results reported in leaderboard for some tasks (e.g. Handover Mic), but nevertheless report official values from the benchmark leaderboard. We replicate the VPP baseline [35] following their official codebase training under identical settings as our FOFPred framework, enabling a direct and fair point of comparison. On our selected task subset, our model achieves an average success rate of 68.6%, significantly outperforming our baseline model, VPP (61.8%), which is trained similarly on web human videos, but with frame prediction instead of motion prediction. Our FOFPred outperforms this baseline consistently across all tasks, demonstrating the strength of motion prediction over frame prediction. In comparison to other baselines, our model notably shows robust performance across tasks (e.g. Place Dual Shoes), highlighting its better generalization ability. We take these results as confirmation on the utility of FOFPred framework for solving complex robotic tasks requiring cooperative bimanual manipulation.

4.2. Motion Control in Video Generation

Our second downstream task focuses on text-to-video generation with focus on language-driven motion control aspects. Table 3 presents results for our evaluation on the validation split of SSv2 dataset following prior work [1]. Our evaluation protocol is identical to [1] with results for baselines [1, 29, 80, 98] reported directly from [1]. We implement and evaluate the CogVideoX baseline [101] under identical settings as our FOFPred framework. We achieve consistent improvements over the CogVideoX baseline across all major metrics. Notably, while most prior controllable-generation baselines (e.g., [1, 80]) rely on auxiliary control signals such as hand or object masks over time—inputs that are not freely available at inference, our framework operates solely from language prompts and visual context. Despite this weaker supervision at test-time, FOFPred attains comparable or superior performance, underscoring its capacity to learn text-conditioned motion representations that generalize without explicit spatiotemporal guidance. We take these results as indication to how motion-aware language grounding alone can yield coherent, controllable dynamics in text-to-video generation.

4.3. Ablations

Next we analyze contributions of each design choice. For all experiments in this section, we use only the SSv2 dataset as the pretraining source for faster experimentation. This dataset provides a broad range of human activity videos with rich motion diversity, making it a suitable testbed for controlled comparisons. For evaluation, we follow the standard CALVIN setup [35, 60], where each evaluation sequence consists of five chained tasks. While our main experiments utilize the full 1000-sequence benchmark to ensure comprehensive assessment, the ablations are performed on a reduced subset of 400 sequences to facilitate efficient experimentation without compromising representativeness. The following analyses dissect three key factors: overall framework, human web video pretraining, and motion disentangling strategy. We evaluate each for its impact

Method	AL ↑
Ours	4.39
No DiT temporal adaptation	4.05 (-0.34)
Replace VLM w/ Sentence-T5	3.95 (-0.44)
No web-video pretraining	3.91 (-0.48)
No image-edit pretraining	4.14 (-0.25)
Stable Diffusion-XL (SD-XL)	4.01 (-0.38)
Stable-Video Diffusion (SVD)	4.10 (-0.29)
CogVideoX	4.31 (-0.08)

Table 4. **Ablations:** We highlight usefulness of our architecture & pretraining design choices, and how tuning alternative Image / Video DiTs did not outperform our modified VLM-diffusion backbone. CALVIN average length (AL) is reported.

on motion guidance prediction through downstream task performance.

Backbone Ablation: Table 4 establishes contribution of our DiT modifications, the VLM component in our architecture, and the pretraining stages. The image diffusion-only SD-XL [73] baseline achieves lower performance compared to our VLM-diffusion architecture [94], indicating its ability to model low-level motion but limited understanding of linguistic context and multimodal reasoning. Alternate video DiTs (SVD [11], CogVideoX [101]) result in reduced robot performance and lower-quality OF (see Appendix E), indicating that language-driven OF prediction requires strong visual-language understanding, text-aligned motion geometry & spatial precision over appearance/texture consistency that video DiTs provide. Our pretraining leverages superior language-grounded spatial priors from high-resolution image editing pretraining, while our DiT modifications effectively create a minimal video-DiT, sufficient for modeling short-horizon OF sequences. Together, these results validate our architectural design: combining VLM and diffusion components benefits generalizable learning of motion guidance prediction from weakly aligned web captions.

Pretraining Ablation: Table 5 investigates the effect of pretraining on large-scale web videos of human activities compared to teleoperated robotic demonstrations by using the full DROID dataset [39]. While DROID provides diverse in-the-wild robot manipulation demonstrations, it remains limited in coverage and motion variability relative to human activity videos such as SSv2. When pretrained for the same number of steps and data samples, our model exhibits a noticeable performance gain (+0.35) when trained on SSv2, confirming that web-scale human motion data significantly improves the model's ability to infer language-conditioned motion patterns.

Ablation on Motion Target Calculation: Table 6 examines how different formulations of motion supervision affect the ability to learn meaningful language-conditioned motion guidance prediction. Removing disentangling and

Pretrain Dataset	Train Steps	Avg Len
DROID	2000	4.04
SSv2	2000	4.39 (+0.35)

Table 5. **Ablation on Human Web Video Pretraining:** We investigate impact of human videos on our framework by pretraining for a common step count on both DROID [39] and SSv2 dataset. DROID is a large-scale "In-the-Wild" robotic manipulation dataset containing diverse scenes and actions in its teleoperated demonstration videos. We use 76K videos for both human data (randomly drawn from SSv2) and robotic data (full DROID). Despite equal dataset size & training steps, human data yielded superior performance. We hypothesize that this is due to higher causal diversity (e.g., complex tool use, varied viewpoints) in human videos.

Method	Avg Len
No Video Pretraining	3.91 (-0.48)
Static Frame Targets	4.28 (-0.11)
Raw Motion Targets	3.89 (-0.50)
Disentangled Motion Targets	4.39

Table 6. **Ablation on Motion Disentangling Strategy:** We examine the importance of motion disentangling for learning language driven dense pixel motion prediction from web videos of human activities. In line with prior work [100], our results demonstrate the difficulty of learning any meaningful signals without suitable disentangling of motion targets using our algorithm. In Row 1, we provide a baseline with no video pretraining, followed by pretraining with static frame targets (similar to VPP [35]), raw motion targets (no camera-object motion separation), and our disentangled motion targets (default) in Rows 2-4 respectively.

using raw optical flow (Row 3) results in significant performance drop, suggesting that camera-induced motion severely hinders learning of object-centric motion guidance. Similarly, static frame targets (Row 2), which ignore motion altogether, lead to underfitting of temporal dynamics. In contrast, our disentangled motion targets (Row 4, default) yield the best results, demonstrating that separating object motion from camera motion provides clean, physically meaningful supervision aligned with linguistic intent. This confirms that our motion disentangling algorithm is critical for grounding model predictions in genuine object dynamics rather than pixel changes from viewpoint shifts.

OF Quality: We also quantitatively measure our OF prediction quality, establishing strong performance as well as a positive correlation in OF quality to downstream task performance. See Appendix E for these results.

5. Conclusion

We presented FOFPred, a new framework for bridging language and motion through dense, pixel-level displacement prediction. Our disentangling algorithm refines noisy web video supervision by separating object and cam-

era motion, yielding cleaner learning signals. The unified VLM–diffusion backbone further enables robust multimodal learning from diverse and noisy caption data. Through extensive experiments across robotics and video generation, we demonstrate that motion-guided representations substantially enhance language-conditioned control and synthesis. We hope this work paves the way toward generalizable, motion-aware world models that understand and act through dynamic visual grounding.

References

[1] Rick Akkerman, Haiwen Feng, Michael J. Black, Dimitrios Tzionas, and Victoria Fernández Abrevaya. Interdyn: Controllable interactive dynamics with video diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2025. 7

[2] Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning optical flow from still images. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 15196–15206, 2021. 3

[3] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. In *Robotics: Science and Systems*, 2022. 2

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025. 4, 5, 15

[5] Simon Baker et al. A Database and Evaluation Methodology for Optical Flow. *International Journal of Computer Vision*, 2007. 18

[6] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. 2

[7] John L. Barron, David J. Fleet, and Steven S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12:43–77, 1992. 3

[8] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables diverse zero-shot robot manipulation. *Eur. Conf. Comput. Vis. (ECCV)*, 2024. 1, 2, 3

[9] Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Rich Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *ArXiv*, abs/2310.10639, 2023. 6

[10] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. $\pi 0$: A vision-language-action flow model for general robot control. *ArXiv*, abs/2410.24164, 2024. 6

[11] A. Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, and Dominik Lorenz. Stable video diffusion: Scaling latent video diffusion models to large datasets. *ArXiv*, abs/2311.15127, 2023. 8, 18

[12] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale. *Robotics science and systems (RSS)*, 2023. 6

[13] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael S. Ryoo, Paul E. Debevec, and Ning Yu. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13–23, 2025. 1, 3, 5, 20

[14] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1577–1585, 2024. 3

[15] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *ArXiv*, abs/2502.02492, 2025. 2, 4

[16] Annie S Chen, Suraj Nair, and Chelsea Finn. Learning Generalizable Robotic Reward Functions from "In-The-Wild" Human Videos. In *Robotics: Science and Systems*, 2021. 2

[17] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *ArXiv*, abs/2505.09568, 2025. 2, 3

[18] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Qiwei Liang, Zixuan Li, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025. 6, 7

[19] Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. Motion-conditioned

diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404*, 2023. 3

[20] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *ArXiv*, abs/2303.04137, 2023. 6

[21] Minkyoung Cho, Ruben Ohana, Christian Jacobsen, Adityan Jothi, Min-Hung Chen, Z Morley Mao, and Ethem Can. Tc-lora: Temporally modulated conditional lora for adaptive diffusion control. *arXiv preprint arXiv:2510.09561*, 2025. 2

[22] Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning Universal Policies via Text-Guided Video Generation. *arXiv:2302.00111*, 2023. 2, 3, 6

[23] Chelsea Finn and Sergey Levine. Deep Visual Foresight for Planning Robot Motion. In *IEEE International Conference on Robotics and Automation*, 2017. 2

[24] Chongkai Gao, Haozhuo Zhang, Zhixuan Xu, Zhehao Cai, and Lin Shao. Flip: Flow-centric generative planning as general-purpose manipulation world model. In *Int. Conf. Learn. Represent.*, 2025. 2

[25] Ruohan Gao, Bo Xiong, and Kristen Grauman. Im2flow: Motion hallucination from static images for action recognition. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5937–5947, 2018. 3, 7

[26] Mathias Gehrig, Manasi Muglikar, and Davide Scaramuzza. Dense continuous-time optical flow from event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4736–4746, 2024. 3

[27] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion prompting: Controlling video generation with motion trajectories. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–12, 2024. 3

[28] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter N. 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 IEEE International Conference on Computer Vision (ICCV)*, pages 5843–5851, 2017. 5

[29] Xianfan Gu, Chuan Wen, Jiaming Song, and Yang Gao. Seer: Language instructed video prediction with latent diffusion models. *ArXiv*, abs/2303.14897, 2023. 7

[30] Prerit Gupta, Shourya Verma, Ananth Grama, and Aniket Bera. Unified multi-modal interactive & reactive 3d motion generation via rectified flow. *arXiv preprint arXiv:2509.24099*, 2025. 2

[31] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *ArXiv*, abs/2010.09709, 2020. 3

[32] Zekun Hao, Xun Huang, and Serge Belongie. Controllable video generation with sparse trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[33] Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *ArXiv*, abs/2505.11709, 2025. 5

[34] Berthold Klaus Paul Horn, editor. *Robot vision*. MIT Press, Cambridge, MA, USA, 1986. 3

[35] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *ICML*, 2025. 2, 3, 5, 6, 7, 8

[36] Team Hunyuan3D, Bowen Zhang, Chunchao Guo, Haolin Liu, Hongyu Yan, Huiwen Shi, Jingwei Huang, Junlin Yu, Kunhong Li, Penghao Wang, et al. Hunyuan3d-omni: A unified framework for controllable generation of 3d assets. *arXiv preprint arXiv:2509.21245*, 2025. 2

[37] Xuan Ju, Weicai Ye, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Qiang Xu. Fulldit: Video generative foundation models with multimodal control via full attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15737–15747, 2025. 2

[38] Yash Kant, Ethan Weber, Jin Kyu Kim, Rawal Khirodkar, Su Zhaoen, Julieta Martinez, Igor Gilitschenski, Shunsuke Saito, and Timur Bagautdinov. Pippo: High-resolution multi-view humans from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16418–16429, 2025. 2

[39] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, P Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Ye Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sung Yul Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean-Pierre Mercat, Abdul Rehman, Pannag R. Sanketi, Archit Sharma, C. Blake Simpson, Quang Uyen Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Da Ling Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Muhammad Zubair Irshad, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosa Maria Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J. Lim, Jitendra Malik, Roberto Martin-Martin, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A

large-scale in-the-wild robot manipulation dataset. *ArXiv*, abs/2403.12945, 2024. 8

[40] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Josh Tenenbaum. Learning to act from actionless videos through dense correspondences. *ArXiv*, abs/2310.08576, 2023. 2, 3

[41] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jia-Liang Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fan Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Peng-Yu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhen Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Z. Xu, Yang-Dan Tao, Qinglin Lu, Songtao Liu, Daquan Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. *ArXiv*, abs/2412.03603, 2024. 5

[42] Mathis Koroglu, Hugo Caselles-Dupr'e, Guillaume Jeanneret Sanmiguel, and Matthieu Cord. Onlyflow: Optical flow based motion conditioning for video diffusion models, 2024. 2

[43] Akshay Krishnan, Xinchen Yan, Vincent Casser, and Abhijit Kundu. Orchid: Image latent diffusion for joint appearance and geometry generation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2

[44] Thanard Kurutach, Aviv Tamar, Ge Yang, Stuart J Russell, and Pieter Abbeel. Learning Plannable Representations with Causal InfoGAN. In *Neural Information Processing Systems*, 2018. 2

[45] Teyun Kwon, Norman Di Palo, and Edward Johns. Language models as zero-shot trajectory generators. *IEEE Robotics and Automation Letters*, 9(7):6728–6735, 2024. 3

[46] Black Forest Labs, Stephen Batifol, A. Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Muller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *ArXiv*, abs/2506.15742, 2025. 4, 5, 15

[47] Hoang Le, Feng Liu, Shu Zhang, and Aseem Agarwala. Deep homography estimation for dynamic scenes. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7649–7658, 2020. 4

[48] Jangwon Lee and Michael S Ryoo. Learning Robot Activities from First-Person Human Videos Using Convolutional Future Regression. In *CVPRW*, 2017. 2

[49] Guojun Lei, Chi Wang, Rong Zhang, Yikai Wang, Hong Li, and Weiwei Xu. Animateanything: Consistent and controllable animation for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[50] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chi-Hou Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. Vision-language foundation models as effective robot imitators. *ArXiv*, abs/2311.01378, 2023. 6

[51] Xiaoqi Li, Jingyun Xu, Mingxu Zhang, Jiaming Liu, Yan Shen, Iaroslav Ponomarenko, Jiahui Xu, Liang Heng, Siyuan Huang, Shanghang Zhang, et al. Object-centric prompt-driven vision-language-action model for robotic manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27638–27648, 2025. 2

[52] Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Yuexian Zou, and Ying Shan. Image conductor: Precision control for interactive video synthesis. *arXiv preprint arXiv:2406.15339*, 2024. 3

[53] Jingyun Liang, Yuchen Fan, Kai Zhang, Radu Timofte, Luc van Gool, and Rakesh Ranjan. Movideo: Motion-aware video generation with diffusion model. In *European Conference on Computer Vision*, 2024. 2

[54] Fanfan Liu, Feng Yan, Liming Zheng, Chengjian Feng, Yiyang Huang, and Lin Ma. Robouniview: Visual-language model with unified view representation for robotic manipulaiton. *ArXiv*, abs/2406.18977, 2024. 6

[55] Haiyang Liu, Zhan Xu, Fa-Ting Hong, Hsin-Ping Huang, Yi Zhou, and Yang Zhou. Video motion graphs. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2

[56] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024. 6

[57] Xiao Liu, Yifan Zhou, Fabian Weigend, Shubham Sonawani, Shuhei Ikemoto, and Heni Ben Amor. Diff-control: A stateful diffusion-based policy for imitation learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7453–7460. IEEE, 2024. 2

[58] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, 1981. 4, 17

[59] Ao Luo, Xin Li, Fan Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Flowdiffuser: Advancing optical flow estimation with diffusion models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19167–19176, 2024. 3

[60] Oier Mees, Lukás 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*, 7: 7327–7334, 2021. 6, 7

[61] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A Universal Visual Representation for Robot Manipulation. In *Conference on Robot Learning*, 2022. 2

[62] Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable

image animation via generative motion field adaptions in frozen image-to-video diffusion model. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[63] Nvidia, Johan Bjorck, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *ArXiv*, abs/2503.14734, 2025. 2

[64] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries. *ArXiv*, abs/2504.06256, 2025. 2, 3

[65] Jyothish Pari, Nur Muhammad Shafiullah, Sridhar Pandian Arunachalam, and Lerrel Pinto. The Surprising Effectiveness of Representation Learning for Visual Imitation. In *Robotics: Science and Systems*, 2022. 2

[66] Bo Peng, Xinyuan Chen, Yaohui Wang, Chaochao Lu, and Yu Qiao. Conditionvideo: Training-free condition-guided video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4459–4467, 2024. 2

[67] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2024. 2

[68] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. Controlmm: Controllable masked motion generation. *arXiv preprint arXiv:2410.10780*, 2024. 2

[69] Kanchana Ranasinghe, Xiang Li, Cristina Mata, Jong Sung Park, and Michael S. Ryoo. Pixel motion as universal representation for robot control. *ArXiv*, abs/2505.07817, 2025. 1, 2, 3, 4, 5, 6, 7, 18

[70] Scott Reed, Konrad Zolna, Emilio Parisotto, and et al. A generalist agent. *Transactions on Machine Learning Research (TMLR)*, 2022. 2

[71] Juntao Ren, Priya Sundaresan, Dorsa Sadigh, Sanjiban Choudhury, and Jeannette Bohg. Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning. *ArXiv*, abs/2501.06994, 2025. 2

[72] Moritz Reuss, Ömer Erdinç Yagmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. *ArXiv*, abs/2407.05996, 2024. 5, 6

[73] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 8

[74] Lin Shao, Toki Migimatsu, Qiang Zhang, Karen Yang, and Jeannette Bohg. Concept2Robot: Learning Manipulation Concepts from Instructions and Human Demonstrations. *IJRR*, 2021. 2

[75] Pratyusha Sharma, Deepak Pathak, and Abhinav Gupta. Third-person visual imitation learning via decoupled hierarchical controller. In *Neural Information Processing Systems*, 2019. 2

[76] Yash Sharma, Yi Zhu, Chris Russell, and Thomas Brox. Pixel-level correspondence for self-supervised learning from video. *ArXiv*, abs/2207.03866, 2022. 3

[77] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Y. Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Da, and Hongsheng Li. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. *ACM SIGGRAPH 2024 Conference Papers*, 2024. 3

[78] Aravind Sivakumar, Kenneth Shaw, and Deepak Pathak. Robotic Telekinesis: Learning a Robotic Hand Imitator by Watching Humans on Youtube. In *Robotics: Science and Systems*, 2022. 2

[79] Daeun Song, Jing Liang, Xuesu Xiao, and Dinesh Manocha. Vl-tgs: Trajectory generation and selection using vision language models in mapless outdoor environments. *IEEE Robotics and Automation Letters*, 2025. 3

[80] Sruthi Sudhakar, Ruoshi Liu, Basile Van Hoorick, Carl Vondrick, and Richard Zemel. Controlling the world by sleight of hand. *ArXiv*, abs/2408.07147, 2024. 2, 3, 7

[81] Shikun Sun, Min Zhou, Zixuan Wang, Xubin Li, Tiezheng Ge, Zijie Ye, Xiaoyu Qin, Junliang Xing, Bo Zheng, and Jia Jia. Minimal impact controlnet: Advancing multi-controlnet integration. In *Int. Conf. Learn. Represent.*, 2025. 2

[82] Shao-Hua Sun, Hyeonwoo Noh, Sriram Somasundaram, and Joseph Lim. Neural program synthesis from diverse demonstration videos. In *International Conference on Machine Learning*, 2018. 2

[83] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020. 3, 4

[84] Shuyuan Tu, Qi Dai, Zhi-Qi Cheng, Han Hu, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motioneditor: Editing video motion via content-aware diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2024. 2

[85] Jacob Walker, Abhinav Kumar Gupta, and Martial Hebert. Dense optical flow prediction from a static image. In *Int. Conf. Comput. Vis. (ICCV)*, pages 2443–2451, 2015. 3

[86] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Xiaofeng Meng, Ningying Zhang, Pandeng Li, Ping Wu, Ruihang Chu, Rui Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wen-Chao Zhou, Wente Wang, Wen Shen, Wenyuan Yu, Xianzhong Shi, Xiaomin Huang, Xin Xu, Yan Kou, Yan-Mei Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhengbin Han, Zhigang Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *ArXiv*, abs/2503.20314, 2025. 5

[87] Boyang Wang, Nikhil Sridhar, Chao Feng, Mark Van der Merwe, Adam Fishman, Nima Fazeli, and Jeong Joon Park.

This&that: Language-gesture controlled video generation for robot planning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12842–12849. IEEE, 2025. 2

[88] Qinghe Wang, Yawen Luo, Xiaoyu Shi, Xu Jia, Huchuan Lu, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–10, 2025. 2

[89] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 3

[90] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 36–54. Springer, 2024. 3

[91] Zhouxia Wang, Fanqing Lin, Xintao Wang, Takuhiro Kaneko, Jianfeng Lu, Chuan He, Fengyun Han, Zhibin Wang, Yuxu Wang, Jie Chen, Jianxiong Xiao, Chao Dong, Zhaoxiang Zhang, and Xiaokang Yang. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Proceedings*, 2024. 1, 3

[92] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *ArXiv*, 2023. 1, 2, 3, 18, 19

[93] Youpeng Wen, Junfan Lin, Yi Zhu, Jianhua Han, Hang Xu, Shen Zhao, and Xiaodan Liang. Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. *ArXiv*, abs/2411.09153, 2024. 6

[94] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation. *ArXiv*, abs/2506.18871, 2025. 2, 3, 4, 5, 8, 16

[95] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pretraining for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023. 6

[96] Weijia Wu, Jiajun Liang, Jianqiao Zhou, Jian Zhang, Qingsheng Han, Jinlin Liu, Xiangyu Fan, Dongdong Yu, Zaidan Zhang, and Changhua Zhou. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision (ECCV)*, 2024. 3, 5

[97] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13294–13304, 2024. 2, 3

[98] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. DynamiCrafter: Animating open-domain images with video diffusion priors. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 399–417, 2024. 7

[99] Mengda Xu, Zhenjia Xu, Yinghao Xu, Cheng Chi, Gordon Wetzstein, Manuela Veloso, and Shuran Song. Flow as the cross-domain manipulation interface. In *Conference on Robot Learning*, 2024. 1, 2, 3, 4, 18

[100] Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, Yuquan Deng, Lars Lidén, and Jianfeng Gao. Magma: A foundation model for multimodal ai agents. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14203–14214, 2025. 2, 3, 4, 8

[101] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. *ArXiv*, abs/2408.06072, 2024. 5, 7, 8, 18, 19, 20, 21

[102] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2024. 7

[103] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Ming Gong, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.10640*, 2023. 1, 3, 5

[104] Xin Yu, Tianyu Wang, Soo Ye Kim, Paul Guerrero, Xi Chen, Qing Liu, Zhe Lin, and Xiaojuan Qi. Objectmover: Generative object movement with video prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17682–17691, 2025. 3

[105] Chengbo Yuan, Chuan Wen, Tong Zhang, and Yang Gao. General flow as foundation affordance for scalable robot learning. *arXiv preprint arXiv:2401.11439*, 2024. 1, 2, 3

[106] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024. 6

[107] Hang Zhang, Jinglin Chen, Chen Ma, and Yuan Lin. Tora: Trajectory-oriented diffusion transformer for video generation. *arXiv preprint arXiv:2404.15617*, 2024. 3

[108] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824, 2023. 2

[109] Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge. *ArXiv*, abs/2507.04447, 2025. 3, 5, 6

[110] Kaifeng Zhao, Gen Li, and Siyu Tang. Dartcontrol: A diffusion-based autoregressive motion model for real-time

text-driven motion control. In *International Conference on Learning Representations*, 2024. 3

[111] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023. 6

[112] Zhide Zhong, Haodong Yan, Junfeng Li, Xiangchen Liu, Xin Gong, Wenxuan Song, Jiayi Chen, and Haoang Li. Flowvla: Thinking in motion with a visual chain of thought. *ArXiv*, 2025. 3, 4

[113] Haitao Zhou, Chuang Wang, Rui Nie, Jinlin Liu, Dongdong Yu, Qian Yu, and Changhu Wang. Trackgo: A flexible and efficient method for controllable video generation. *arXiv preprint arXiv:2408.11475*, 2024. 3

Future Optical Flow Prediction Improves Robot Control & Video Generation

Supplementary Material

Reproducibility Statement

To ensure the reproducibility of our results, we have made our code, trained model weights, and data configurations publicly available. Our method builds upon open-source pretrained models retrieved from the Hugging Face Hub. Detailed implementation specifics, hyperparameter settings, and experimental protocols are thoroughly documented in the main text and the Appendix.

Acknowledgments

We would like to thank the broader research team at Salesforce for support, feedback and guidance. In particular, we thank Shelby Heinecke, Juntao Tan, Srinath Meadusani, Eric Xu, Matthew Fernandez, Jim Jagielski, Joyce Zheng, and Jinxuan Xu for technical feedback and support; Jeanette Berberich and Jennifer McCallion for legal coordination; Janna Remperas, Ian Thomas, Mitra Mitchell, and Alex Dalton for logistical help and encouragement throughout the project.

Contributions

KR led the project by implementing the preliminary idea of language-driven OF prediction, building the codebase for experimentation, and performing most of the evaluations. HZ discussed all aspects of the project, contributed to several key design choices (on model architecture, OF calculation, video evaluations), and helped debug several technical issues. YF discussed multiple aspects of the robotic evaluation pipelines, helped implement the RoboTwin evaluations, and performed several robotic baseline evaluations. LY contributed to building the video evaluation pipeline, helped investigate our novelty against prior works, and discussed several aspects of our video generation pipeline. LX proposed using the unified VLM-Diffusion architecture, provided feedback on our data and training pipeline implementations, and discussed several early results. RX helped streamline early exploration of the idea, supported scaling up our training pipeline, and provided several critical feedback on project design. CX & SS provided the strategic vision for the project, oversaw the research environment, and shaped the high-level framing of the research problem. MR set the direction for robotic downstream tasks and discussed multiple aspects of the project idea, scope, and implementation. JN organized the overall project, set the research direction, and discussed all aspects of the project idea, scope, and implementation.

Appendix Contents

A. Additional Architectural Details 15

B. Optical Flow Representation 16

C. Optical Flow Calculation 16

D. Motion-Guided Frame Sampling 16

E. OF Quality Evaluation 17

F. Additional Ablations 18

G. Limitations and Future Work 18

H. Visualizations 19

A. Additional Architectural Details

We provide additional details of our architecture in this section, focusing on the conditional input processing and the modifications made to the Diffusion Transformer (DiT). To recap, our overall architecture contains 3 main components: the VLM (3B parameters), VAE (83M parameters), and DiT (4B parameters). The DiT (diffusion transformer) module is the key component that we modify and train to construct our FOFPred model.

A.1. Conditional Input Processing

The core architecture utilizes two distinct conditional features: $\boldsymbol{f}_c$ (textual) and $\boldsymbol{f}_v$ (visual), which are passed through Multi-Layer Perceptrons (MLPs) to ensure dimensional compatibility with the Diffusion Transformer.

- **Textual Feature Projection:** The textual feature $\boldsymbol{f}_c$, obtained from the Qwen2.5-VL [4] Vision-Language Model (VLM), has an initial channel dimension of 2520. This feature is projected via an MLP layer to a common channel dimension D. The VLM input consists of the natural language caption c paired with the interleaved visual inputs $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_t$.
- **Visual Feature Projection:** The visual feature $\boldsymbol{f}_v$, obtained from the Flux.1 VAE [46] encoder, has an initial channel dimension of 16. This feature is reshaped using 2×2 grid to obtain $64 = 16 \times 2 \times 2$ channel dimension vectors, followed by projected via a different MLP layer to the same common channel dimension D.

- **Initial Noise Vector:** The noise vector, sampled from the VAE latent space, is also projected into the DiT channel dimension using the same reshaping operation and visual feature project MLP layer.

We enforce the condition that the output dimensions of the two MLPs are equal to $D = 2520$ which is the DiT input channel dimension. These projected conditional features, $\hat{\boldsymbol{f}}_c$ and $\hat{\boldsymbol{f}}_v$, are simply appended to the input noise sequence of the DiT, providing comprehensive context for the diffusion process.

A.2. DiT Modifications for Temporal Modeling

Our Diffusion Transformer is based on the OmniGen architecture [94], adapted to explicitly handle the temporal nature of our sequence prediction task (future optical flow $\hat{\boldsymbol{y}}_{t+1:t+4}$ from inputs $\boldsymbol{x}_{t-1}, \boldsymbol{x}_t$).

Time-Aware 3D RoPE: We implemented a modified RoPE module to handle the required temporal dimension within the sequence. The original OmniGen RoPE provides 3-axis position encoding (L, H, W) for text length (L), image height (H), and image width (W). Our modification interprets these three axes as:

- **Axis 1 (Text/Image Shift):** Used to assign unique base position IDs for the text tokens and for each separate image (or image sequence). This axis is now leveraged to encode the **temporal offset** or frame index for the input sequences $(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t)$ and the latent noisy output $\boldsymbol{y}$.
- **Axis 2 and 3 (Image H, W):** Encode the spatial positions within each frame's patch grid.

Specifically, for multiple frames (e.g., $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_t$ representations), the position shift variable is incremented after processing each frame. This ensures that while text tokens and $\boldsymbol{x}_{t-1}$ tokens receive their respective base position IDs, the tokens corresponding to $\boldsymbol{x}_t$ receive a unique, subsequent base position ID, which acts as a temporal index. Overall, this ensures that the tokens are timestamped with temporal position using RoPE, allowing the transformer to distinguish between and model the relationship across the consecutive time steps in both the input and output frame sequences.

Spatio-Temporal Attention: The DiT transformer blocks are updated to perform *full spatio-temporal attention* over the output tokens. This means the self-attention mechanism is configured to operate over the entire input sequence $\boldsymbol{f}_{in} = [\hat{\boldsymbol{f}}_c, \hat{\boldsymbol{f}}_v, \boldsymbol{f}_y]$, where $\boldsymbol{f}_y$ is the latent sequence of the noisy optical flow. This design choice allows the DiT to explicitly capture the motion dynamics and sequence dependencies required for accurate predictions of future optical flow frame sequences.

B. Optical Flow Representation

The visualization algorithm begins by converting the Cartesian optical flow field, denoted as $F \in \mathbb{R}^{2\times H\times W}$ with components f_x and f_y, into a polar representation. The magnitude of the flow is computed as $M = \sqrt{f_x^2 + f_y^2}$ and normalized by a scaling factor $\eta = 64.0$. This normalized magnitude, $\hat{M}$, is clamped to the range $[0, 1]$. The directional angle is derived using the four-quadrant inverse tangent function, $\phi = \arctan2(f_y, f_x)$, and is shifted by π to ensure the final angle θ lies within the interval $[0, 2\pi]$.

Subsequently, these polar coordinates are mapped to the HSV color space to generate an RGB image. The hue channel H is assigned the flow angle θ, encoding the direction of motion, while the saturation channel S is defined by the normalized magnitude $\hat{M}$. The value channel V is set to a constant maximum intensity of 1 (formulated in the algorithm as $\hat{M} + (1 - \hat{M})$) - this eliminates large color variances across consecutive frames due to outliers. Finally, the resulting HSV tensor is converted into RGB format using a differentiable color space transformation.

C. Optical Flow Calculation

The proposed method (Algorithm 1) compensates for camera motion within dense optical flow fields by leveraging the flow data itself to derive dense correspondences. Rather than relying on computationally expensive sparse feature extraction, we uniformly sample grid points from the source frame and project them into the target frame using the raw flow vectors $\mathbf{F}_{\text{raw}}$. These calculated point pairs serve as inputs for a RANSAC-based homography estimation, which computes a transformation matrix $\mathbf{H}$ representing the global camera motion. A dense "camera flow" field, $\mathbf{F}_{\text{cam}}$, is subsequently synthesized by applying $\mathbf{H}$ to the entire coordinate grid and calculating the displacement vectors. Finally, the object-centric motion is isolated by subtracting the estimated camera flow from the raw flow, such that $\mathbf{F}_{\text{obj}} = \mathbf{F}_{\text{raw}} - \mathbf{F}_{\text{cam}}$. A post-processing magnitude threshold is applied to $\mathbf{F}_{\text{obj}}$ to suppress residual noise and artifacts arising from imperfect alignment.

We reiterate that this one-time process is run offline. For our 500,000 training videos, we are able to complete processing using 4 A100 GPUs in roughly 30 hours. This timing also includes the motion-guided frame sampling (negligible time for this operation compared to relative flow calculation) that we describe in the next section.

D. Motion-Guided Frame Sampling

Training video prediction models on natural sequences requires careful data curation, as large portions of raw video may contain static scenes or imperceptible motion that contribute little to the learning process. To address this, we em-

Algorithm 1 Relative Optical Flow Calculation

Require:
$\mathbf{F}_{\text{raw}} \in \mathbb{R}^{B\times 2\times H\times W}$: Default optical flow tensors
τ_{ransac}: RANSAC reprojection threshold (default 5.0)
s: Sampling stride (default 8)
τ_{noise}: Post-compensation noise threshold (default 0.5)
Ensure:
$\mathbf{F}_{\text{comp}}$: Camera-motion compensated flow tensors

procedure COMPENSATEFLOW($\mathbf{F}_{\text{raw}}, s, \tau_{\text{ransac}}, \tau_{\text{noise}}$)
  $B, C, H, W \leftarrow$ shape($\mathbf{F}_{\text{raw}}$)
  $\mathbf{F}_{\text{comp}} \leftarrow$ List()
  $\mathcal{G} \leftarrow \{(x, y) \mid 0 \le x < W, 0 \le y < H\}$
  $\mathcal{S} \leftarrow$ Sample($\mathcal{G}, s$); $\mathbf{p}_0 \leftarrow$ Coords($\mathcal{S}$)
  for $i \leftarrow 0$ **to** $B - 1$ **do**
    $\mathbf{f}_i \leftarrow \mathbf{F}_{\text{raw}}[i]$
  Derive Correspondences
    $\mathbf{v} \leftarrow$ Extract($\mathbf{f}_i, \mathcal{S}$)
    $\mathbf{p}_1 \leftarrow \mathbf{p}_0 + \mathbf{v}$
  Estimate Homography
    $\mathbf{H} \leftarrow$ FindHomography($\mathbf{p}_0, \mathbf{p}_1$, RANSAC, τ_{ransac})
    if $\mathbf{H}$ is valid **then**
    Compute Camera Flow & Compensation
      $\mathcal{G}' \leftarrow$ PerspectiveTransform($\mathcal{G}, \mathbf{H}$)
      $\mathbf{f}_{\text{cam}} \leftarrow \mathcal{G}' - \mathcal{G}$
      $\mathbf{f}_{\text{obj}} \leftarrow \mathbf{f}_i - \mathbf{f}_{\text{cam}}$
    Post-Compensation Thresholding
      if Threshold Enabled **then**
        $\mathbf{M} \leftarrow \|\mathbf{f}_{\text{obj}}\|_2$
        $\mathbf{f}_{\text{obj}}[\mathbf{M} < \tau_{\text{noise}}] \leftarrow 0$
      end if
      $\mathbf{f}_{\text{final}} \leftarrow \mathbf{f}_{\text{obj}}$
    else
      $\mathbf{f}_{\text{final}} \leftarrow \mathbf{f}_i$
    end if
    Append $\mathbf{f}_{\text{final}}$ to $\mathbf{F}_{\text{comp}}$
  end for
  return Stack($\mathbf{F}_{\text{comp}}$)
end procedure

ploy a two-stage filtering pipeline, detailed in Algorithm 2, which acts as a computational gate to prioritize high-motion segments before generating expensive ground-truth labels. Calculating dense optical flow (e.g., using RAFT or PWC-Net) for every frame pair in a large-scale dataset is computationally prohibitive. To circumvent this bottleneck, we utilize a lightweight approximation strategy (Lines 4–6). We first spatially downsample all input frames to a resolution of 32×32 pixels. This reduction removes high-frequency textures and compression artifacts while preserving dominant structural motion. On these low-resolution pairs (I^{low}), we apply the Lucas-Kanade method [58]. Unlike deep learning

Algorithm 2 Motion-Aware Frame Selection

Require: Video Sequence $V = \{I_1, I_2, \ldots, I_N\}$, Threshold τ, Top-Percentile k
Ensure: Selected Frame Indices S
1 Initialize selected set $S \leftarrow \emptyset$
2 **for** $t \leftarrow 1$ **to** $N - 1$ **do**
3 ▷ Step 1: Spatial Downsampling for efficiency
4 $I_t^{low}, I_{t+1}^{low} \leftarrow$ Resize($I_t, I_{t+1}, 32 \times 32$)
5 ▷ Step 2: Fast Optical Flow Approximation
6 $F_{LK} \leftarrow$ LucasKanade(I_t^{low}, I_{t+1}^{low})
7 ▷ Step 3: Compute Motion Proxy
8 $M \leftarrow \|F_{LK}\|_2$ ▷ Calculate L2 norm of flow vectors
9 $\mu_{proxy} \leftarrow$ Percentile(M, k) ▷ Extract top-k% value
10 ▷ Step 4: Threshold Filtering
11 **if** $\mu_{proxy} > \tau$ **then**
12 $S \leftarrow S \cup \{t\}$
13 **end if**
14 **end for**
15 **return** S

approaches, Lucas-Kanade relies on local least-squares optimization, which converges rapidly on small spatial grids (32×32), allowing for high-throughput processing of millions of frames. To robustly distinguish meaningful scene activity from background noise, we derive a scalar motion proxy, μ_{proxy}, from the raw flow field (Lines 8–9). We compute the magnitude (L2 norm) of the flow vectors and select the top-k percentile value (k = 10) rather than the mean. This percentile-based approach ensures that the metric is driven by the moving objects within the scene, rather than being diluted by large static background regions. Frame pairs are included in the final training dataset only if this proxy value exceeds an empirical threshold τ (Lines 11–12), ensuring the model focuses on sequences with significant temporal dynamics.

The result of this motion-based frame filtering is the elimination of seemingly static regions of the video. In practice, we observe that certain static regions are removed while most frames in other motion-heavy regions remains. The resulting filtered videos still produce coherent video segments without breaking the natural temporal continuity of the motion. The low threshold for motion filtering (we use a 5 pixel length as threshold for 256×256 images) is crucial for this. We set this based on empirical observation over a randomly selected set of videos. We highlight again that tune the filtering process is important to ensure coherent frame sequences, instead of generating frame sequences that may contain large discontinuities.

E. OF Quality Evaluation

We quantitatively measure the quality of our predicted future optical flow and report these results in Table A.1. Our

Method	EPE ($\times 100$) ↓	CS ↑
LTM [69]	137	0.243
Im2Flow2Act [99]	78.3	0.0
SVD [11]	2.29	0.61
CogVideoX [101]	1.84	0.64
CogVideoX† [101]	7.64	0.52
FOFPred (Ours)	**1.48**	**0.74**

Table A.1. **OF Quality:** We report End Point Error (EPE) [5] & Cosine-Similarity (CS) between predicted & GT OF, calculated on SSv2 (val set). All models are trained to predict OF, except CogVideoX† (off-the-shelf weights) where OF is calculated from predicted RGB frames.

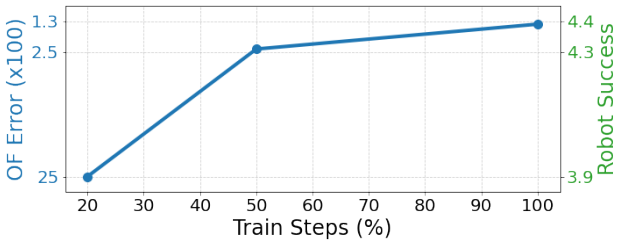

Figure A.1. **OF $\propto$ downstream:** Lower end-point-error [5] (SSv2) in predicted OF in pretraining leads to improved downstream robotic performance (AL metric on CALVIN).

model predicts better OF compared to prior work exploring language driven OF prediction [69, 99] as well as video diffusion networks adapted to predict OF (SVD [11] and CogVideoX [101]). We also explore a naive baseline of using off-the-shelf CogVideoX model (for RGB future frame prediction), and calculating OF from the prediction RGB frames. Our proposed FOFPred framework outperforms all of these baselines.

We also use different checkpoints of our FOFPred model from earlier steps of our training process (i.e. later checkpoints perform better OF prediction) to investigate the correlation between OF prediction quality and downstream task performance. We illustrate these results in Figure A.1. OF prediction error and robotic task performance (AL metric on CALVIN) are plotted. Note the separately scaled axes and flipped OF error axis, done for simpler visualization. We take this result as indication of a strong correlation between OF prediction quality and downstream robotic task performance.

F. Additional Ablations

We provide two new ablations on using dense optical flow as our representation and the significance of the motion signal passed into our robot action policy in Table A.2 and Table A.3 respectively.

Dense vs. Sparse Motion: We investigate the impact of dense vs. sparse motion representations on the downstream robot control task in Table A.2, reporting average length on the CALVIN benchmark. Training and evaluation settings are identical to our ablations in the main paper.

To isolate the value of our dense representation, we compare our method against two sparse variants: a naive baseline where our predicted future optical flows are sub-sampled to a 16×16 grid, and a re-implementation of the sparse trajectory model, ATM [92], trained on our dataset. The results demonstrate a distinct advantage for spatially dense motion information. Our default dense model achieves an average length of 4.39, significantly outperforming the ATM baseline (2.92) and the sub-sampled variant (1.24). This suggests that the fine-grained, pixel-level dynamics captured by dense optical flow are essential for precise robotic manipulation, whereas sparse representations fail to capture the complete motion context required for complex tasks.

Ablations on Motion Forecast Conditioning: We investigate the usefulness of motion forecasting as opposed to generic visual representations for our downstream robot control task in Table A.3. We compare our full framework against two baselines: one where motion input is entirely removed (note that our VLM input is a representation of both images and text), and another where the motion input is replaced by static visual embeddings from our VAE encoder. The results are decisive; the policy fails almost entirely without motion input (0.02) and shows only marginal improvement when provided with the static visual features (0.52). In contrast, conditioning on our predicted future optical flow embeddings yields a score of 4.39. This confirms that the predicted optical flow provides unique, critical dynamic information that static visual representations cannot substitute.

Video Generation Upper Bound: We explore a possible upper bound for our video generation results in Table 3 which uses optical flow calculated from each ground truth (GT) video in the test set=. Results in Table A.4 indicate a significant gap in the upper bound compared to both baselines and our method. Given the multi-modal output nature of OF prediction from a single frame, this upper bound should be viewed cautiously because the oracle OF (calculated from GT video) has the same mode as GT video. Predicted flows can be plausible or even optimal yet score lower simply by following a different valid mode.

G. Limitations and Future Work

Our FOFPred model currently contains several limitations. It is sensitive to the text prompt (i.e. slightly different text prompts could lead to wrong predictions). For example, changing a text "moving from right to left" → "moving left" results in wrong predictions for some samples. In addition,

Method	Motion	Avg Len ↑
Naive Baseline	Sparse	1.24 (-3.15)
ATM	Sparse	2.92 (-1.47)
Ours (default)	Dense	4.39

Table A.2. **Dense vs. Sparse Motion:** We report the average length metric (Avg Len) on CALVIN benchmark. We demonstrate the significance of our dense motion representation. The first row uses a naive sub-sampling of our predicted future optical flows to obtain motion for points on a uniform 16×16 grid on the image. This variant uses training and inference identical to ours. The second row re-implements the ATM [92] approach with training on our same data but inference similar to the original work. We highlight the clear improvements arising from our proposed method.

Method	Motion Input	Avg Len ↑
No motion input	✗	0.02 (-4.37)
Static visual input	✗	0.52 (-3.87)
Ours (default)	✓	4.39

Table A.3. **Motion Input Ablation:** We report the average length metric (Avg Len) on CALVIN benchmark. For our robot control extension, we ablate the motion input under two settings. First (row 1) we remove the motion input leaving only the state and text goal inputs. Next (row 2) we replace the motion input with an embedding (from our VAE) of the visual input. Our motion inputs in the form of future optical flow clearly leads to a performance improvement.

Method	FVD ↓	KVD ↓	LPIPS ↓
CogVideoX	78.47	12.46	30.3
Ours	75.39	11.38	28.5
Upper Bound	17.58	1.92	22.5

Table A.4. **Upper bound for downstream video generation**: Large gap since upper-bound uses GT videos for OF which has same mode in the the multi-modal output space.

the current FOFPred is a relatively large model (∼7B parameters in total) that is expensive to deploy in a real-time setting and requires considerable compute (at least 24GB of GPU) for inference.

On the other hand, our model captures diversity of motion patterns quite well and generates meaningful future optical flow in as less as 1 reverse diffusion (i.e. denoising) iteration. Across different seeds for the same frame-caption pair, our model generates a diverse distribution of mostly meaningful future optical flow: we attribute this strength to the diffusion based training. In Figure A.2, we explore the diversity of our FOFPred base model predictions, highlighting limitations as well. Across different seeds, we notice

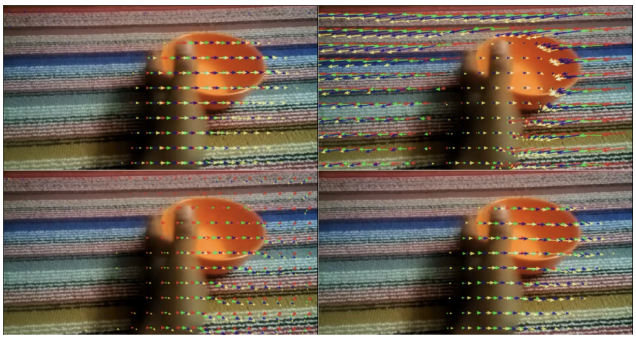

Figure A.2. **Sensitivity to seed:** We visualize 4 FOFPred predictions for the same image and prompt, `"Moving the bowl from left to right"`, but using 4 different starting noise vectors for the reverse diffusion process. Notice how the upper-right corner conflates the object motion with a camera motion instead; however this camera motion does correspond to the object motion described in the provided prompt. In the lower left image, in addition to the desired object motion, we again observe a slight amount of corresponding camera motion.

unexpected camera motion sometimes. While often relevant to the provides motion prompt, this does not explicit capture the object motions that we desire from our model.

On fast convergence (during reverse diffusion), we note how our model is not distilled or trained specifically for this purpose; RGB image generation with identical architecture and training pipeline usually requires at least 20 sampling steps for meaningful generation. We hypothesize that optical flow is a less complex distribution compared to natural images (i.e. lies on a lower dimensional manifold and contains less high frequency information such as textures or patterns): this allows even a single sampling step of reverse diffusion to generate meaningful outputs.

In future work, we plan to explore training data augmentation with automated text-label re-phrasing and model distillation into light-weight architectures to address our key limitations. We also aim to further investigate and quantify the diversity aspect of FOFPred, analyze the fast convergence aspect, and explore possibility of real-time inference.

H. Visualizations

We first visualize predicted optical flow from our base FOFPred for several selected image pairs in Figure A.3. We select image pairs to highlight the *language-driven* future optical flow generation capability of our FOFPred framework.

We next visualize some videos generated with our framework for pairs of first frames and motion-focused captions obtained from the SSv2 dataset. Figure A.4 presents this qualitative comparison between the ground truth (GT) video, our T2V baseline from CogVideoX [101], and our

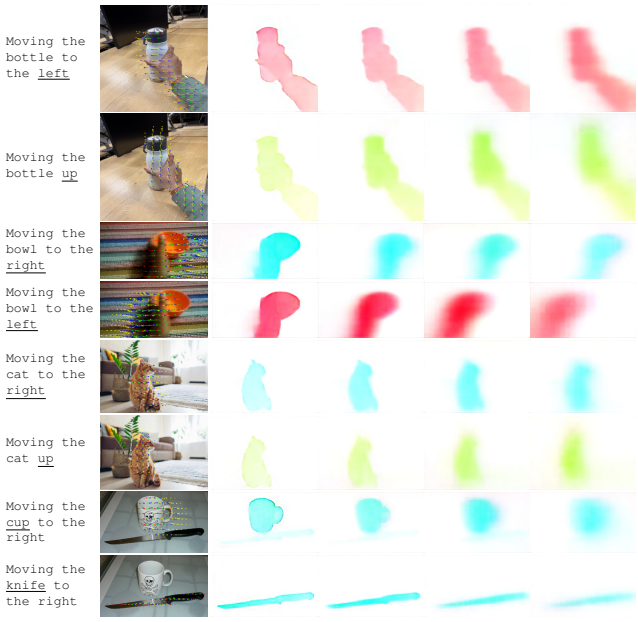

Figure A.3. **OF Quality:** We visualize our predicted future OF (4 timesteps) as sparse arrows (left: color for timestep) & raw prediction (right: color for direction). We use pairs of the same image with different prompts. Best viewed zoomed.

framework, FOFPred[1]. The samples are selected from the SSv2 validation split to demonstrate generation capabilities conditioned on the first frame (outlined in green) and a specific text goal. As observed across the samples, FOFPred consistently exhibits superior motion adherence compared to the baseline, which frequently struggles to generate significant movement or directionally accurate dynamics (e.g., the "Moving glue stick" and "Moving cycle" examples). While FOFPred successfully executes the semantic requirements of the text prompts, we also visualize a partial failure case in the second row ("Pulling toy car"). In this instance, while our model correctly adheres to the motion instruction, our extended pipeline exhibits a trade-off in visual fidelity, resulting in slight distortion of the object's appearance compared to the baseline. However, the baseline in this case moves the car in the wrong direction.

[1]In our work, we explore Text-to-Video (T2V) generation using a two-stage pipeline that connects FOFPred with the existing video synthesis model Go-with-the-Flow (GWTF) [13]. GWTF is built on top of CogVideoX [101] and extends it to additionally accept a user-provided motion prompt as input.

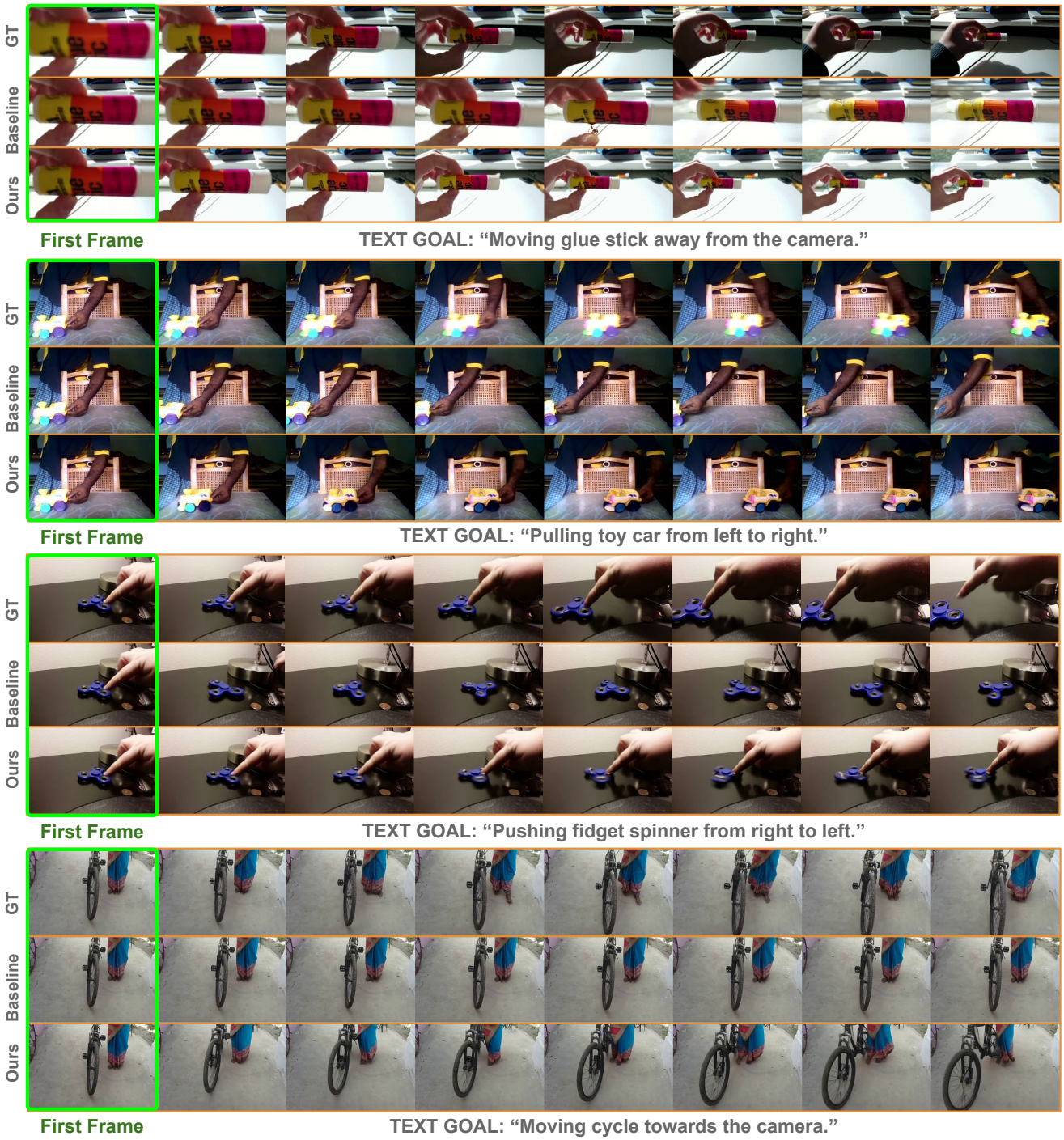

Figure A.4. **Visualization of success and failure cases for Text-to-Video (T2V) generation:** We visualize some success and failure cases for our framework over the baseline, CogVideoX [101]. Examples are drawn from the SSv2 validation split. We note that our method consistently improves motion adherence over the baseline. However, in some cases our framework distorts the visual appearance of objects although they undergo correct movement (e.g., see “toy car” in Row 2).