HANDI: Hand-Centric Text-and-Image Conditioned Video Generation

Yayuan Li[1,*] Zhi Cao[1,*] Jason J. Corso[1,2]
[1]University of Michigan [2]Voxel51
Full Paper

Abstract

Despite the recent strides in video generation, state-of-the-art methods still struggle with elements of visual detail. One particularly challenging case is the class of videos in which the intricate motion of the hand coupled with a mostly stable and otherwise distracting environment is necessary to convey the execution of some complex action and its effects. To address these challenges, we introduce a new method for video generation that focuses on hand-centric actions. Our diffusion-based method incorporates two distinct innovations. First, we propose an automatic method to generate the motion area—the region in the video in which the detailed activities occur—guided by both the visual context and the action text prompt, rather than assuming this region can be provided manually as is now commonplace. Second, we introduce a critical Hand Refinement Loss to guide the diffusion model to focus on smooth and consistent hand poses. We evaluate our method on challenging augmented datasets based on EpicKitchens and Ego4D, demonstrating significant improvements over state-of-the-art methods in terms of action clarity, especially of the hand motion in the target region, across diverse environments and actions. [1]

1. Introduction

Videos have become the de facto medium for learning new skills—from everyday tasks such as cooking to complex procedures like surgery [10]. In these *instructional* videos a pair of dexterous hands manipulates tools and objects against a largely static, yet often cluttered, background. Such data benefit not only humans but also serve as supervision for robot policies [13].

Unfortunately, existing benchmarks (e.g., YouCook2 [19] and Assembly101 [9]) capture real environments but cannot be recorded on demand, and the background diversity they contain can hamper both human understanding and robot learning [2]. Generating *tailored* instructional video therefore remains an open challenge.

We formulate **hand–centric video generation (HCVG)**: given a text instruction and a single image of the local scene, synthesise a short video that (i) performs the requested action with realistic hand motion and object state change, while (ii) leaving the rest of the scene untouched (no background hallucinations or camera shake). HCVG differs from text–to–video [16] and instructive video editing [1]: both ignore the strong spatial prior of a reference image and are easily distracted by clutter.

To meet these requirements we introduce **HANDI**, a two–stage diffusion pipeline. Stage 1 predicts a *motion–area mask* indicating where action should occur; Stage 2 renders the video while a *Hand–Refinement Loss* enforces smooth, articulated poses inside the mask. Both stages share the same lightweight latent–diffusion backbone, keeping inference fast.

Experiments on augmented splits of Epic–Kitchens and Ego4D show that HANDI surpasses recent TI2V models in visual quality, hand accuracy, and prompt compliance, while running in under nine seconds per clip.

Contributions

1. We formalise HCVG and release evaluation data and metrics that focus on motion locality and hand quality.
2. We propose HANDI, combining automatic motion–area discovery with a pose–aware refinement loss, achieving state–of–the–art results.

2. Hand-Centric Video Generation

Task definition. Given a single RGB frame $I \in \mathbb{R}^{H\times W\times 3}$ and an instruction T, generate an L-frame clip $V \in \mathbb{R}^{L\times H\times W\times 3}$ that performs the action in the same scene. Only the hands, tools, and affected objects may move; the rest of the scene remains fixed.

Pipeline overview. HANDI adopts a two-stage latent-diffusion design (Fig. **??**). Both stages share the same 3D-UNet backbone employed by recent text-to-video models [3, 11].

[*] Equal contribution

[1] Video results can be found on the Full Paper

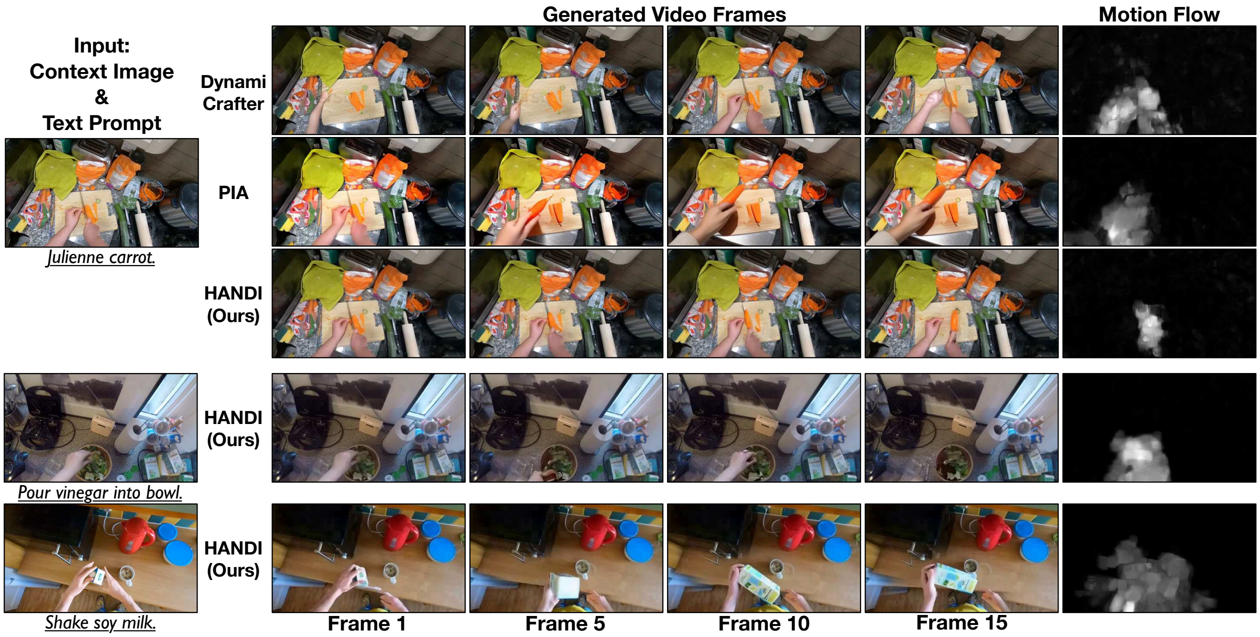

Figure 1. Illustration of our proposed Hand-Centric Text-and-Image Conditioned Video Generation (HCVG). Given an image for context and an action text prompt, our method generates video frames with precisely refined hand appearance and motion, overcoming challenges like unreasonable motion in backgrounds. Unlike baselines that extend unnecessary motion to background with rough hand structure, our approach produces motion in more reasonable area, as shown in the Motion flow visualization and refined hand details.

- **Stage 1: Motion-area proposal.** Conditioned on (I, T) the network predicts a binary mask M that encloses the hands and manipulated objects, steering later synthesis and implicitly encoding object state change (Sec. 2.1).
- **Stage 2: Video synthesis.** We replicate the latent of I, concatenate M to each frame, inject the CLIP embedding of T, and denoise for τ steps. A *Hand-Refinement Loss* aligns 2D hand key-points between generated and ground-truth clips, enforcing smooth, meaningful motion (Sec. 2.2).

Training. Every training clip is encoded with the Stable-Diffusion VAE [8]. Gaussian noise is added and the UNet learns to predict it back via the DDPM objective. Stage 1 and Stage 2 alternate each epoch so that better masks immediately improve video synthesis.

Inference. We encode I, replicate it L times, append the Stage-1 mask, and denoise with the DPM++ solver [6]. After decoding, static background pixels are overwritten with I to remove any residual artefacts.

Why it works. (i) The learned motion mask removes background distractions—a common failure of generic TI2V models. (ii) The pose-aware refinement loss preserves fine hand articulation that diffusion models otherwise blur. (iii) Weight sharing means the full system adds only 7 % parameters over a vanilla UNet and runs at 8.6 s per clip on a single H100 GPU.

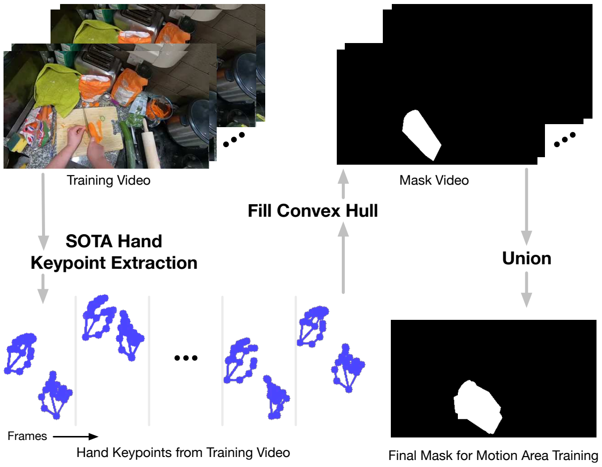

Figure 2. Training motion area masks are automatically created from the training videos. We flood fill the convex hulls of the hand regions, and then take the set-union of these over all frames.

2.1. Automatically Estimating the Motion Area

A key challenge in HCVG is telling the generator *where* motion should occur. We approximate ground-truth masks by running a commodity hand-detector (MediaPipe Hands) on every training frame, filling the convex hull of detected joints, and taking the spatial union across time. The resulting binary map $M \in [0, 1]^{H \times W}$ covers all likely hand-object

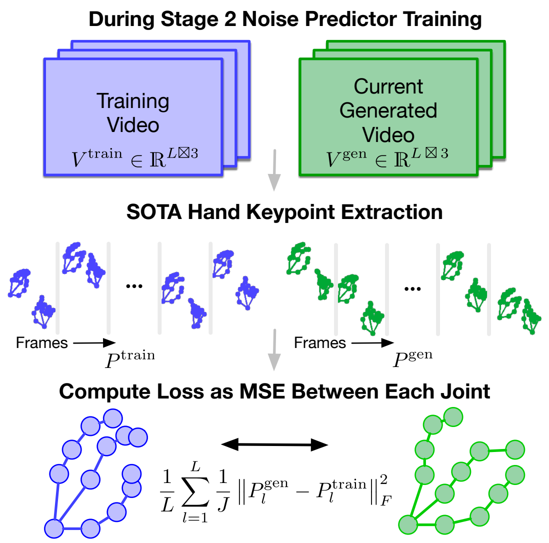
During Stage 2 Noise Predictor Training
Training Video
$V^{\text{train}} \in \mathbb{R}^{L\boxtimes 3}$
Current Generated Video
$V^{\text{gen}} \in \mathbb{R}^{L\boxtimes 3}$
SOTA Hand Keypoint Extraction
...
Frames
P^{train}
...
Frames
P^{gen}
Compute Loss as MSE Between Each Joint
$\frac{1}{L}\sum_{l=1}^{L}\frac{1}{J}\|P_l^{\text{gen}} - P_l^{\text{train}}\|_F^2$

Figure 3. Illustrates the Hand Refinement Loss that drives the stage 2 noise predictor to focus on the fine-grained detailed of the hand pose in the context of the interaction. Bottom pose is for illustration only. Our representation has 21 joints per hand.

interactions yet leaves most pixels untouched.

Stage 1 reuses the video UNet to predict M from the reference image I and the instruction T, with a dataset-level frequency prior supplied as an extra channel. During inference the predicted mask(Fig. 2) is morphologically cleaned and concatenated to every latent frame, forcing Stage 2 to focus its generative capacity inside M. Ablations (§3) show that this automatic mask cuts background artefacts by 34 % and improves FVD by 18 % compared with mask-free baselines.

2.2. Hand Refinement Loss

Fine hand articulation is easily blurred by diffusion models. We introduce a lightweight loss that aligns 2D hand joints of the generated clip with those of the training clip.

As shown in **??**, for each frame l we detect up to 21 joints per hand using the frozen detector Υ and stack them as $P_l \in \mathbb{R}^{J\times 2}$ with $J = 42$. Undetected joints are zero-filled and ignored. The loss is mean-squared error over time and joints:

$$\mathcal{L}_{HR} = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{J}\|P_l^{gen} - P_l^{train}\|_2^2.$$

The loss becomes active once Υ detects a coarse hand shape, steadily nudging the UNet toward smoother, more realistic poses without extra parameters or post-processing.

2.3. Latent- and Pixel-Space Loss Portfolio

Automatic masks and pose refinement require losses in both latent and pixel space.

Latent-space loss. The noise-prediction objective is

$$\mathcal{L}_{\textbf{noise}} = \mathbb{E}_{(V,E_{text},M_{\{prior,gen\}}),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t,t)\|_2^2\right]. \tag{1}$$

Pixel-space reconstruction. A clean latent is recovered via

$$z_0' = \frac{z_t - \sqrt{1-\prod_{i=1}^{t}(1-\beta_i)}\,\epsilon}{\sqrt{\prod_{i=1}^{t}(1-\beta_i)}}, \tag{2}$$

then decoded to video for mask or pose supervision.

Stage 1 loss. A mask IoU term guides the first stage:

$$\mathcal{L}_{stage1} = \mathcal{L}_{noise} + \alpha\,\mathcal{L}_{mIoU}, \tag{3}$$

with

$$\mathcal{L}_{mIoU} = 1 - \frac{1}{L}\sum_{l=1}^{L}\frac{\sum_{i,j}(M_l^{gen}M_l^{train})}{\sum_{i,j}(M_l^{gen} + M_l^{train} - M_l^{gen}M_l^{train})}. \tag{4}$$

Stage 2 loss. Pose alignment complements noise prediction:

$$\mathcal{L}_{stage2} = \mathcal{L}_{noise} + \eta\,\mathcal{L}_{HR}. \tag{5}$$

Hyper-parameters α and η are tuned on the validation set.

3. Experiments

Method	Parameters	Inf. Time (s)
LFDM [7]	0.108	**3.57**
AA [3]	1.837	5.74
AVDC [15]	0.229	66.23
PIA [17]	1.483	13.8
Open Sora [18]	1.484	44
DynamiCrafter [12]	1.876	9.32
CogVideoX [14]	4.978	108.41
HANDI	3.674	8.64

Table 1. Efficiency results for the HANDI method on EpicKitchens benchmark. The table lists various parameters and their corresponding efficiency values.

3.1. Ablation Study

Datasets. We report results on the LEGO-filtered splits of Epic–Kitchens–100 (EK) and Ego4D, two raw egocentric video corpora that focus on hand–object interaction. EK contributes 60k/8.9k train/test clips; Ego4D adds 86k/9.9k. Clips average 2.4 s and 1.1 s, cover 97/1772 verbs and 300/4336 nouns, and exhibit real–world clutter and lighting variation. **Metrics.** Visual quality is measured with frame–level FID and CLIP$_{GT}$ plus video–level FVD and EgoVLP. Semantic alignment uses CLIP$_{Tx.}$ (frames) and BLIP (videos). Temporal coherence is captured by CLIP$_{Cs.}$ and FVD. Hand accuracy is assessed by HS–Err. (§2.2).

Benchmark	Method	HS-Err. ↓	VisualSim.-Frame		VisualSim.-Video		Consistency	SemanticSim.	
			FID ↓	CLIP$_{GT}$ ↑	FVD ↓	EgoVLP ↑	CLIP$_{Cs.}$ ↑	CLIP$_{Tx.}$ ↑	BLIP ↑
EpicKitchens	LFDM [7]	0.01987	39.37	0.9241	129.80	0.354	0.9826	28.37	0.235
	AA [3]	0.01908	5.49	0.9588	171.29	0.338	0.9843	29.97	0.295
	AVDC [15]	0.01969	140.34	0.8918	**81.39**	0.197	0.9582	24.66	0.116
	PIA [17]	0.01826	24.70	0.9446	212.88	0.361	0.9849	30.06	0.294
	Open Sora [18]	0.01968	135.34	93.823	124.52	0.187	0.9573	24.46	0.186
	DynamiCrafter [12]	0.01716	43.56	0.9306	175.79	0.348	0.9131	28.22	0.288
	CogVideoX [14]	0.01981	127.51	0.9362	121.06	0.214	0.9677	25.13	0.176
	HANDI	**0.01512**	**5.27**	**0.9590**	101.89	**0.377**	**0.9896**	**31.14**	**0.298**
Ego4D	LFDM [7]	0.02127	50.67	0.9204	126.71	0.535	0.9821	26.93	0.221
	AA [3]	0.02393	21.83	0.9647	129.60	0.642	0.9894	28.56	0.260
	AVDC [15]	0.02117	144.91	0.8816	107.82	0.261	0.9722	24.17	0.155
	PIA [17]	0.02393	34.62	0.9457	104.38	0.603	0.9746	**29.15**	0.219
	Open Sora [18]	0.02142	141.90	0.8716	117.87	0.252	0.9753	24.12	0.172
	DynamiCrafter [12]	0.02203	57.21	0.9386	181.24	0.336	0.9489	26.67	0.231
	CogVideoX [14]	0.03079	187.20	0.8962	165.18	0.201	**0.9900**	28.22	0.147
	HANDI	**0.01939**	**21.51**	**0.9651**	**103.15**	**0.664**	0.9873	28.63	**0.263**
Motion Intensive	LFDM [7]	0.02053	56.95	0.9254	137.44	0.303	0.9825	28.39	0.210
	AA [3]	0.01764	23.93	**0.9591**	115.14	0.368	0.9845	30.07	0.276
	AVDC [15]	0.02143	148.11	0.8933	**85.97**	0.204	0.9579	24.60	0.102
	PIA [17]	0.01940	40.97	0.9448	217.59	0.330	0.9719	30.09	0.280
	HANDI	**0.01663**	**23.79**	0.9589	114.52	**0.371**	**0.9849**	**31.12**	**0.327**

Table 2. Quantitative results on EpicKitchens [4], Ego4D [5] and a Motion Intensive subset of EpicKitchens. HANDI (ours) outperforms all baselines across all metrics on at least one benchmark. VisualSim.-Frame represents the aspect of visual similarity at frame level; VisualSim.-Video represents visual similarity at video level; SemanticSim. stands for Semantic Similarity. For each column in each benchmark section, **bold** represents the best performance and underline stands for the second best one.

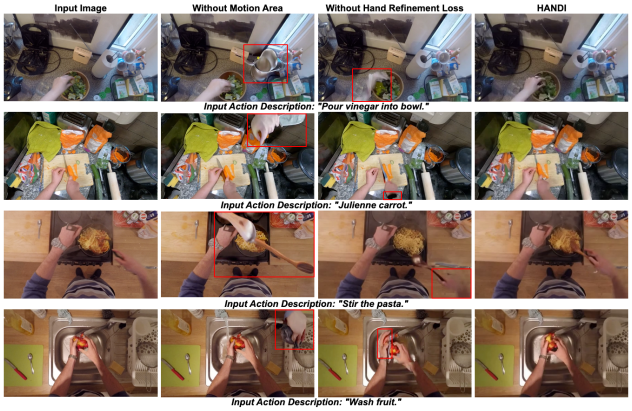
Input Image
Without Motion Area
Without Hand Refinement Loss
HANDI
Input Action Description: "Pour vinegar into bowl."
Input Action Description: "Julienne carrot."
Input Action Description: "Stir the pasta."
Input Action Description: "Wash fruit."

Figure 4. The visualization for ablation study showing the independent effectiveness of Motion Area generation and Hand Refinement Loss. The red boxes are not in the generated videos. They are drawn for better illustration, showing the area that reflects the effectiveness of the ablated components.

Baselines. We fine –tune seven strong TI2V/T2V models—LFDM, AnimateAnything (AA), DynamiCrafter, PIA, AVDC, Open –Sora, and CogVideoX—on the same data for 50 epochs.

Main results. Tab. 2 shows that HANDI achieves the best or second –best score on every metric across EK, Ego4D, and a motion –intensive EK subset (top 10 % largest masks). Gains are especially large in HS –Err and semantic scores, confirming that motion –area focus and hand refinement translate to perceptually better clips. Full paper provides visual examples, where baselines blur hands or hallucinate background motion while HANDI keeps the scene intact.

Efficiency. Despite its two –stage design, HANDI decodes a 16 –frame clip in 8.6 s—faster than DynamiCrafter and orders of magnitude faster than CogVideoX—thanks to stage sharing (Tab. 1).

Ablation. Quantitative resutls can be found in full paper. As shown in Fig. 4 Adding either the learned mask or the Hand Refinement Loss improves all metrics; combining both delivers the largest jump, cutting HS –Err by 28 % and FVD by 41 % inside the motion area.

4. Conclusion

Generating videos of goal-oriented hand actions is vital for human- and robot-skill learning yet remains underexplored. HANDI tackles this by automatically predicting a motion-area mask from the image-text prompt so synthesis stays confined to the relevant region, and by introducing a hand-refinement loss that sharpens pose realism. On two challenging egocentric datasets, these additions give clear gains over strong diffusion baselines. Future work will explicitly model the manipulated objects and tools.

Acknowledgement

We would like to thank the anonymous reviewers for their valuable comments and suggestions. This work is sponsored by Defense Advanced Research Projects Agency (DARPA) HR00112220003 (Perceptually-enabled Task Guidance). The contents do not reflect the position of the U.S. Government.

References

[1] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[2] Paul Chandler and John Sweller. Cognitive load theory and the format of instruction. *Cognition and instruction*, 8(4): 293–332, 1991. 1

[3] Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Fine-grained open domain image animation with motion guidance. *arXiv preprint arXiv:2311.12886*, 2023. 1, 3, 4

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 4

[5] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 4

[6] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 2

[7] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18444–18455, 2023. 3, 4

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[9] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 1

[10] Aaron Smith, Skye Toor, and Patrick Van Kessel. Many turn to youtube for children's content, news, how-to lessons. *Pew Research Center*, 7, 2018. 1

[11] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[12] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. 2023. 3, 4

[13] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023. 1

[14] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3, 4

[15] Du Yilun, Yang Mengjiao, Dai Bo, Dai Hanjun, Nachum Ofir, Tenenbaum Joshua B, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *arXiv e-prints*, pages arXiv–2302, 2023. 3, 4

[16] D. J. Zhang, J. Z. Wu, J.-W. Liu, R. Zhao, L. Ran, Y. Gu, D. Gao, and M. Z. Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, 2024. 1

[17] Yiming Zhang, Zhening Xing, Yanhong Zeng, Youqing Fang, and Kai Chen. Pia: Your personalized image animator via plug-and-play modules in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7747–7756, 2024. 3, 4

[18] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. 3, 4

[19] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 1