Populate-A-Scene: Affordance-Aware Human Video Generation

Mengyi Shan[1,2], Zecheng He[2], Haoyu Ma[2], Felix Juefei-Xu[2], Peizhao Zhang[2],
Tingbo Hou[2], Peter Vajda[2], Ching-Yao Chuang[2]
[1]University of Washington, [2]Meta

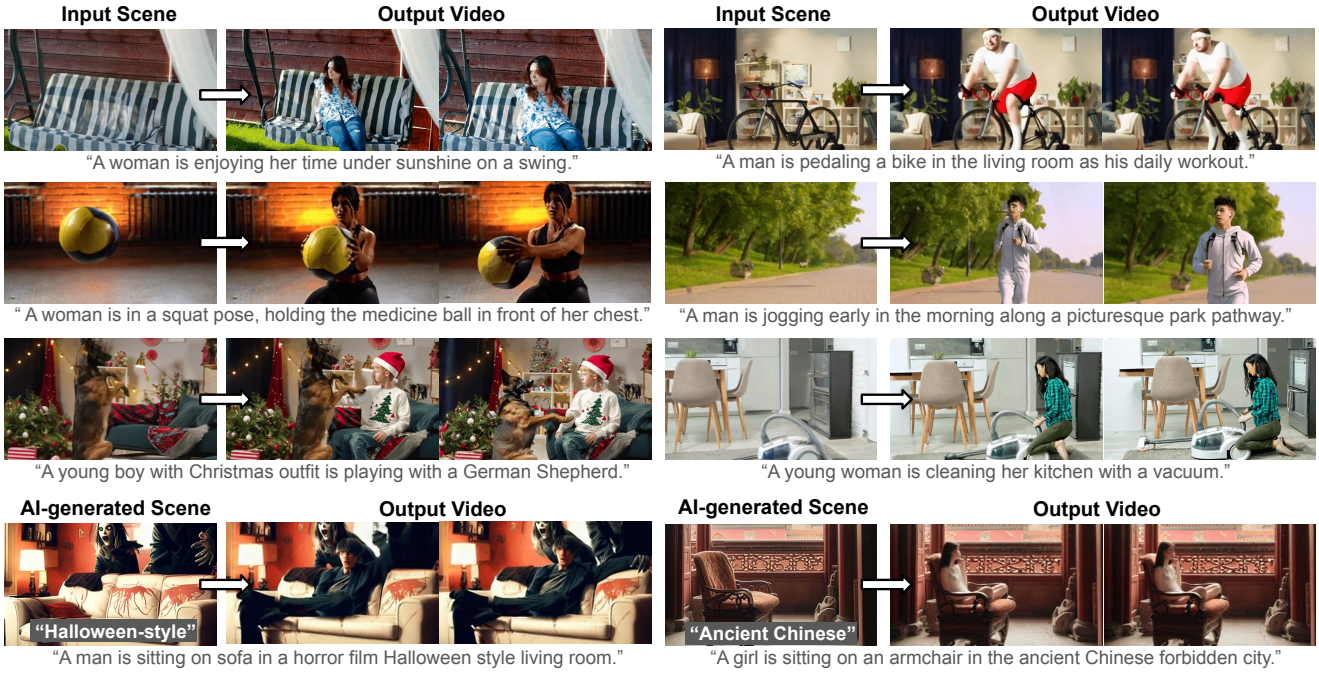

Figure 1. We repurpose a text-to-video generation model as a human-world interaction simulator. Given a scene image and a prompt, our model inserts a person into the environment and generates a video of them naturally interacting with the scene. The scene can be real images (top) or synthesized by image generative models (bottom). Notably, there is no need for any mask, location bounding boxes, or pose sequences to guide the human insertion – our method takes care of affordance prediction entirely within the video model.

Abstract

Can a video generation model be repurposed as an interactive world simulator? We explore the affordance perception potential of text-to-video models by teaching them to predict human-environment interaction. Given a scene image and a prompt describing human actions, we fine-tune the model to insert a person into the scene, while ensuring coherent behavior, appearance, harmonization, and scene affordance. Unlike prior work, we infer human affordance for video generation (i.e., where to insert a person and how they should behave) from a single scene image, without explicit conditions like bounding boxes or body poses. An in-depth study of cross-attention heatmaps demonstrates that we can uncover the inherent affordance perception of a pre-trained video model without labeled affordance datasets.

1. Introduction

Scaling data, compute, and model parameters in video generation models presents a promising avenue for developing highly capable simulators that can accurately replicate complex physical worlds [8, 48], complete with diverse objects and people that interact and coexist within them. Nevertheless, humans are not merely passive observers, but rather active participants in the world. Human understanding of affordance [21, 34, 45] enables purposeful engagement with surroundings and adaptive behavior by recognizing potential actions afforded by an object's physical properties. It remains unclear whether video generation models can interpret and replicate intricate semantic aspects of the world, such as contextual understanding and dynamic behavior, beyond the capabilities of traditional graphics pipelines.

Affordance, or "opportunities for interaction" [21], has inspired extensive research in vision and psychology. Traditional affordance prediction relies on data-driven approaches using 3D information [25], specifically labeled datasets [12, 15, 19, 23, 63], or one-shot large foundational models [38]. However, these methods rely on domain-specific annotations, which are challenging to obtain. In contrast, recent advancements in generative models offer the potential to create realistic human-scene media content using vast amounts of in-the-wild media data. [36], for example, predicts a human's pose and appearance in a scene but is restricted to static images with a given position mask.

In this work, we demonstrate that throughout the intricate process of video generation, the model learns to generate human activities and motions that adhere to the affordance constraints dictated by the physical world. To better study affordance modeling, we propose augmenting a pre-trained text-to-video model [48] with an additional scene conditioning branch. This modification formulates the problem as a conditional video generation task: given a scene represented by an image, the model is tasked with introducing natural human motion and interactions to the scene. We discover that pre-trained video generation models can rapidly adapt to this new task by fine-tuning on a relatively small-scale scene conditioning dataset. We then validate the affordance perception abilities through an extensive study of the cross-attention feature heatmaps, a key module that enables the model to follow language prompts.

Unlike prior work, our model does not require input masks, bounding boxes, or pose sequences to specify regions or patterns of human behavior, which makes it an interaction *simulator* that reasons about semantics and affordance properties in the scene, instead of merely a human *renderer* that turns given pose signals into pixels. During inference, the model can process a wide array of environment-action combinations to generate diverse interactive videos, not limited to interaction with the single, salient object in a complicated scene. Fig. 1 demonstrates results of our model without aggressive cherry-picking. In particular, the last row of Fig. 1 illustrates a "movie studio" pipeline where input scenes are generated using a text-to-image model [14], and our model seamlessly integrates actors into these scenes without requiring 3D capture. Our results lower the barrier for amateur AI video creators by eliminating the need for explicit body poses signals, as they are required in most AI human video models but challenging to synthesize.

In short, this work makes the following contributions:

- We address affordance-aware human video generation, where we generate video of subject(s) interacting with a given environment image, *without* telling the model where the subject(s) are and how their poses look like.
- We apply the dual-stream conditioning mechanism with a minimal grounding module to model affordance, and thus reveal the affordance capabilities of video generation models through in-depth analysis.
- We demonstrate our model's ability to generalize across diverse environments and actions through a synthetic benchmark created with vision-language models.

2. Related Works

Text-to-video generative models. Text-to-video generation aims to synthesize plausible, temporally coherent, and optionally condition-aligned video sequences from textual prompts. Recent rapid advancements in text-to-video models have been phenomenal [3, 4, 7, 11, 18, 20, 26, 27, 48, 55, 61]. Current work explores replacing the traditional U-Net with a Transformer [60] architecture [7, 24, 44, 48], inspired by the promising text-to-image generative results from DiT [47]. We start from a pre-trained Transformer-based text-to-video model Movie Gen [48] and explore its ability to perceive affordance through minimal fine-tuning on human-scene interaction data. Some text-to-video tasks augment the model with an image as a starting frame and use prompts to describe the style or motion in the video [22, 50, 72]. Our task differs in that we give the model an empty scene frame that is not supposed to appear in the video, but provides a "playground" for population.

Human video generation. Human video generation evolves alongside rapidly advancing generic video generative models. Generating realistic human content is inherently challenging due to complex body topology, strong priors on interaction plausibility, and audiences' sensitivity to even minor artifacts. Existing methods use motion guidance to improve video faithfulness, leveraging signals such as OpenPose [28, 62], DensePose [31, 67], SMPL [74], or a driving video [71]. These works focus on human video generation with the subject as the sole salient element, without modeling human-environment interaction. Our work differs in that we reason about natural human-scene interaction without compromising human quality. Notably, our method requires no auxiliary conditions such as position bounding boxes [36, 56] or motion sequences, relying instead on the internal affordance inference potential of video models.

Human-scene interaction modeling. A fundamental task in human-environment modeling is motion prediction in 3D scenes [32, 39, 64]. Related work in 2D explores interaction image and video generation from a scene, mostly using some location or body pose signals as condition [29, 46, 53, 69]. Kulal [36] and Cao [9] claim to predict affordance by inserting a human subject into a static scene, but they require a bounding box as input indicating the position. Shan [54] insert moving humans into a street scene, but restrict actions to predefined walking motions. Singh [56] predict fine-grained masks for human insertion based on scene and text descriptions but do not explicitly model human-scene interaction. Jin [30] builds on similar

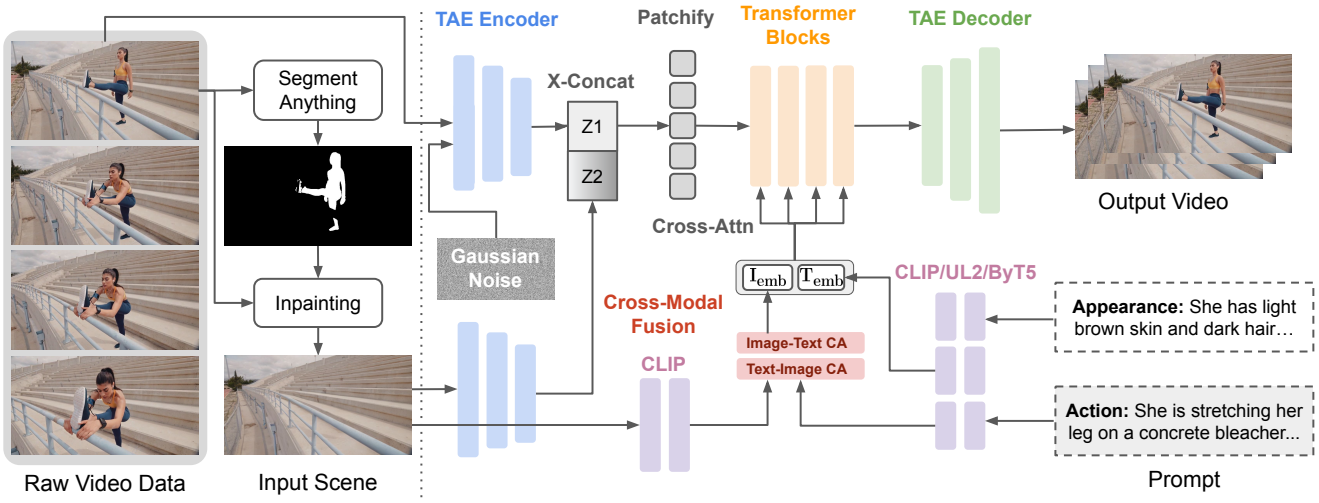

Figure 2. We start by removing humans from raw frames to create synthetic empty-scene and human-video data pairs. We employ a dual-conditioning mechanism, using channel concatenation and cross-attention, to condition the T2V model on an additional scene image. We design a fusion module to facilitate interactions between image and action-text features while locating the desired action position. The fine-tuning pipeline trains a Transformer architecture with flow matching.

ideas as ours, but focus on static images with non-human objects, which in nature lack intricate interactive dynamic behaviors. Our work instead requires no semantic priors for where and how human-scene interaction occurs.

Affordance. Psychologist J.J. Gibson defines *affordance* as the possibilities an environment offers an individual [21, 45] and views affordance perception as essential to socialization. Inspired by this concept from cognitive psychology, computer vision research explores scene and object affordance prediction [13, 57] and affordance learning from human-scene interactions [15, 19, 63]. Inspired by this ongoing discussion, we study how generative models perceive affordance by creating interactive videos.

3. Preliminary: Text-to-Video Generation

In this work, we leverage Movie Gen [48] as our base text-to-video model. Due to resource limitations. we conduct our experiments on a 4B-parameter model that generates 128-frame 256p videos as a proof of concept, instead of training the official 30B-parameter model that operates at 1080p. We highlight key architectural and training aspects incorporated into our experiments in this following section. Refer to the supplementary material for more details.

Temporal autoencoder. Our model encodes RGB videos and images into a learned spatiotemporally compressed latent space using a Temporal Autoencoder (TAE) and generates videos in this space. The TAE encoder is designed by inflating the image autoencoders in [51], adding an 1D temporal convolution after each 2D spatial convolution and an 1D temporal attention after each spatial attention.

Video generation backbone. The model generates videos within a learned latent space as defined by the TAEs. The latent video code is segmented into patches via a 3D convolutional layer [16], then flattened into a 1D sequence as input to the generation backbone. The generation backbone consists of Transformer [60] blocks with cross-attention modules inserted between self-attention and feed-forward networks, enabling text conditioning via text prompt embeddings. The model employs UL2 [58], ByT5 [68], and Long-prompt MetaCLIP [66] as text encoders, enabling both semantic- and character-level text understanding.

Flow matching. The model is trained with Flow Matching [5, 41], which iteratively transforms a prior Gaussian distribution into a sample from the target data distribution. During inference, an ordinary differential equation (ODE) solver transforms random noise into video latents. We use this training and inference framework for all experiments.

4. Affordance-Aware Video Generation

Our full pipeline is illustrated in Fig. 2. We define the problem, explain data processing and model architecture below.

4.1. Task Definition

Let I be an image of a static scene, and let T_h and T_a be text prompts describing a human's appearance and action. We generate a video V that depicts the given scene I with an inserted human matching the appearance described by T_h and performing the action in T_a. During training and inference, we provide no explicit guidance for the human's position or pose in the scene, allowing the generative model full freedom to position the action, simulate body movements, and render the video. Note that this is not image animation; the scene image serves only as a reference for the background appearance and the presence of semanti-

cally meaningful objects. We do not require the image to appear as a frame in the video, nor do we treat the scene as a static background for pasting the human without environmental animations or camera viewpoint changes.

4.2. Training Data

In this section, we explain our full data processing pipeline. Representative data samples are shown in Fig. 3.

Human filtering. We curate our dataset by selecting human-related videos from the ShutterStock [1] text-video dataset. We apply human detection to each middle video frame and retain only those with one or two detected persons.

Full body filtering. We apply OpenPose [10] to videos that pass the previous stage, retaining those where knees' keypoints are visible or face's height and width falls below a threshold to avoid half-body or close-up shots.

Pure background filtering. We compute the color variance of background pixels in the middle frame of each video, retaining only those exceeding a threshold of 200. We also scan video captions and exclude those containing keywords like "a pure green background." This helps eliminate studio-recorded videos that lack background interaction.

Human removal. We take the first and last frames of each video, with GroundingDINO [42] detecting the central human subject and language-guided SAM [33] segmenting the human mask. We dilate the mask by 50 pixels to fully cover the human region and apply a text-to-image inpainting model with the negative prompt "human" for removal. For two-person videos, we remove one person at a time, creating two data samples from a single video. This results in a training dataset of `(text, image, video)` tuples representing `(action, scene, interaction)`, including 217,530 samples for single-person data and 29,700 for two-person data. We handpick 300 samples per category for validation and detail the post-processing steps for the synthetic validation benchmark in Sec. 6.1.

Prompt rewriting. We use LLaMA 3 [17] to rewrite video captions, separating out human-related prompts (T_a and T_h) and removing sentences that pertain solely to the background. This allows the model to learn background information purely from the visual modality rather than text, promoting multimodal information fusion.

4.3. Conditioning Mechanism

During fine-tuning, we aim to keep the original structure as much as possible, while exploring conditioning strategies to unfold a text-to-video model's innate ability on perceiving affordance from a scene image. We discuss key strategies to condition the model on an additional image input.

Masked latent concatenation. To maintain background consistency with the given image I, we concatenate the image latent Z_1 with the noisy latent Z_2 along the channel dimension before feeding them into the Transformer's backbone. Since our model is not an image animation model, we allow environmental updates driven by both the action prompt T_a and natural effects such as camera movements. To achieve this, we progressively add Gaussian noise to the conditional image latent Z_2 with a temporal scaling factor of $\gamma = 0.8$, weakening control as the video progresses until the last frame is fully masked. This decay-based control strategy ensures the scene initially matches the given image while allowing camera movement and human interaction to modify scene elements over time.

Figure 3. Representative samples of our dataset. Top row shows single-person data, while the bottom row shows double-person data. Within each row, the top figure presents the raw first frame of the video, while the bottom figure(s) show the result after detecting and removing humans from the scene. The background remains unchanged while the subject is removed.

Fused text-image feature enhancer. We augment the cross-attention conditioning branch with a fusion module that practices mutual attention on embeddings of image and action text, drawing inspiration from [40, 42]. Following the original model, we concatenate three types of text embeddings (ByT5, UL2, MetaCLIP) to form a unified textual representation and use the CLIP image feature map before pooling it into a spatial-aware image embedding. We apply deformable self-attention [65] to enhance image features and standard self-attention for text features. To promote cross-modal alignment, we introduce separate image-to-text and text-to-image cross-attention layers for feature fusion. This fusion module enhances the text-to-video model's grounding ability, allowing it to 'locate' corresponding action regions within the image. We concatenate the fused image embedding with raw textual embeddings and input them into each Transformer block in the text-to-video model via cross-attention, as in Sec. 3.

Controlled guidance scale. Following the practice of InstructPix2Pix [6], we leverage a controlled multi-scale

guidance mechanism to control the strength of background scene image and action prompt. A higher image strength preserves scene consistency, while a higher text strength emphasizes human action and promotes plausible environmental updates. Training with dummy condition images helps maintain the pre-trained model's text-to-video capability and prevents overfitting to a specific dataset domain.

4.4. Implementation Details

We use the base text-to-video model Movie Gen [48] with 4B parameters, as described in Sec. 3. We train with landscape 256p, 16 frames per second, eight seconds per video. We fine-tune the full model with the text encoders frozen. We use a per GPU batch size of 1, and a learning rate of 1e-5. The training takes two days on 32 H100 GPUs.

5. Unveiling Implicit Affordance Capability

We comprehensively analyze the implicit affordance modeling capabilities of our proposed model. In Sec. 5.1 we justify that affordance perceiving information can be unveiled by investigating the the cross-attention modules, specifically which processes and regulates the CLIP text conditions. In Sec. 5.2 we apply our model on a real-world affordance prediction dataset. As a preliminary, the primary objective of cross-attention is to select appropriate values $\mathbf{V}$ using the attention scores $\mathbf{S}$ determined by

$$\mathbf{S} = \text{softmax}(\mathbf{Q}\mathbf{K}^T/\sqrt{d}) \in \mathbb{R}^{n\times m}$$

Here, $\mathbf{Q} \in \mathbb{R}^{n\times d}$ represents the projected and flattened intermediate diffusion features. $\mathbf{K} \in \mathbb{R}^{m\times d}$ and $\mathbf{V} \in \mathbb{R}^{m\times d}$ are the projected features of the input text embedding. The attention map $\mathbf{S}$ provides a physical interpretation where each entry (i, j) indicates the saliency of interaction between a spatial location i and a token j in the prompt. This saliency reflects how strongly a particular spatial feature is influenced by or associated with a specific word, guiding the model in generating contextually relevant outputs.

5.1. Predicting Affordances via Cross-Attention

We explore the implicit affordance reasoning capability of video modela by visualizing the j-th entry of the attention map $\mathbf{S}$ where the j-th token corresponds to an action-related term in the prompt. For example, given the input prompt "a woman holding the rope and riding a horse", we focus on visualizing the attention heatmap associated with the verb "holding" and "riding".

The top half of Fig. 4 shows the attention scores of the pre-trained T2V model. Trained exclusively on text-video pairs, the model exhibits a reasonable ability to perceive affordances while generating high-quality, faithful content. The heatmaps align well with action regions, highlighting the model's ability to associate generated spatial features

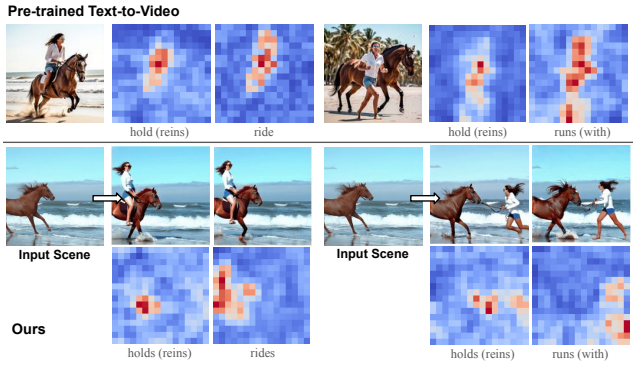

Figure 4. Cross-attention maps of the video models. Top half is the pre-trained model where the presented scene is generated by the model, and bottom half is our scene-conditioned model with a real image as input. Attention is averaged across timesteps.

with actions. However, this correlation appears to be a by-product of video model training, as the heatmaps are conceptually intermediate steps in *synthetic* video generation.

Building on this observation, we propose that conditioning the model on an additional scene enables it to perceive affordances in a *given, real* image. The bottom half of the figure shows that the model accurately identifies action locations in input images and the specific environmental elements involved in the interaction. Our heatmaps reveal internal affordance knowledge, capturing interaction opportunities in real images rather than merely serving as by-products of synthetic content generation.

5.2. Real-World Affordance Prediction Experiment

We subsequently analyze our model's affordance perception using classical 2D affordance detection datasets. We filter the Purpose-Driven Affordance (PAD) dataset [43], retaining only images with no person and action verb-object pairs representing human actions (e.g., push, hit) and discarding passive object verbs (e.g., contain). This leaves us with 24 action verb categories, totaling 235 images with corresponding affordance masks. We create the prompts based on the affordance verb with LLaMA [17], and pass in the image and prompts as inputs for our model.

In Fig.5, we present heatmap visualizations, derived similarly to those in Sec.5.1. We also compute the spatial accuracy (defined as pixel-wise IoU) between the binarized attention map and the ground-truth affordance mask across different layers and diffusion inference steps. We observe slightly higher scores in the initial layers, likely because the model processes semantic information early in generation. Even in the early steps, our model consistently predicts affordance through attention features. Accuracy decreases in later steps as the model shifts from perceiving high-level semantics to refining details of generated content. Peaks in the attention heatmap gradually transition from interaction regions to human content. The spatial alignment

of heatmap and ground-truth affordance maps highlights the video model's ability to perceive affordance in alignment with real-world data. Our model even outperforms the ground-truth by predicting object parts relevant to each action rather than entire objects. For example, it identifies the seat of a bench where people sit, rather than its legs.

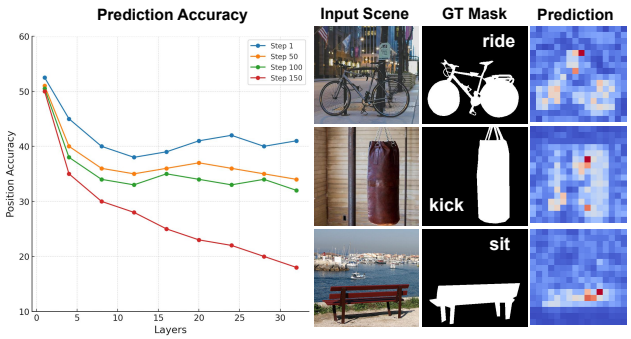

Figure 5. Affordance position accuracy across different steps and layers on a subset of PAD [43]. The attention scores indicate strong predictive ability, and visualizations show that our model accurately locates detailed affordance information.

6. Results

We present quantitative and qualitative results of our proposed affordance-aware human video generation models.

6.1. Evaluation Dataset

We aim to generate *diverse* actions interacting with more than one parts of the environment, even within a fixed scene. To address this, we curate synthetic prompt sets based on real scene images. Specifically, we use a pre-trained vision-language model to generate two prompts per scene by asking, "What might a person do in this scene?" This process yields an evaluation set of 300 images, each paired with one original and two synthetic prompts. These prompts emphasize different objects or positions within a complex environment, allowing us to assess whether our model's generative ability extends beyond central, salient objects. We repeat this process for two-person scenarios, prompting interactions with both the scene and the existing person. Fig. 6 illustrates our benchmark pipeline. We will release this evaluation benchmark for follow-up comparisons.

6.2. Baselines and Ablation

Baselines. To the best of our knowledge, there is no existing work on generating human videos in a scene without location or pose control. We therefore compare our methods with generic image/video editing and image-to-video solutions not tailored for humans. We compare with three image-based models: (1) **InstructPix2Pix [6]** where we directly apply an image editing model on the empty scene image with the prompts. (2) **Flux Editing** which trains instructional image editing on Flux [37]. (3) **Flux Inpainting** where we provide a groundtruth human mask as the inpainting position. We then compare with instruction-based video editing method (4) **AnyV2V [35]** where the scene is repeated for 2 seconds to a video, and then edited based on a prompt. We additionally compare with one open-source and two commercial video generation models (5) **CogVideoX [70]**, (6) **Runway Gen-3 [52]** and (7) **Luma AI Ray-2 [2]** where we apply image-to-video on the scene with a prompt. For (1), (2), (3), we attach a CogVideoX image animation model to the image results, exploring their potential of generating interaction videos in the same setting as ours. Note that we only do small scale visual comparison and human evaluation on (6) and (7) without quantitative metrics as they do not have a free API available.

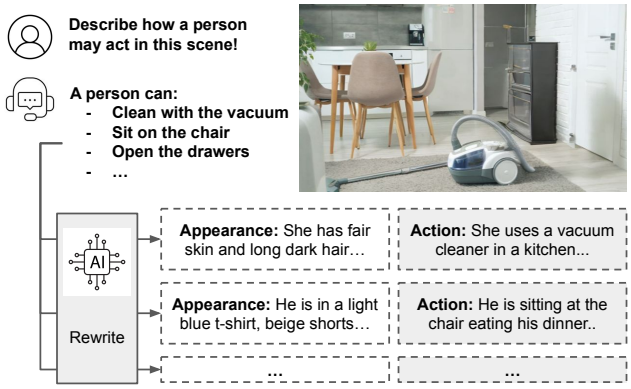

Figure 6. The synthetic action descriptions generated through our prompting process. We use a vision language AI agent to decide palusible actions in a scene, and rewrite the action into prompts.

Ablation studies. We compare with alternative designs of our model that remove key features including latent concatenation, fused cross-attention, and Gaussian noise decay. Due to space limits, parts of the ablation results are presented in the supplementary material.

6.3. Qualitative Evaluation

Human-scene interaction. Fig. 1 presents inserting a human into a scene based on an action prompt. The model maintains pixel-level scene consistency while placing the subject correctly without a predefined mask. The generated video features natural camera movements, object updates in response to human actions, and scene animations.

Diverse affordance. In scenes with complex layouts and multiple interaction possibilities, our model inserts subjects while accounting for diverse scene elements and action-affording objects. Fig. 7 illustrates how our model determines subject placement and imagines body poses for different actions (e.g., riding vs. standing beside a horse).

Human-human interaction. Fig. 8 shows adding a subject to interact with both the scene and an existing person

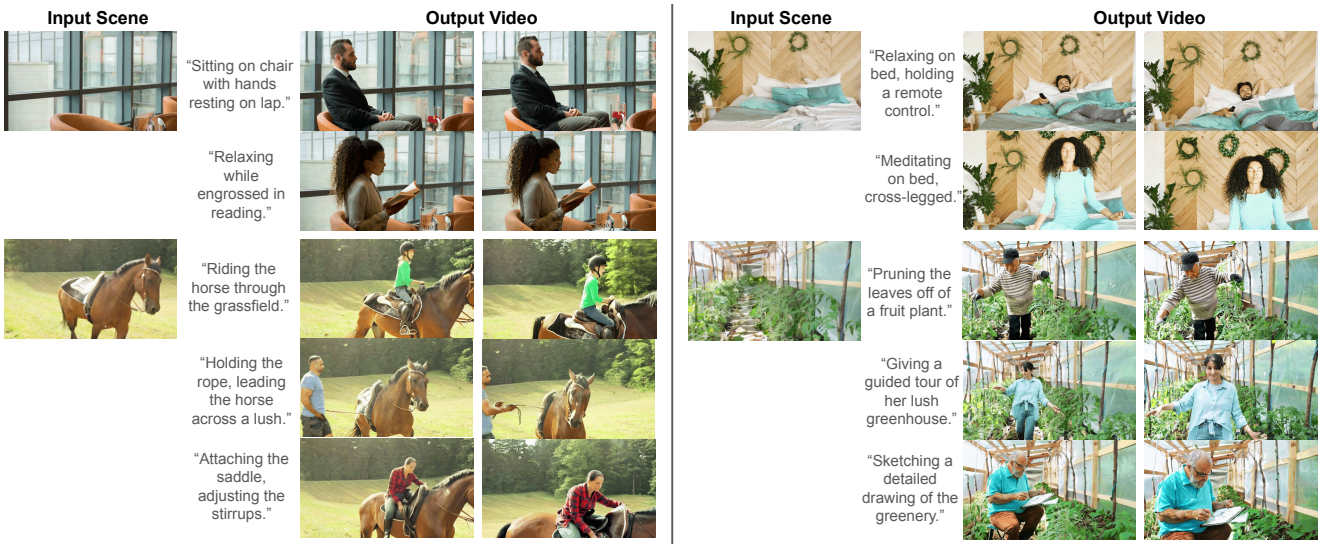

Figure 7. Our model generates diverse videos with multiple action prompts given the same scene. It identifies the correct way for an inserted subject to interact with the scene, and infers location, pose, action, spatial relationship without a pre-defined human mask prior.

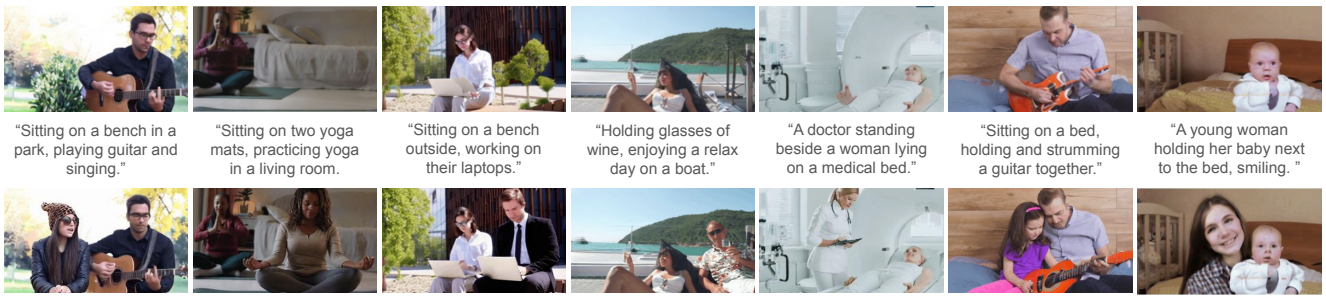

Figure 8. Our model is able to add an extra subject to a scene that contains one person. Here we consider the existing person as an organic part of the environment, and are able to synthesize interactions respecting both the background and the human in the scene. Top row is input scene image, middle row is the action prompt, and the bottom row is middle frame of the generated video.

who is considered a part of the scene.

Baseline comparison. Fig. 9 demonstrates that our model achieves the highest semantic alignment and visual fidelity. Instruction-based editing methods like InstructPix2Pix and AnyV2V [6, 35] generate distorted human bodies and misattribute prompt concepts (e.g., applying "pink" to the treadmill or curtain instead of clothing). Editing methods based on better image model Flux [37] does not preserve scene styles and generates cartoon videos. Flux [37] Inpainting distorts human bodies even when provided with an additional mask and fails to preserve pixel details in masked background regions (the yellow pillow disappears). Current best open-sourced and commercial image-to-video models like CogvideoX[70], Runway Gen-3 [52] and Luma Ray-2 [2] all misinterpret the treadmill's affordance, place subjects in the wrong direction, and hallucinate another treadmill on the left. Our models stand out by successfully preserving the background and simulating natural interactions between the subject and the treadmill.

6.4. Quantitative Evaluation

We evaluate our model based on human video faithfulness, text-video alignment, and action quality. This corresponds to three major quantitative metrics: (i) **FVD (Fréchet Video Distance)** [59], which quantifies the similarity between real and synthetic video embedding distributions. (ii) **CLIP** [49] similarity, which computes the average embedding similarity between the input prompt and each generated frame to assess prompt alignment. (iii) **Action Score**, computed by querying a pre-trained VQA model [73] with "What action is the person performing in this video?" and measuring the CLIP similarity between the recognized motion and the ground-truth action prompt. The Action Score helps isolate interaction accuracy by reducing the influence of appearance. For image-only baselines, we compute metrics on the animated video sequence using CogVideoX [70], one of the best open-sourced image animation models.

We quantitatively compare our model with baselines and

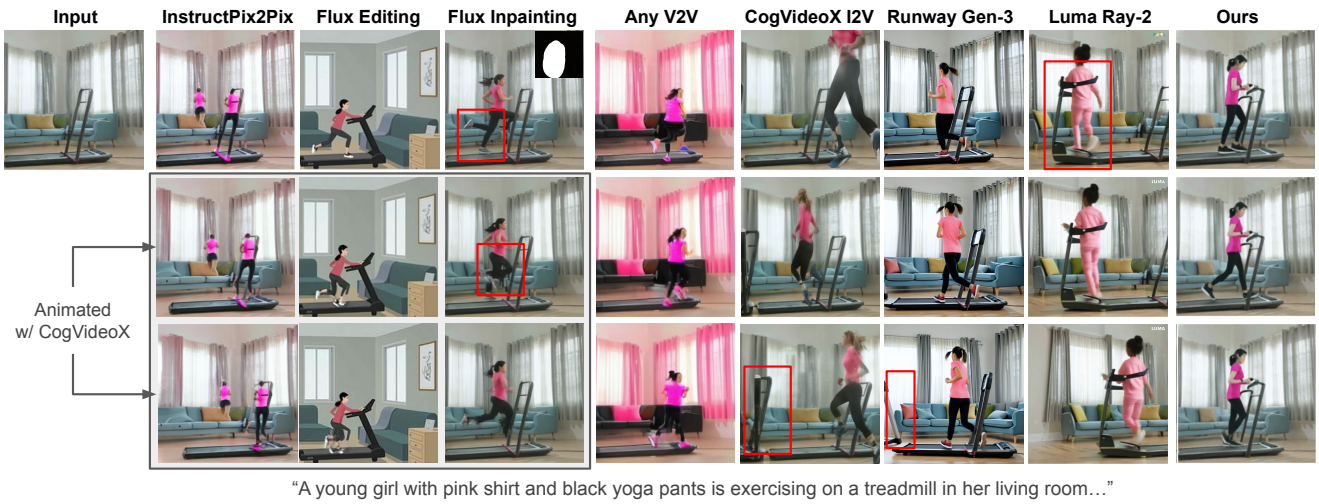

Figure 9. Comparison with baseline methods. Three rows are the first, middle and last frames for each method. The left three columns' models edit a static frame and animate it. The next four edit video directly. Note that Flux [37] Inpainting requires a user-defined mask as input, which eases the task and greatly assists the model in predicting human position. Yet, our model outperforms baselines in terms of human placing, motion simulation and appearance rendering. See video results and more comparison in the supplementary materials.

Table 1. Quantitative evaluation shows our method consistently outperforms baselines and ablation methods.

Model	CLIP ↑	FVD ↓	Action Score ↑
InstructPix2Pix	0.19	302	0.14
Flux Inpainting	0.40	174	0.65
Flux Editing	0.23	332	0.63
AnyV2V	0.23	290	0.33
CogVideoX	0.38	199	0.69
w/o x-concat	0.46	185	0.76
w/o cross-attn	0.59	220	0.55
w/o fusion	0.65	171	0.85
Ours	**0.67**	**168**	**0.88**

ablated variants. Results in Tab. 1 show that our model consistently outperforms others in human video quality, text alignment, and action faithfulness.

6.5. Human Evaluation

We supplement our analysis with a structured A/B test human evaluation. We assess the results based on four criteria: **(i) Scene consistency (SC)** evaluates how well the video preserves the original scene, even with flexible camera angles and scene motions. **(ii) Human quality (HQ)** assesses the realism of the generated human body. **(iii) Text-prompt alignment (PA)** evaluates how accurately the generated action and appearance match the given prompt. **(iv) Affordance prediction (AP)** assesses the interaction plausibility between the subject and the scene. Tab. 2 presents the percentage of subjects preferring each model, demonstrating that our model is consistently perceived as more realistic, natural, and capable of producing reasonable interactions compared to baselines and ablations.

Table 2. Human evaluation preference comparison with baseline approaches. Percentage rounded to integer shows how many human subjects prefer our model over the baseline or ablated model. 100% means our model is always perceived as better.

Model	SC (%)	HQ (%)	PA (%)	AP (%)
InstructPix2Pix	100	98	100	96
Flux Editing	87	94	99	97
Flux Inpainting	95	79	60	57
AnyV2V	100	100	100	98
CogVideoX	68	87	74	89
Runway Gen-3	54	65	67	70
Luma Ray-2	55	59	69	75
w/o x-concat	99	48	53	76
w/o cross-attn	73	89	61	69
w/o fusion	54	52	58	60
w/o noise decay	76	48	56	53

7. Conclusion

We explore the ability of text-to-video models to perceive affordance and reason about interaction through the task of populating empty scenes with moving humans. Beyond a creative application, we show that video generative models implicitly learn affordance and can simulate affordance-aware activities through extensive analysis of attention features. We provide preliminary insights into effectively leveraging video generative models beyond appearance rendering toward interaction simulation.

References

[1] Shutterstock video. https://www.shutterstock.com/video. 4

[2] Luma AI. Ray2: Advanced image-to-video generation model. https://lumalabs.ai/ray, 2024. Accessed: 2025-03-07. 6, 7

[3] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation, 2024. 2

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[5] Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. Flow map matching, 2024. 3

[6] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 4, 6, 7

[7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2

[8] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators, 2024. 1

[9] Xuanqing Cao, Wengang Zhou, Qi Sun, Weilun Wang, Li Li, and Houqiang Li. Disa: Disentangled dual-branch framework for affordance-aware human insertion. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2025. Just Accepted. 2

[10] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 4

[11] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024. 2

[12] Joya Chen, Difei Gao, Kevin Qinghong Lin, and Mike Zheng Shou. Affordance grounding from demonstration video to target image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6799–6808, 2023. 2

[13] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 975–983, 2018. 3

[14] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 2

[15] Vincent Delaitre, David F. Fouhey, Ivan Laptev, Josef Sivic, Abhinav Gupta, and Alexei A. Efros. Scene semantics from long-term observation of people. In *Computer Vision – ECCV 2012*, pages 284–298, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 2, 3

[16] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[17] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhos-

ale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. 4, 5

[18] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models, 2023. 2

[19] David F. Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei A. Efros, Ivan Laptev, and Josef Sivic. People watching: Human actions as a cue for single-view geometry. In *Proc. 12th European Conference on Computer Vision*, 2012. 2, 3

[20] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models, 2024. 2

[21] J. J. Gibson. *The Senses Considered as Perceptual Systems*. George Allen and Unwin LTD, 1996. 1, 2, 3

[22] Litong Gong, Yiran Zhu, Weijie Li, Xiaoyang Kang, Biao Wang, Tiezheng Ge, and Bo Zheng. Atomovideo: High fidelity image-to-video generation, 2024. 2

[23] Abhinav Gupta, Scott Satkin, Alexei A. Efros, and Martial Hebert. From 3d scene geometry to human workspace. In *Computer Vision and Pattern Recognition(CVPR)*, 2011. 2

[24] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models, 2023. 2

[25] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[26] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. 2022. 2

[27] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022. 2

[28] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 2

[29] Li Hu, Guangyuan Wang, Zhen Shen, Xin Gao, Dechao Meng, Lian Zhuo, Peng Zhang, Bang Zhang, and Liefeng Bo. Animate anyone 2: High-fidelity character image animation with environment affordance. *arXiv preprint arXiv:2502.06145*, 2025. 2

[30] Jian Jin, Yang Shen, Xinyang Zhao, Zhenyong Fu, and Jian Yang. Unicanvas: Affordance-aware unified real image editing via customized text-to-image generation. *International Journal of Computer Vision*, pages 1–25, 2025. 2

[31] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. 2023. 2

[32] Hyeonwoo Kim, Sangwon Beak, and Hanbyul Joo. David: Modeling dynamic affordance of 3d objects using pre-trained video diffusion models, 2025. 2

[33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023. 4

[34] K. Koffka. *Principles of Gestalt Psychology*. Routledge, 1999. 1

[35] Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024. 6, 7

[36] Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A. Efros, and Krishna Kumar Singh. Putting people in their place: Affordance-aware human insertion into scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[37] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 6, 7, 8

[38] Gen Li, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. One-shot open affordance learning with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3086–3096, 2024. 2

[39] Lei Li and Angela Dai. GenZI: Zero-shot 3D human-scene interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[40] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 4

[41] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. 3

[42] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 4

[43] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. One-shot affordance detection. In *IJCAI*, 2021. 5, 6

[44] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 2

[45] Donald A. Norman. *The Design of Everyday Things*. MIT Press., 2013. 1, 3

[46] Mirela Ostrek, Soubhik Sanyal, Carol O'Sullivan, Michael J. Black, and Justus Thies. Environment-specific people, 2023. 2

[47] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 2

[48] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff

Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models, 2024. 1, 2, 3, 5

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 7

[50] Weiming Ren, Harry Yang, Ge Zhang, Cong Wei, Xinrun Du, Stephen Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024. 2

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[52] Runway. Introducing gen-3 alpha: A new frontier for video generation. https://runwayml.com/research/introducing-gen-3-alpha, 2024. Accessed: 2025-03-07. 6, 7

[53] Nirat Saini, Navaneeth Bodla, Ashish Shrivastava, Avinash Ravichandran, Xiao Zhang, Abhinav Shrivastava, and Bharat Singh. Invi: Object insertion in videos using off-the-shelf diffusion models, 2024. 2

[54] Mengyi Shan, Brian Curless, Ira Kemelmacher-Shlizerman, and Steve Seitz. Animating street view. In *Proceedings of ACM SIGGRAPG Asia 2023*, 2023. 2

[55] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022. 2

[56] Jaskirat Singh, Jianming Zhang, Qing Liu, Cameron Smith, Zhe Lin, and Liang Zheng. Smartmask: Context aware high-fidelity mask generation for fine-grained object insertion and layout control. *arXiv preprint arXiv:2312.05039*, 2023. 2

[57] Jiajin Tang, Ge Zheng, Jingyi Yu, and Sibei Yang. Cotdet: Affordance knowledge prompting for task driven object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3068–3078, 2023. 3

[58] Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. Ul2: Unifying language learning paradigms, 2023. 3

[59] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *ArXiv*, abs/1812.01717, 2018. 7

[60] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2, 3

[61] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report, 2023. 2

[62] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*, 2023. 2

[63] X. Wang, Rohit Girdhar, and Abhinav Kumar Gupta. Binge watching: Scaling affordance learning from sitcoms. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3366–3375, 2017. 2, 3

[64] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say, interact as you can: Language-guided human motion generation with scene affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[65] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Dat++: Spatially dynamic vision transformer with deformable attention. *arXiv preprint arXiv:2309.01430*, 2023. 4

[66] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Vasu Sharma Russell Howes, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. 2023. 3

[67] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. 2024. 2

[68] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022. 3

[69] Zhangsihao Yang, Mengyi Shan, Mohammad Farazi, Wenhui Zhu, Yanxi Chen, Xuanzhao Dong, and Yalin Wang. Amg: Avatar motion guided video generation, 2024. 2

[70] 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. 6, 7

[71] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8466–8476, 2024. 2

[72] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. *arXiv:2311.10982*, 2023. 2

[73] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 7

[74] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance, 2024. 2