Overview

Our Appendix includes the following content.

Sec. A shows videos generated by our seven models. Figs. 12 to 26 (pp. 19–33) show samples generated by all the models for five examples.

Sec. B (pp. 34–42) provides further details from our data preparation pipeline. Fig. 27 shows a high-level overview of the filtering process. Tabs. 4 and 5 list additional details of the dataset construction (p. 34), while Figs. 36 and 37 compare video duration and resolution across the WYD, TikTok and TED-Talks datasets (p. 38). Figs. 28 to 34 (pp. 35–37) show sample videos that were removed as part of the filtering process, and Fig. 35 (p. 37) shows more examples from the final WYD dataset. Figs. 38 to 40 (pp. 40–41) show our UIs for video categorization, segmentation and pose annotations. Fig. 41 displays the overlap in videos between any two categories.

Sec. C reports additional results from our experiments. Fig. 42 (p. 43) shows the overall performance of depth- and edge-conditioned models on WYD_{16}. Fig. 43 shows the difference in errors of depth- and edge-conditioned models when adding captions as an additional source of guidance. Figs. 44 to 46 (pp. 43–44) instead provide quantitative support for the ablations of pose-guided models discussed in Sec. 5. Moreover, Figs. 47 to 50 (pp. 44–46) report and discuss category-level performance of our top-performing models (MimicMotion, ControlNeXt and VACE).

Sec. D includes further details about our human evaluation protocols. We report our instructions and setup for side-by-side human evaluations in p. 47, and show our UI in Fig. 51 (p. 46). Figs. 52 to 54 (pp. 47–48) present and discuss how metrics that we considered to measure different aspects of video generation score our evaluated models.

Finally, we share some ethical considerations related to controllable human video generation in Sec. E (p. 48), where we additionally remark that our WYD dataset is meant to be used for academic research purposes only.

A. Samples of generated videos

An overview of the evaluated models is shown in Tab. 3. Figs. 12 to 26 show and discuss the limitations of samples generated by all the models for five WYD examples.

Model	Condition	Extractor	Training data	Close-up single-person videos?	WYD overlap?
MagicAnimate [83]	Dense pose	Detectron2	TikTok [32]	Yes	No
MagicPose [13]	2D pose	OpenPose	TikTok [32]	Yes	No
MimicMotion [92]	2D pose	DWPose	Internal	Yes	No
ControlNeXt-SVD-v2 [56]	2D pose	DWPose	Internal	No	N/A
VACE-Wan2.1 [34]	2D pose	DWPose	Internal	N/A	N/A
Control-A-Video [16]	Depth / Canny	MiDaS / OpenCV	WebVid [1] subset* + internal	No	No
TF–T2V [75]	Depth	MiDaS	WebVid [1] subset* + internal	No	No
Ctrl-Adapter [41]	Depth / Canny	MiDaS / OpenCV	Panda-70M [15]	No	No

Table 3. **Overview of evaluated models.** We list models' conditions and used extractors, their training data and whether it mostly consists of close-up single-person videos, and whether any of the video datasets used in WYD were used by them during training. We thank the authors for clarifying information about their training data and confirming the absence of overlap with our evaluation videos. * Note that different models rely on different subsets of WebVid.

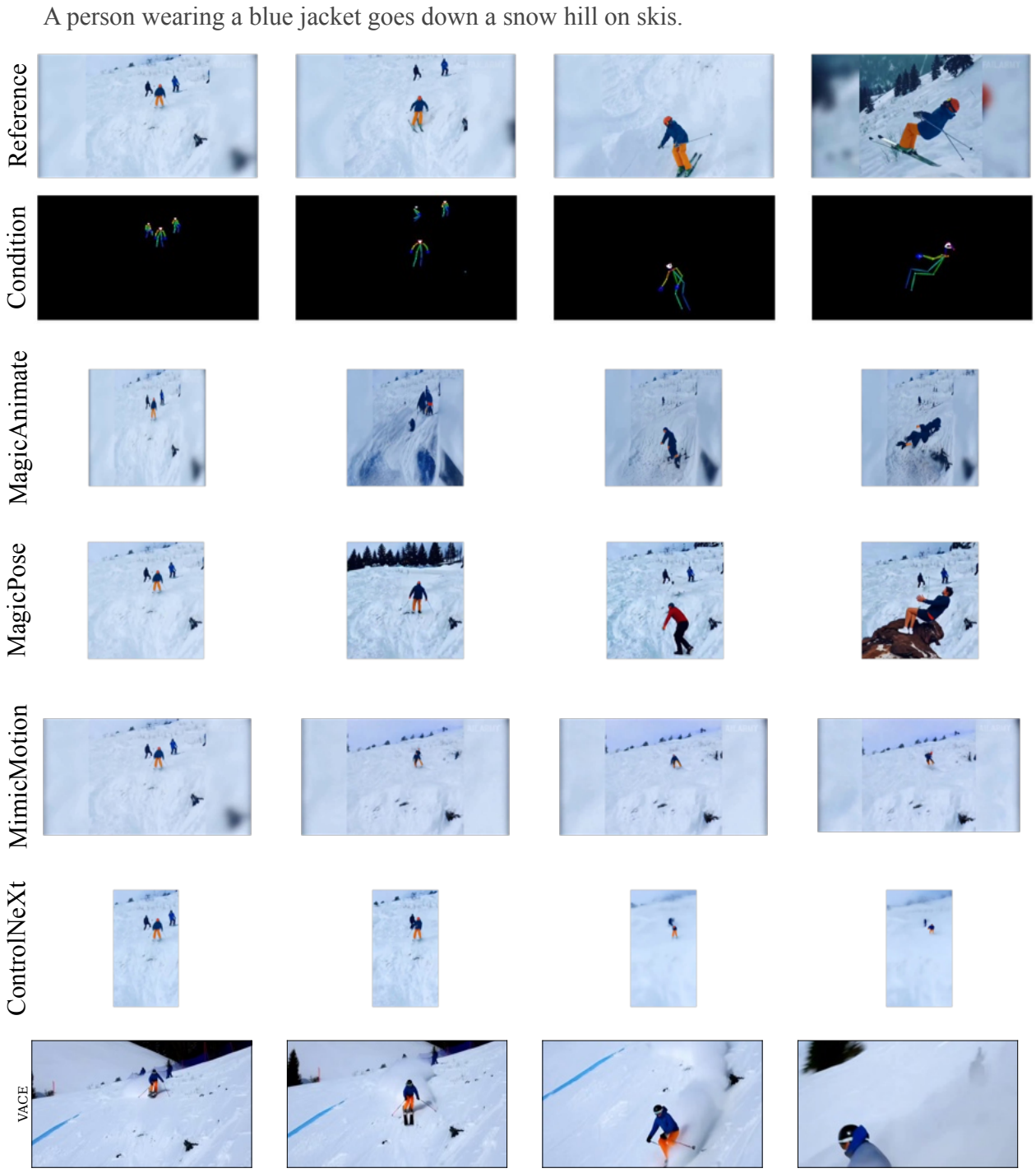

Figure 12. Example generations of our evaluated pose-conditioned models (MagicAnimate uses dense poses). We can see how people's appearance changes in MagicPose, although matching the human movements the best. We can also see the size mismatches in ControlNeXt and MimicMotion.

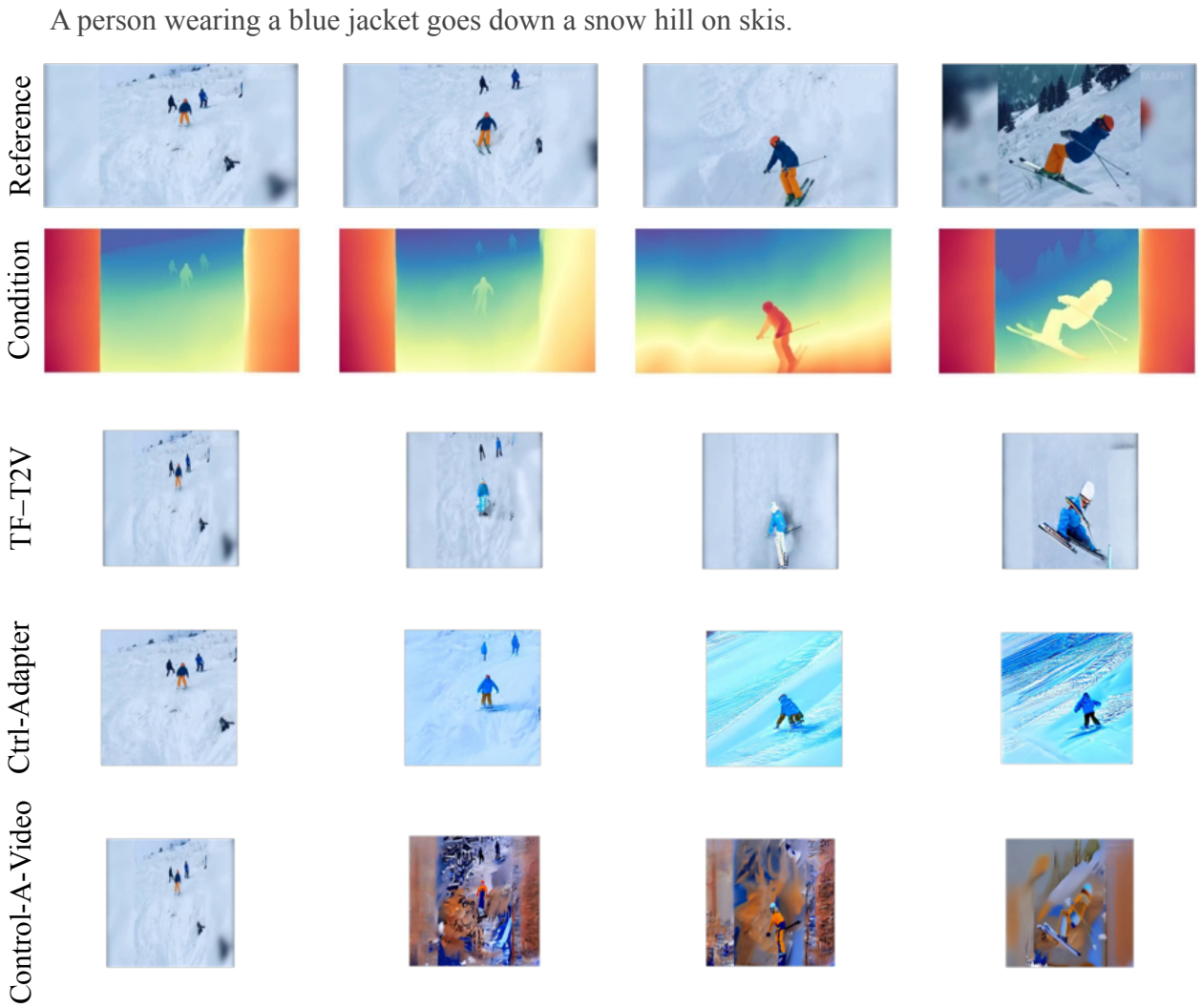

Figure 13. Example generations of our evaluated depth-conditioned models. We can see how people's appearance changes in TF-T2V, increasing saturation in Ctrl-Adapter and distortions in Control-A-Video.

Figure 14. Example generations of our evaluated edge-conditioned models. We can see increasing saturation in Ctrl-Adapter and distortions in Control-A-Video.

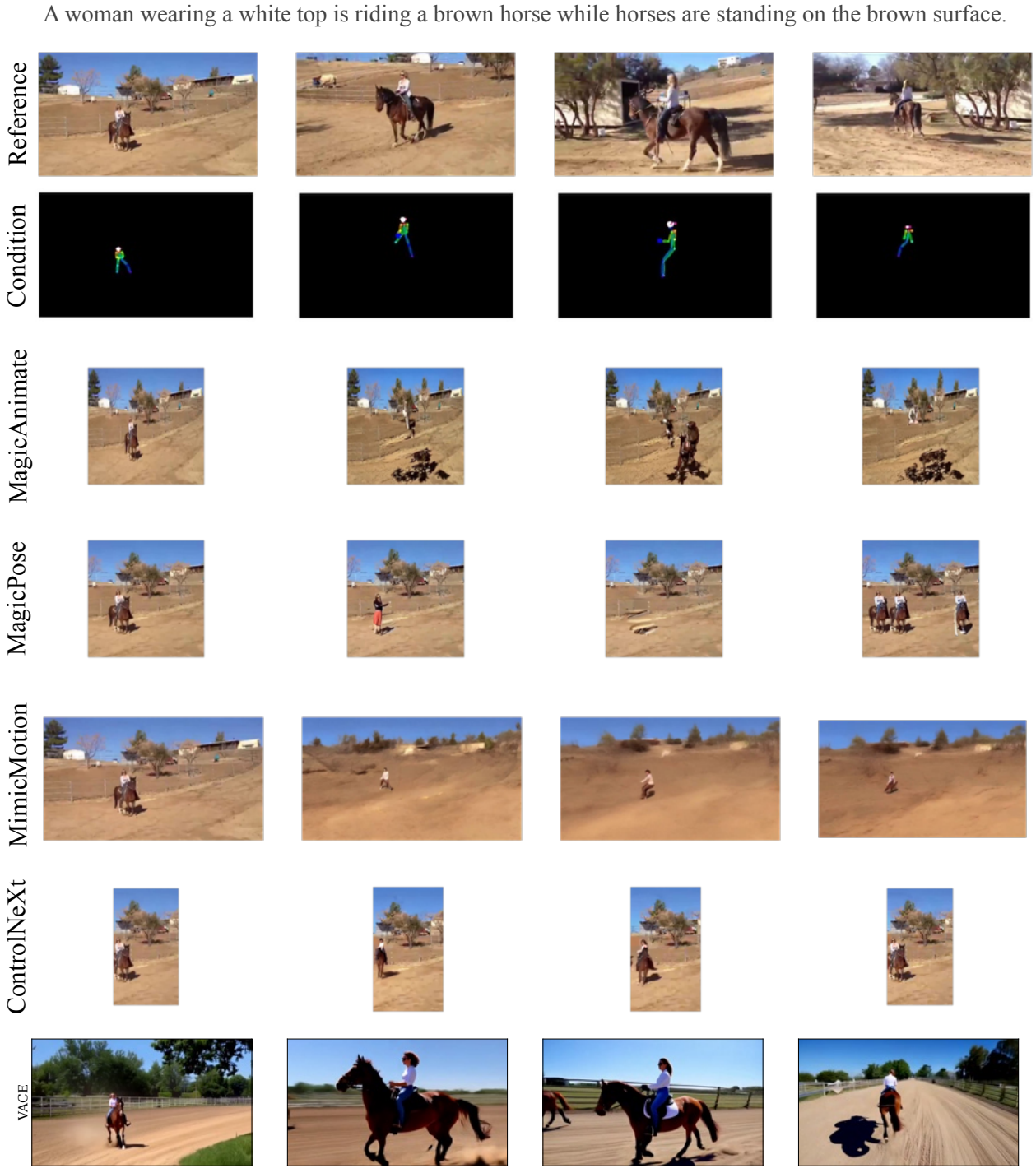

Figure 15. Example generations of our evaluated pose-conditioned models (MagicAnimate uses dense poses). We note the challenges in camera motion for all models, the distortions of characters in MagicAnimate, and flickering effects in MagicPose, as well as horse disappearance in MimicMotion.

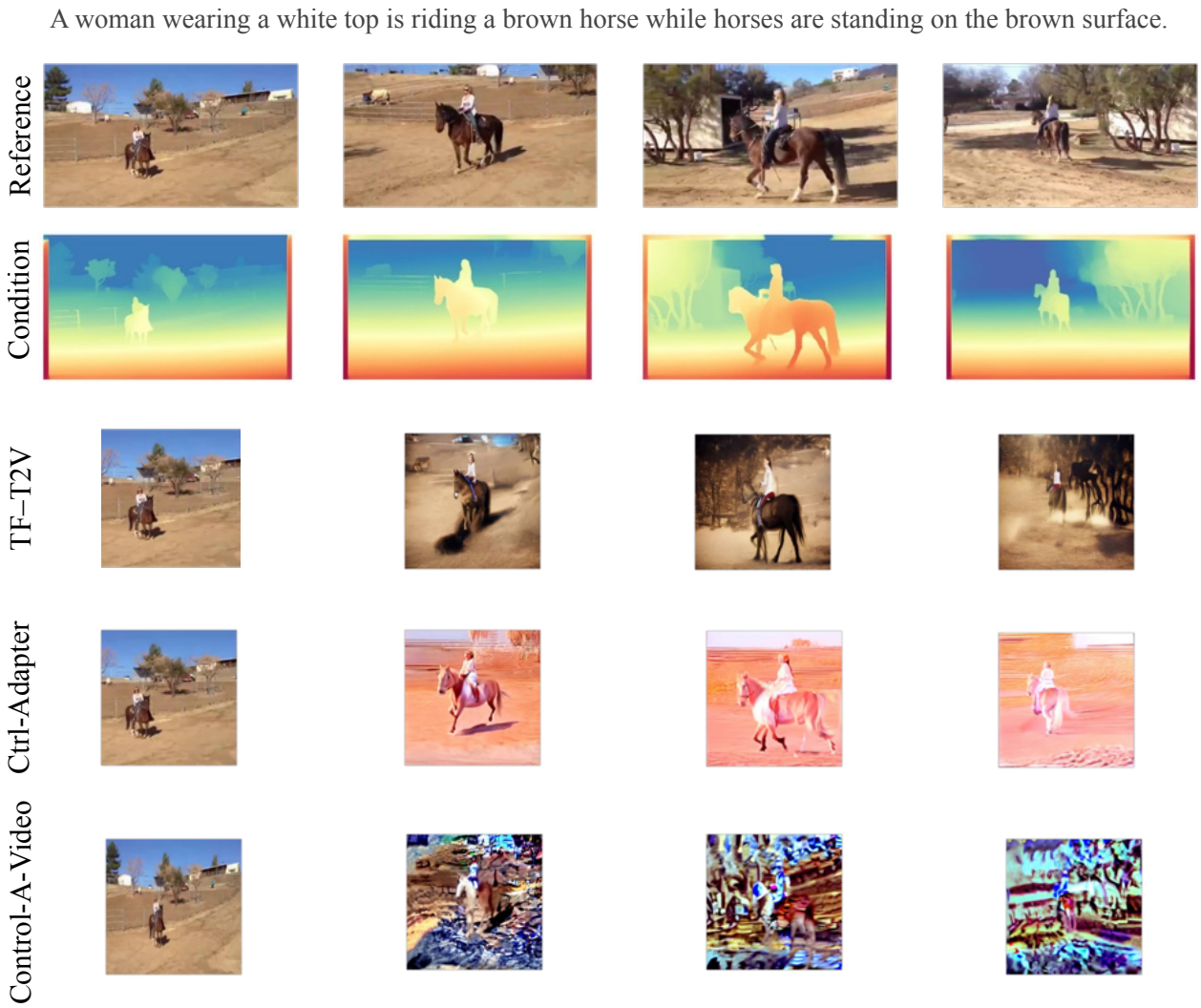

Figure 16. Example generations of our evaluated depth-conditioned models. We can see increasing saturation in Ctrl-Adapter and distortions in Control-A-Video, while TF-T2V best matches the overall scene.

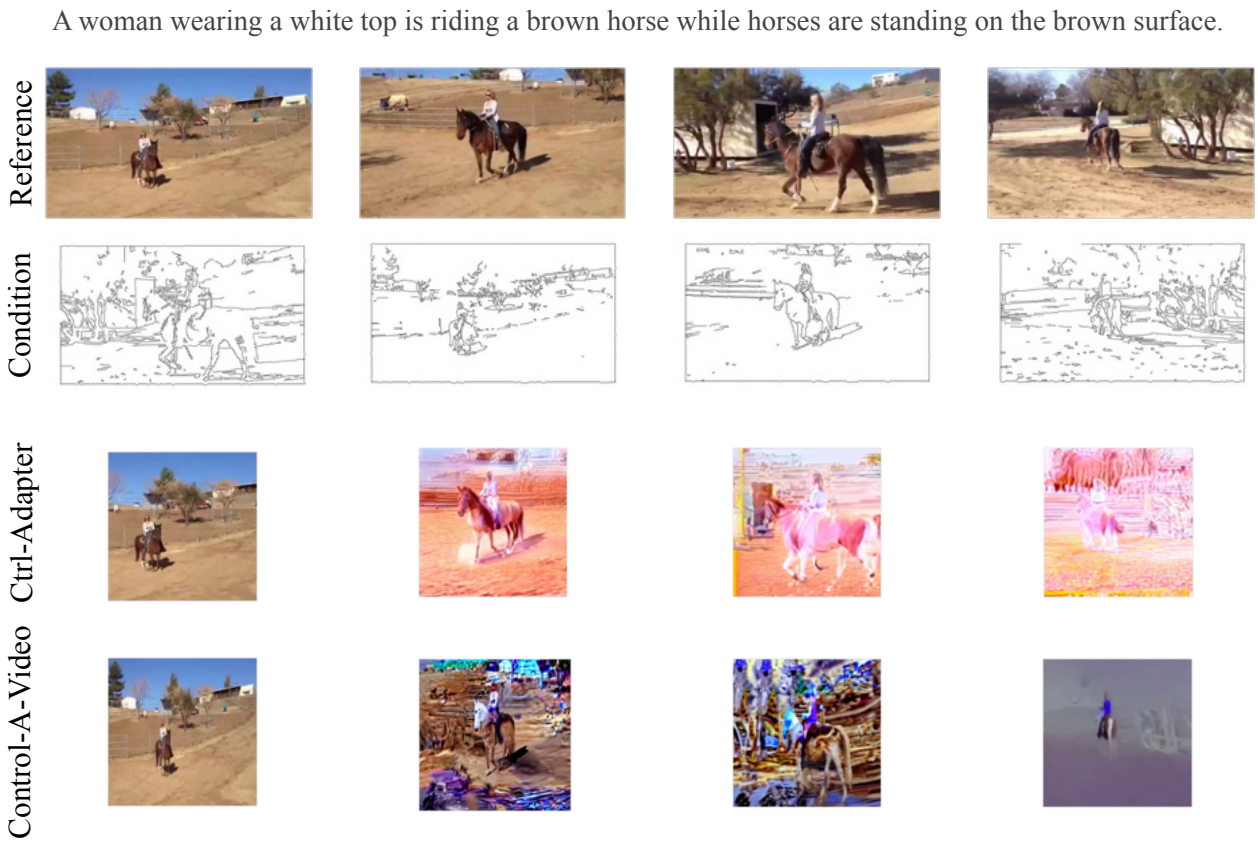

Figure 17. Example generations of our evaluated edge-conditioned models. We can see increasing saturation in Ctrl-Adapter and distortions in Control-A-Video.

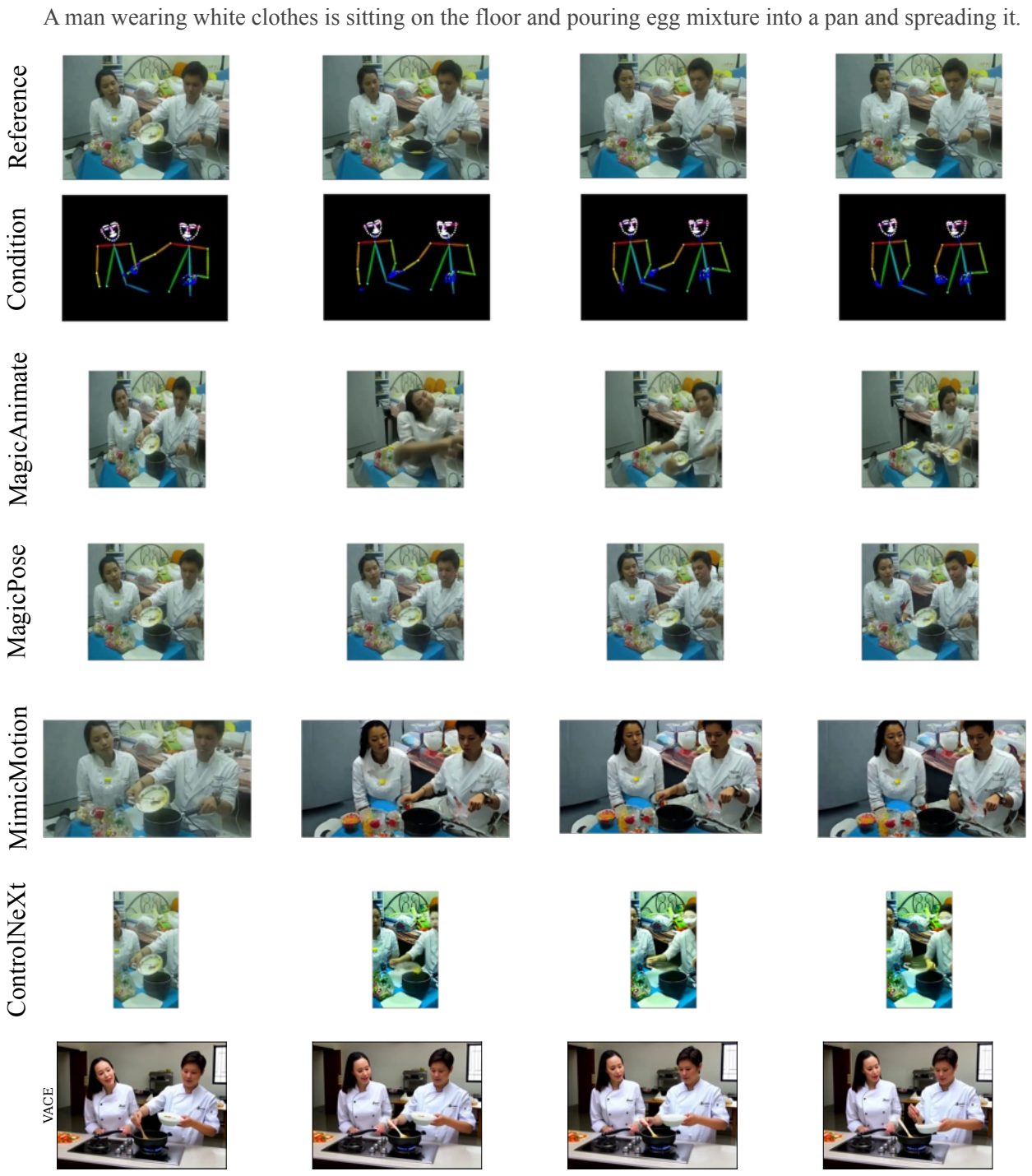

Figure 18. Example generations of our evaluated pose-conditioned models (MagicAnimate uses dense poses). We can see how MimicMotion changes the facial traits of humans towards specific age and beauty standards, and how it also fails to make the man interact with the pan. Due to its pre-processing, ControlNeXt misses the face of the man in the first frame and later creates a different one.

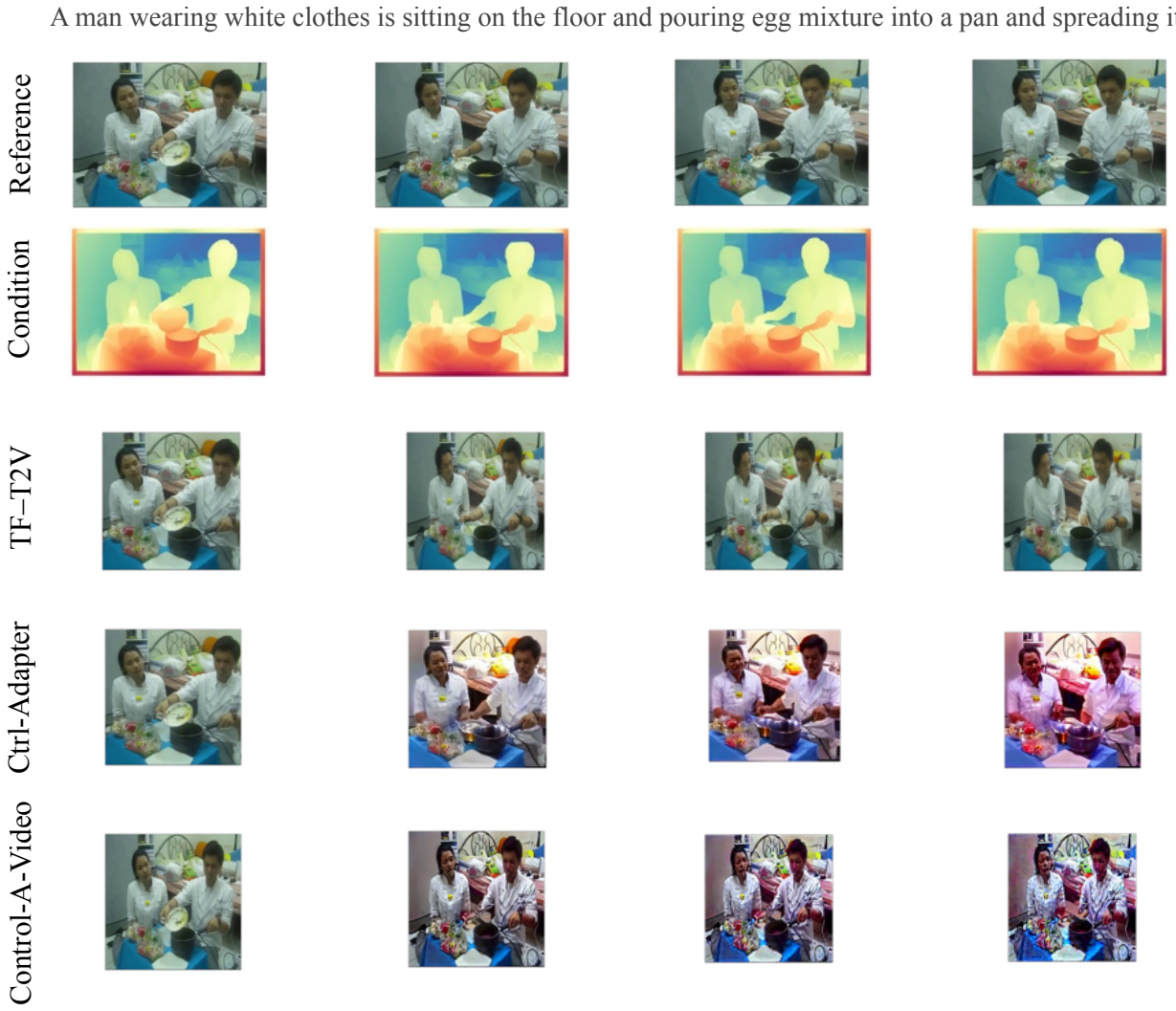

Figure 19. Example generations of our evaluated depth-conditioned models. We can see how Ctrl-Adapter change the facial traits of humans towards specific age and beauty standards. We can still see increasing saturation in Ctrl-Adapter and distortions in Control-A-Video, although less than in previous, dynamic examples.

Figure 20. Example generations of our evaluated edge-conditioned models. Similar to their depth-conditioned counterparts, Ctrl-Adapter shows increasing saturation and Control-A-Video presents distortions.

Figure 21. Example generations of our evaluated pose-conditioned models (MagicAnimate uses dense poses). Besides the large distortions and artifacts in MagicAnimate, all pose-conditioned models fail to generate the fish. MimicMotion again changes the man's appearance by removing his beard.

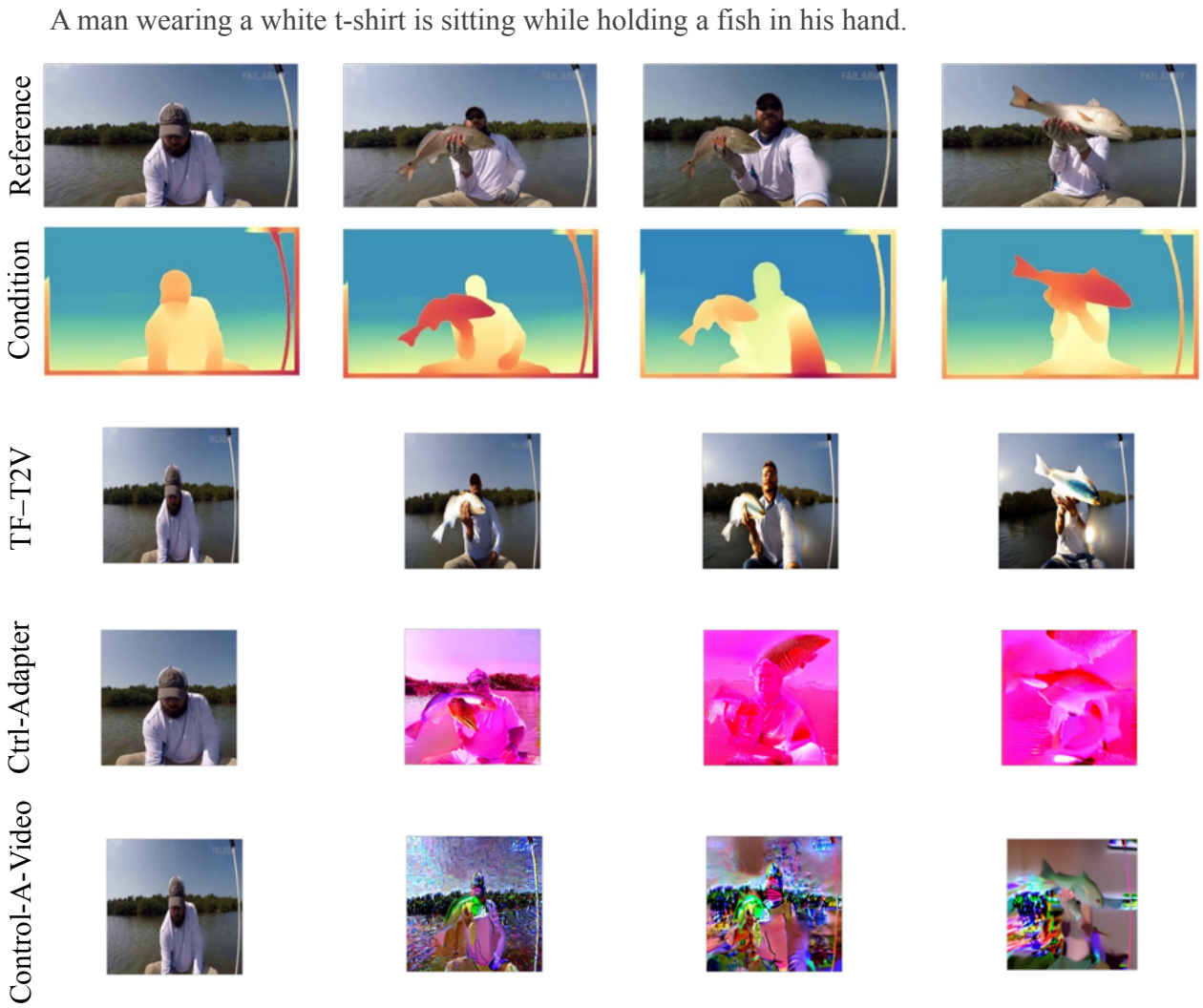

Figure 22. Example generations of our evaluated depth-conditioned models. We see high levels of saturation and distortions in Ctrl-Adapter and Control-A-Video, respectively. TF-T2V is the only model that synthesizes the fish well.

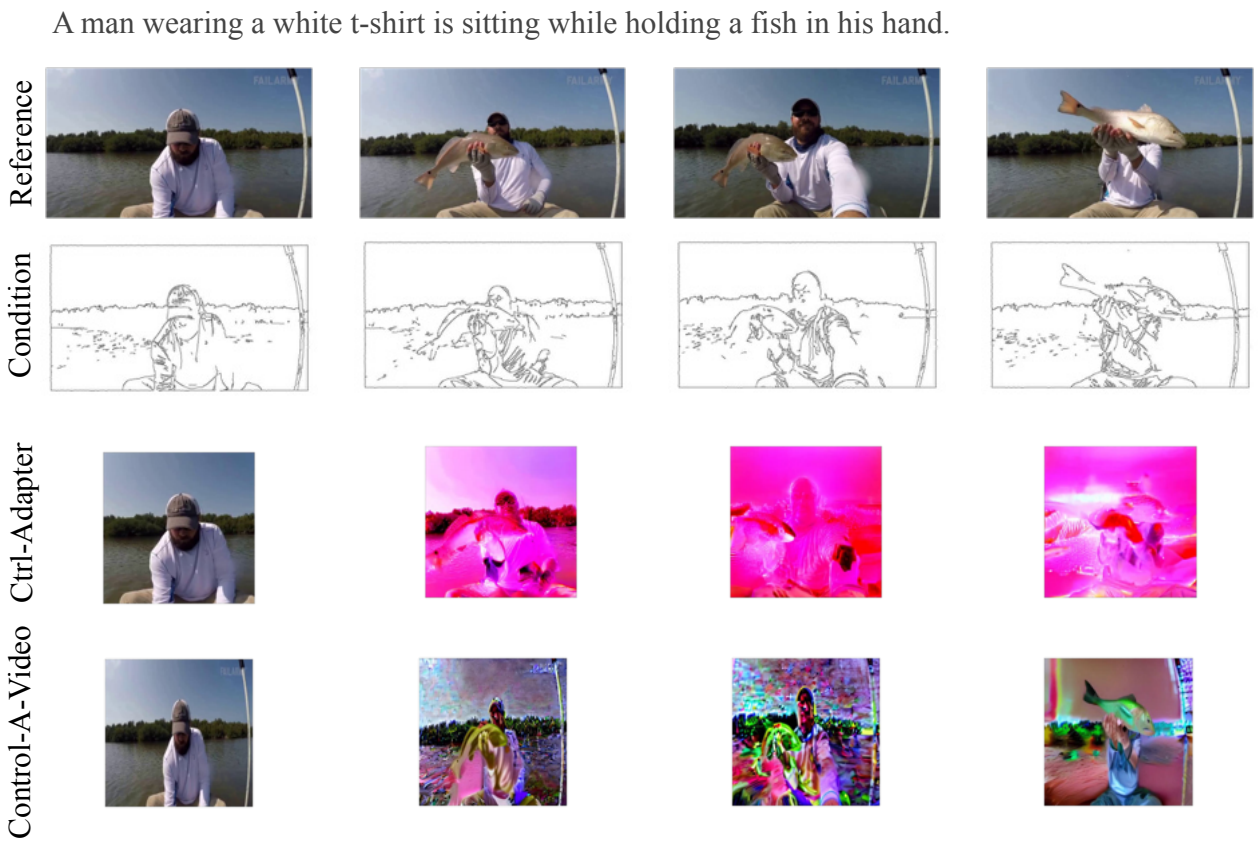

Figure 23. Example generations of our evaluated edge-conditioned models. Ctrl-Adapter and Control-A-Video generates videos with very high levels of saturation and distortion.

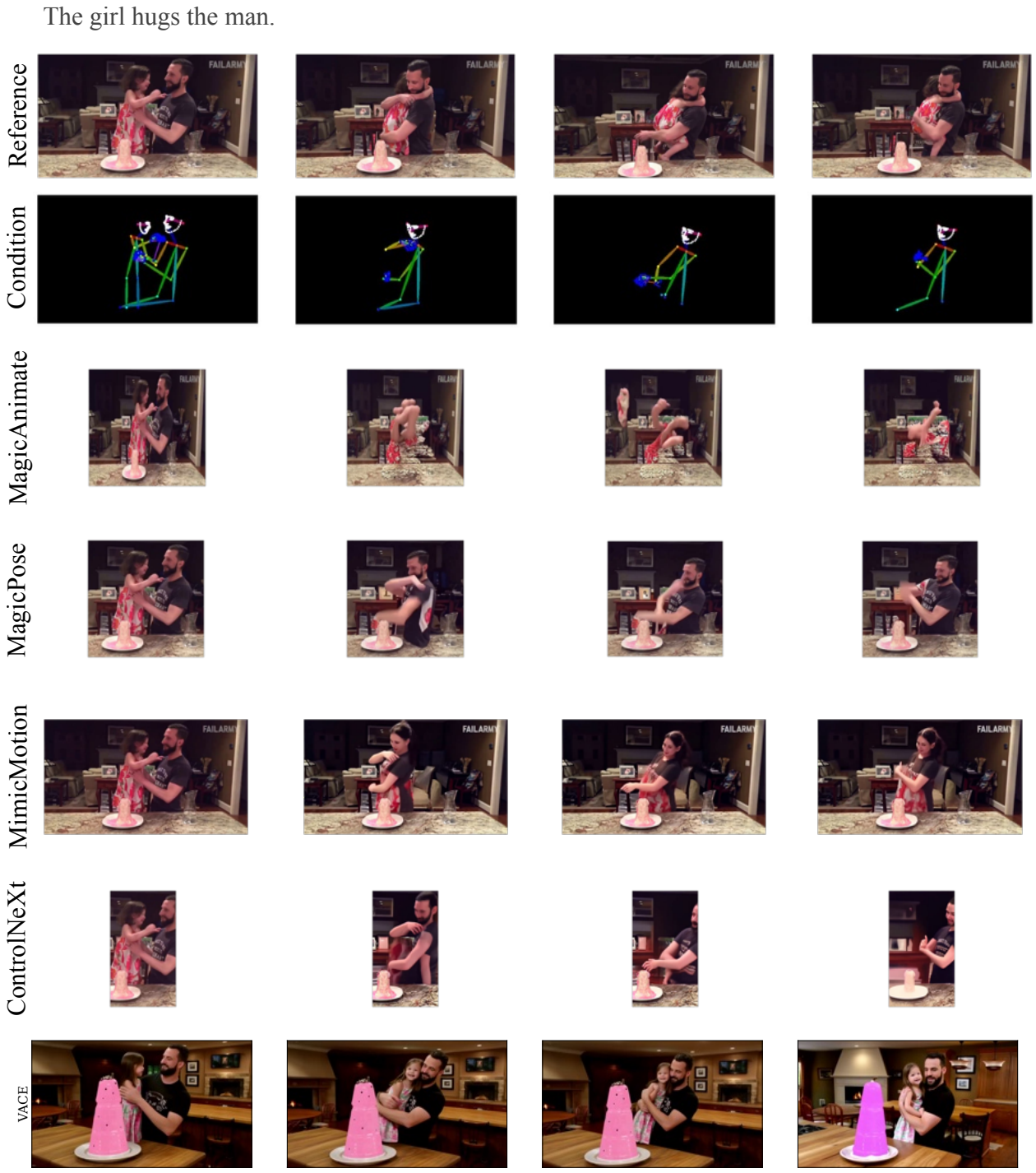

Figure 24. Example generations of our evaluated pose-conditioned models (MagicAnimate uses dense poses). All models struggle with generating multiple humans interacting with each other consistently due to the limitations of the pose extractor. For example, MagicPose and ControlNeXt make the girl disappear, while MimicMotion merges the girl and the man into a woman.

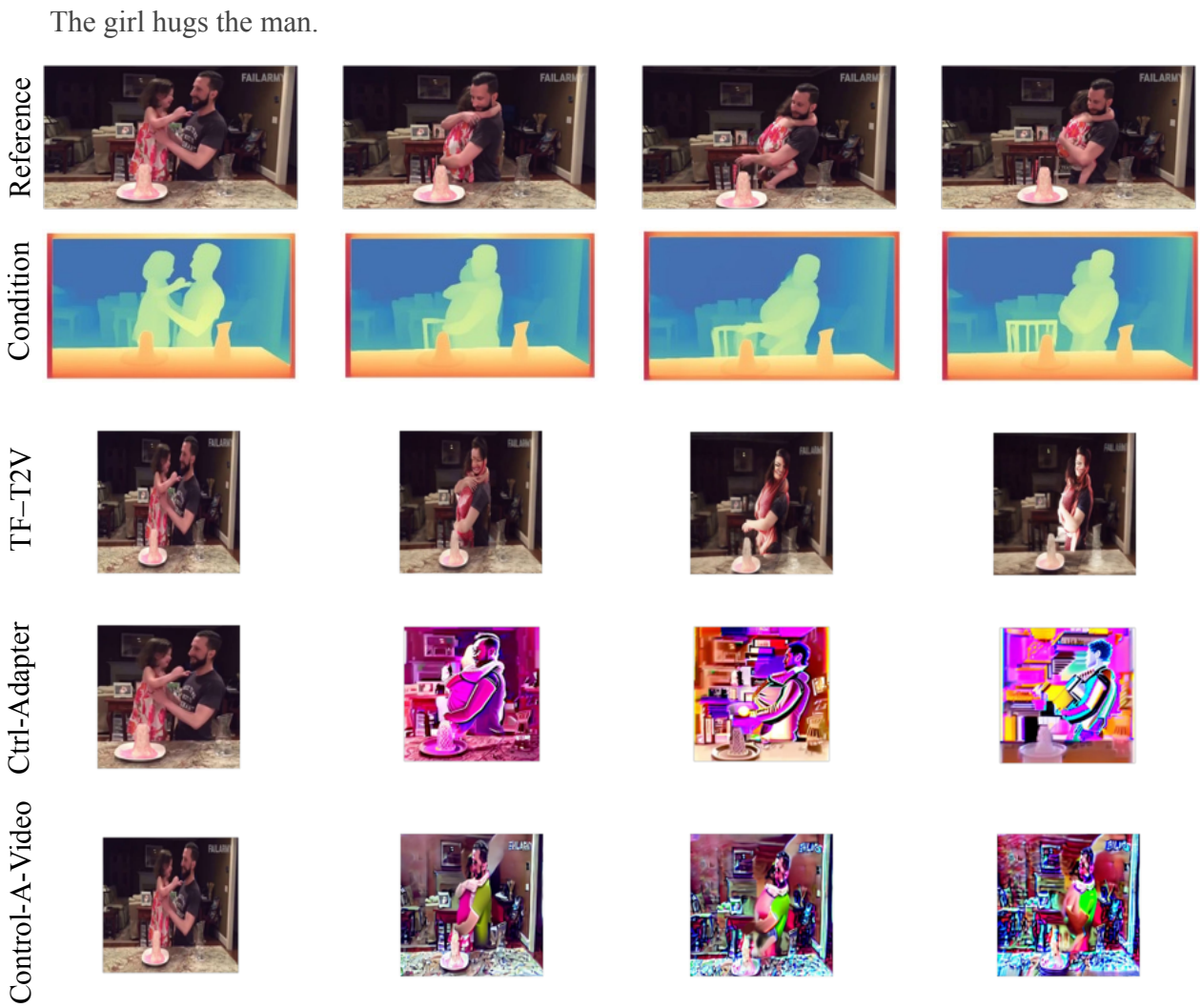

Figure 25. Example generations of our evaluated depth-conditioned models. We can see that even TF-T2V struggles to make the two humans interact with each other, resulting in a video where the man and the girl are merged into a single person with low-quality facial traits.

Figure 26. Example generations of our evaluated edge-based models. Once again, the generations of Ctrl-Adapter and Control-A-Video are affected by high levels of saturation and distortion, despite being a static scene.

Processing step	*Dev*			*Test*		
	UVO	**Oops**	**DiDeMo**	**UVO**	**Oops**	**DiDeMo**
StoryBench	2332	3472	2607	3895	3726	2319
Videos of humans	2085	3009	1479	3497	3168	1356
Scene cuts	1960	2962	1497	3264	3126	1342
Human visibility	996	1800	847	1638	1938	783
Video length	941	1545	706	1550	1650	670
Text alignment	545	859	550	870	953	511
Video resolution	545	859	501	870	953	462
Manual verification (WYD)	235	323	97	346	424	119

Table 4. Number of StoryBench entries in our datasets after each step in the WYD data preparation pipeline (see Fig. 27 and Sec. 3.1).

B. Data annotation details

In this section, we provide additional details, statistics and samples of our data preparation and categorization processes.

Data licenses. We rely on three publicly available datasets to source videos for our benchmark. Kinetics [36] is released under a CC BY 4.0 license, DiDeMo [25] is released under a CC BY-NC-SA 2.0 license, and Oops [18] is released under CC BY-NC-SA 4.0 license. Moreover, we use captions and metadata collected in StoryBench [7], which are also released under a CC BY 4.0 license.

B.1. Data filtering

High-quality text descriptions are necessary to accurately evaluate text-guided video models. For the datasets above, StoryBench [7] includes human-written captions for video generation, as well as useful metadata, including event boundaries and actor identification (*i.e.*, the entities with a key role) in a video [70]. WYD leverages StoryBench annotations, but we treat each video segment separately (rather than as part of a sequence) in our 7-step pipeline (see Fig. 27). Tab. 4 details how each step of our pipeline affected the number of entries of StoryBench, and Figs. 28 to 34 show examples that were dropped at each step.

1. Videos with human actors. We start by filtering out videos where the main actors are not humans. For Kinetics and Oops videos, we use human-labeled metadata in StoryBench, which associates each caption to its main actor (*e.g.*, "man with white t-shirt"). For DiDeMo, we extract the main actors of each caption with an instruction-tuned LLM [22] (note that it is not trivial to use object detectors to extract the main actors due to non-salient humans in a video) with the prompt: "Which living being is performing the main action in the following caption? Reply with one word." From 600+ actors, we manually identify 224 referring to humans (Tab. 5).

2. Removing scene cuts. Most of the original videos are single shot, but we found a few of them with multiple scenes. We use a shot detector [11] and whenever it detects 2–4 cuts, we remove the video if none of the parts lasts for at least 80% of the original duration. Otherwise we replace the original video with that part. Doing so allows us to remove scene cuts while preserving the actions described in the captions.

3. Ensuring visible humans. People performing the main action in the video need to be visible, especially in the first frame, to evaluate how image-to-video systems can generate them. We annotate our videos with a state-of-the-art human pose estimator [85], and keep only those in which people are 'mostly visible' (defined as 11/18 body keypoints detected) in the first, and at least 70% of the frames.

4. Removing short and long videos. Given that our primary goal is to evaluate human video generation, we take into account the capabilities and the computational resources of most existing models to date, and opt to limit videos in our benchmark to a duration between 1.5s and 15s.

5. High text alignment. The captions for Kinetics and Oops videos were originally made from the perspective of a specific actor [70]. As a result, some captions fail to naturally describe the most salient actors in a video. To minimize such cases, we only keep the actor whose video–caption pairs have the highest similarity according to a fine-grained contrastive VLM [20]. With the same approach, we then remove the bottom 25% of the videos to maximize text controllability.

6. Minimum resolution. We further filter down our data to only include videos with smaller edge of at least 360 pixels, to ensure references of higher quality while keeping enough samples for statistical significance.

7. Manual verification. Finally, we meticulously scrutinize the quality of the resulting videos and remove those with (i) significant blur, (ii) poor lighting, (iii) unstable camera, (iv) low motion, (v) unclear captions, or (vi) where the first frame does not capture the main actors. This process, in conjunction with the video categorization below, was done in multiple rounds and took over 500 hours of annotation time. The authors validated the annotations and sought to ensure that the videos had high diversity and would be challenging for current video generation models.

At the end of this pipeline, WYD contains 1,544 high-

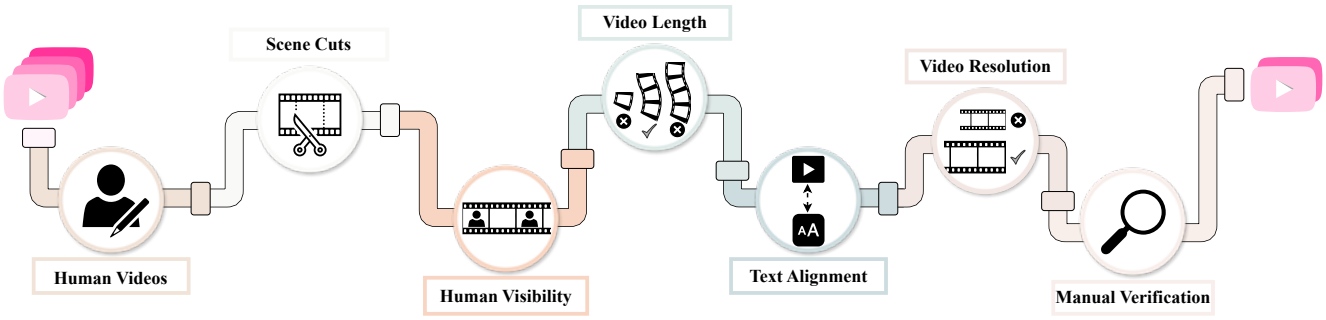

Figure 27. **WYD data filtering pipeline.** Our pipeline includes 7 steps: identifying videos with human actors, removing scene cuts, ensuring human visibility, removing short/long videos, keeping videos with high text alignment, removing low-res videos and manual verification.

Baby girl one	Bikers	Daddy	Group of People Two	Lady	Mother	Persons	Tourists
A baby	Boy	Dancers	Group of children	Lady one	Musicians	Pianist	Twins
A boy	Boy	Divers	Group of men	Lady three	Officer	Player	User
A girl	Boy Four	Drummer	Group of musicians	Lady two	Old lady one	Players	Woman
A man	Boy One	Everyone	Group of people	Man	Old lady two	Police officer	Woman
A old man	Boy Three	Family	Group of people	Man	Old man	Priest	Woman
A person	Boy Two	Fighter	Group of people one	Man Five	Old woman	Rangers	Woman One
A person	Boy five	Fighters	Group of people one	Man Four	Others	Referee two	Woman Three
A woman	Boy four	Fourth man	Group of people two	Man One	Passengers	Rest	Woman Two
Baby	Boy one	Gentleman	Group two	Man One	People	Rider	Woman five
Baby	Boy one	Girl	Guitarist	Man Three	Peoples	Rider one	Woman four
Baby Boy	Boy three	Girl	Guy	Man Two	Performer	Rider one	Woman one
Baby boy	Boy two	Girl three	He	Man five	Person	Rider two	Woman one
Baby girl	Boy two	Girl One	Judge	Man four	Person	Riders	Woman three
Baby girl one	Bride	Girl Two	Kid	Man one	Person Four	Runner	Woman two
Baby girl two	Bridegroom	Girl five	Kid	Man one	Person One	Second man	Woman two
Baby girl two	Child	Girl four	Kid Four	Man one	Person Three	Singer	Women
Baby one	Child	Girl one	Kid Three	Man six	Person Two	Singers	Women
Baby two	Child one	Girl one	Kid Two	Man three	Person five	Someone	Women one
Band	Child one	Girl third	Kid four	Man three	Person four	Speaker	Young man
Batter	Child two	Girl three	Kid one	Man two	Person one	Swimmers	kid
Bichon frise	Child two	Girl two	Kid one	Man two	Person one	They	kid one
Bicyclist	Couple	Girl two	Kid three	Marchers	Person three	Third	kid three
Bicyclists	Cyclist	Girls	Kid two	Mascot	Person three	Third boy	kid two
Bike rider	Cyclist Two	Group	Kid two	Members	Person two	Third girl	person
Biker	Cyclists	Group of People One	Kids	Men	Person two	Third man	woman

Table 5. List of unique human actors extracted from StoryBench annotations.

quality videos (from the original 18,351) which enable the fine-grained analyses described next at a tractable runtime. Figs. 36 and 37 show the distribution of video duration and resolution for WYD, TikTok and TED-Talks.

Manual video quality verification and labeling were carried through an extensive period (over two months). Most of the resulting videos (99.5%) have a size of at least 512p. Moreover, we note that a fraction of the frames display motion blur, which is unavoidable in videos with high motion.

B.2. Video categorization

A key goal of our benchmark is to enable fine-grained understanding of the capabilities of video generation models to synthesize humans across different facets; rather than reporting a single aggregated score, as done with other datasets. To achieve this, we annotate our data with nine categories that capture important aspects for synthesizing videos of humans. Each category in turn contains sub-categories (see Fig. 2 for an overview), each with at least ≈100 samples so as to provide sufficient statistical power for our analyses.

Number of human actors. For each video in WYD, we

Figure 28. Discarded samples due to no human actor.

Figure 29. Discarded samples due to scene cuts.

manually label the exact number of humans performing the main actions (*i.e.*, salient for generation), and then group them in three groups (1, 2, 3–8). Notably, each sub-category presents specific challenges, from consistently generating multiple people to more dynamic videos with a single person.

Human actor size. The size of human actors can affect how well a video generation model performs. We manually estimate the area covered by the human actors in each video, and categorize them into seven splits of actor size (Fig. 2).

Human occlusion. Object consistency is crucial in generated videos, and humans need to keep their appearance despite partial or full occlusions. We measure the average number of body keypoints detected by the pose estimator [85], and categorize our videos into five ranges of human actor occlusion (*i.e.*, percentage of keypoints that are not visible).

Human actions. The ability to perform a wide range of actions is a distinctive characteristic of humans, and different actions may require disparate generation capabilities (*e.g.*, swimming vs. eating). We manually assign action labels to each video in multiple rounds, adjusting the levels of

Figure 30. Discarded samples due to low human actor's visibility.

Figure 31. Discarded samples due to video duration.

specificity after each round. This process yields sixteen subcategories of visually similar actions (as shown in Fig. 2).

Human locomotion. We manually classify human body movements into three categories: full-body, partial-body, and hand-focused. While full-body motion indicates actors changing location in the video, partial-body motion involves only moving part of the body (*e.g.*, the arms). We label videos where hands' motion is crucial separately, as existing models often struggle to generate hands [39, 47, 88].

Camera motion. We manually label each video as dynamic if the camera follows the actor, or static otherwise.

Video motion. The primary aspect that differentiates videos from images is the motion within them. We use an optical flow model [67] to estimate the amount of motion in each video, which we then use to study how this key aspect affects human video generation across seven motion ranges.

Actor interactions. Humans often interact with their environment, either through inanimate objects, animals or other humans. Object interactions are often hard to generate, as they require a deeper understanding of shapes, texture and

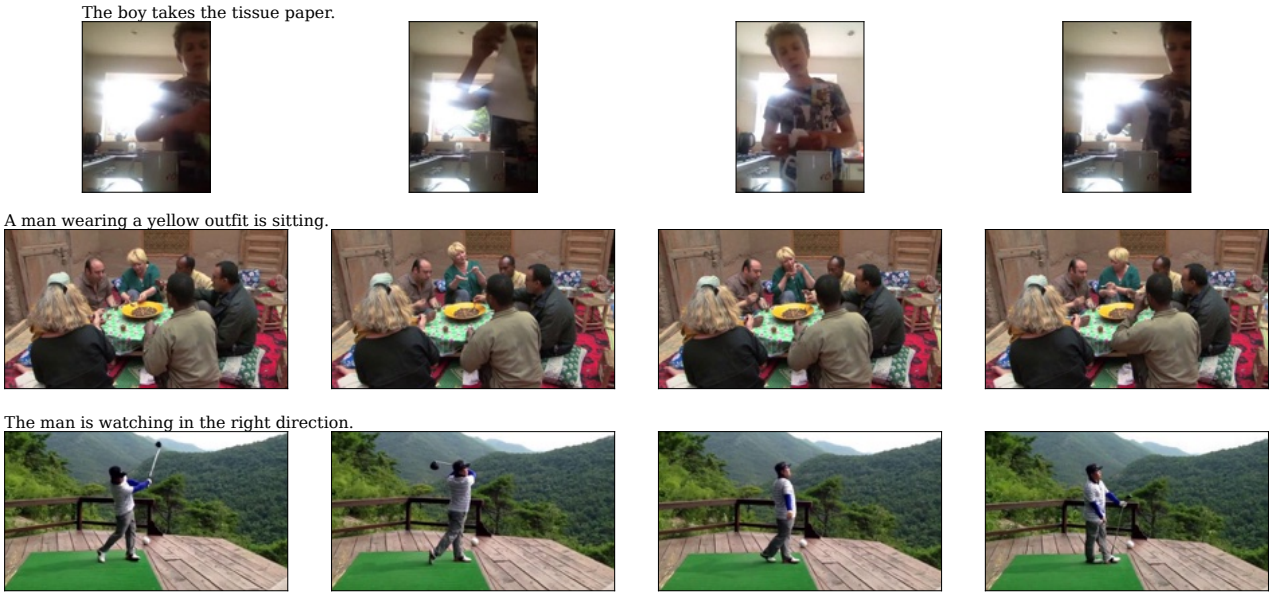

Figure 32. Discarded samples due to low video–text alignment.

Figure 33. Discarded samples due to low video resolution.

the physical laws of the world. While previous work only evaluates solo actions [12, 32, 63], our annotations show that most of the videos in our dataset involve interactions.

Scene. Different actions are associated with different environments (*e.g.*, swimming) and video generation models should be able to synthesize a variety of environments. We manually annotate the videos in WYD with nine different scenes where actions take place (both indoors and outdoors).

Discussion. Fig. 38 shows our UI for video categorization. We find that only a few categories overlap with each other significantly (see Fig. 41 in Sec. B). Namely, actions and interactions with animals; and video and camera motion, where high-motion videos come from dynamic camera, and videos with low motion correspond to static camera. Interestingly, high-motion videos do not necessarily involve fewer people, but small people are often associated with high motion and full-body movement of a single actor.

Figure 34. Discarded samples due to manual verification.

B.3. Dense annotations

Video segmentation masks. In addition to labeling our videos with 9 categories, we also annotate each human actor in a video with tracked segmentation masks. To do this, we first identify people in the first frame of each video via bounding boxes using OWLv2 [53]. After selecting and refining the bounding boxes corresponding to the actors only, we feed them as input to SAM 2 [59], which returns video segmentation tracks for each of the actors. These tracks are further verified and manually corrected at the frame level by the authors, an effort that took over 1000 hours. We use them to define new automatic metrics for WYD by analyzing model performance at the human level, as discussed next. Fig. 39 shows our UI for verifying and fixing video segmentation masks for each actor. Annotations were made at every frame using a brush to extend or delete pixels for the automatically generated segmentation masks for each human actor.

Actor 2D pose key-points. We further collect human-annotated 2D pose key-points for the (184) actors across 100 randomly selected videos (UI shown in Fig. 40). Human raters are given the option to modify existing keypoints extracted using DWPose [85] or to discard them and annotate human skeletons from scratch. This data collection campaign also took over 1000 hours of human rating time. We use these skeletons to verify the trustworthiness of our evaluation framework.

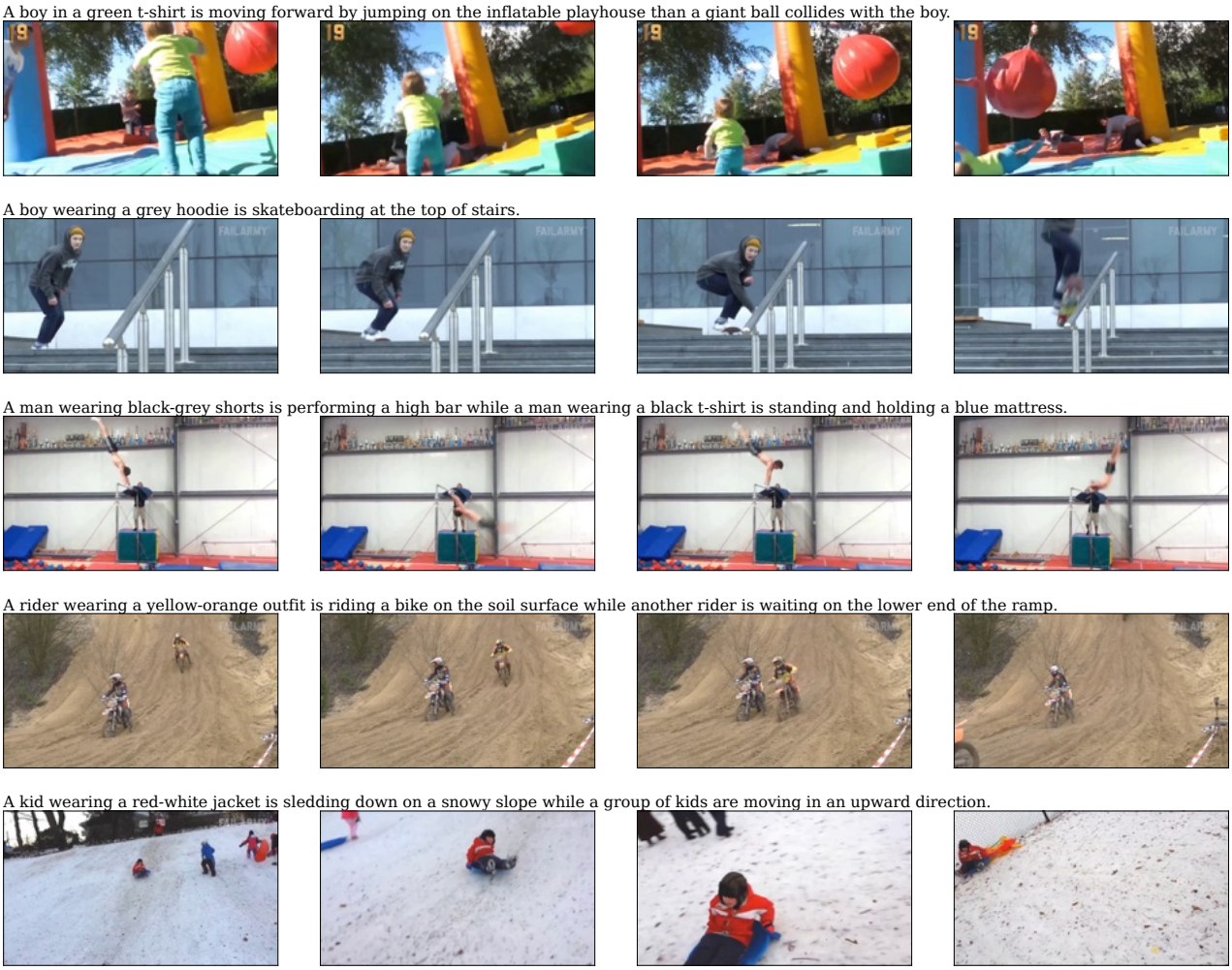

Figure 35. Samples from the final WYD dataset.

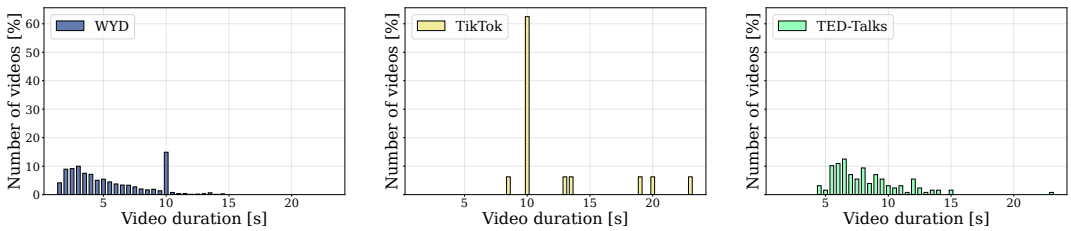

Figure 36. **Distribution of video duration in WYD, TikTok and TED-Talks.** WYD covers actions lasting a few seconds and up to 15s.

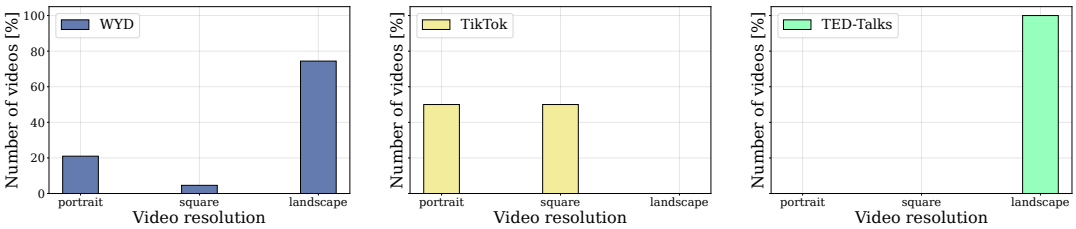

Figure 37. **Distribution of video resolution in WYD, TikTok and TED-Talks.** WYD contains videos that have more diverse aspect ratios.

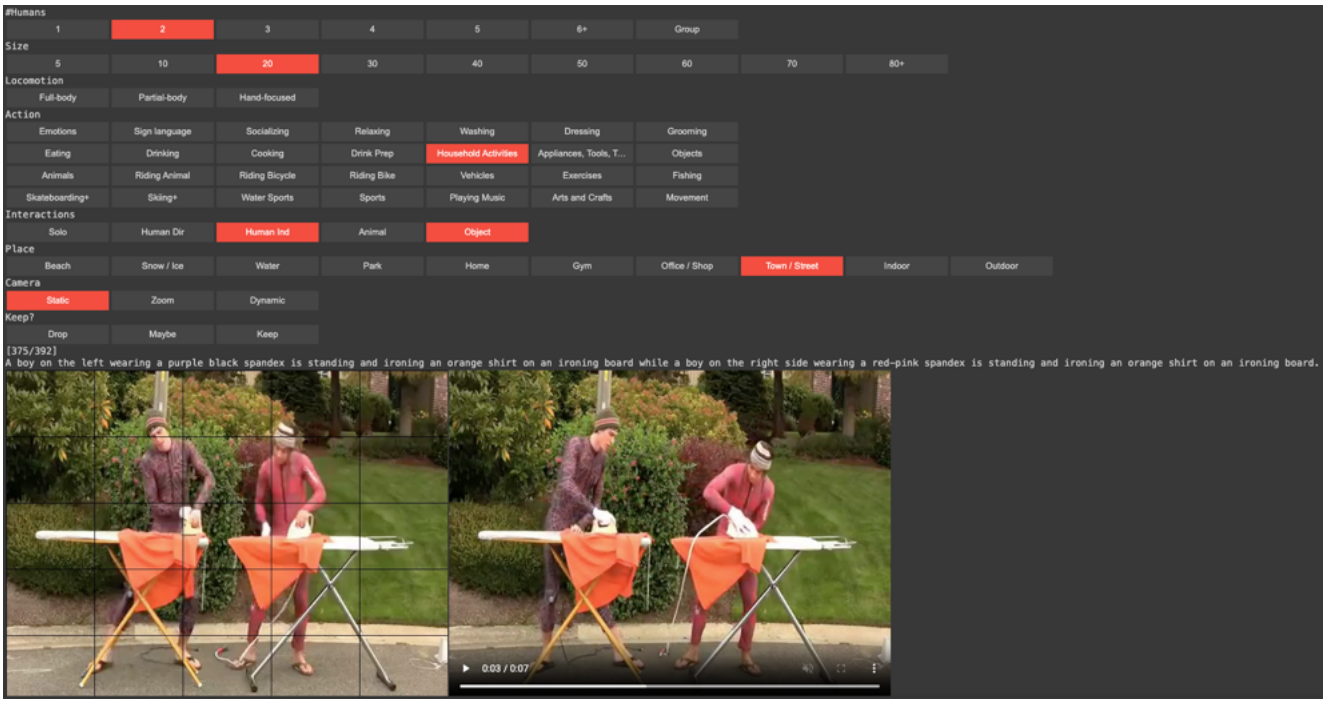

Figure 38. UI used to manually filter and label videos in WYD according to different categories.

Figure 39. UI used to manually fix video segmentation masks in WYD through a brush to select pixels corresponding to an actor's mask.

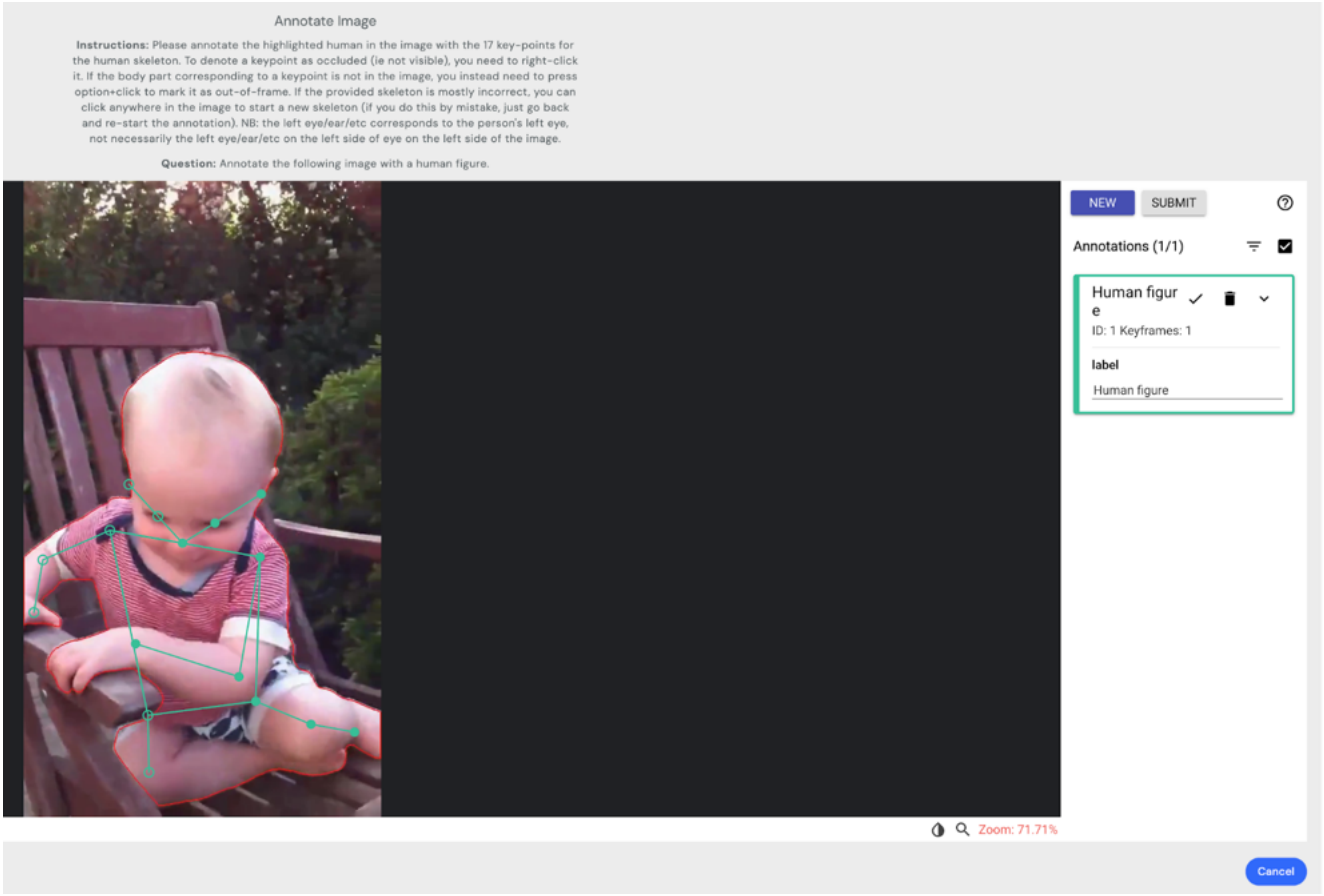

Figure 40. UI used to manually annotate 2D body key-points in WYD.

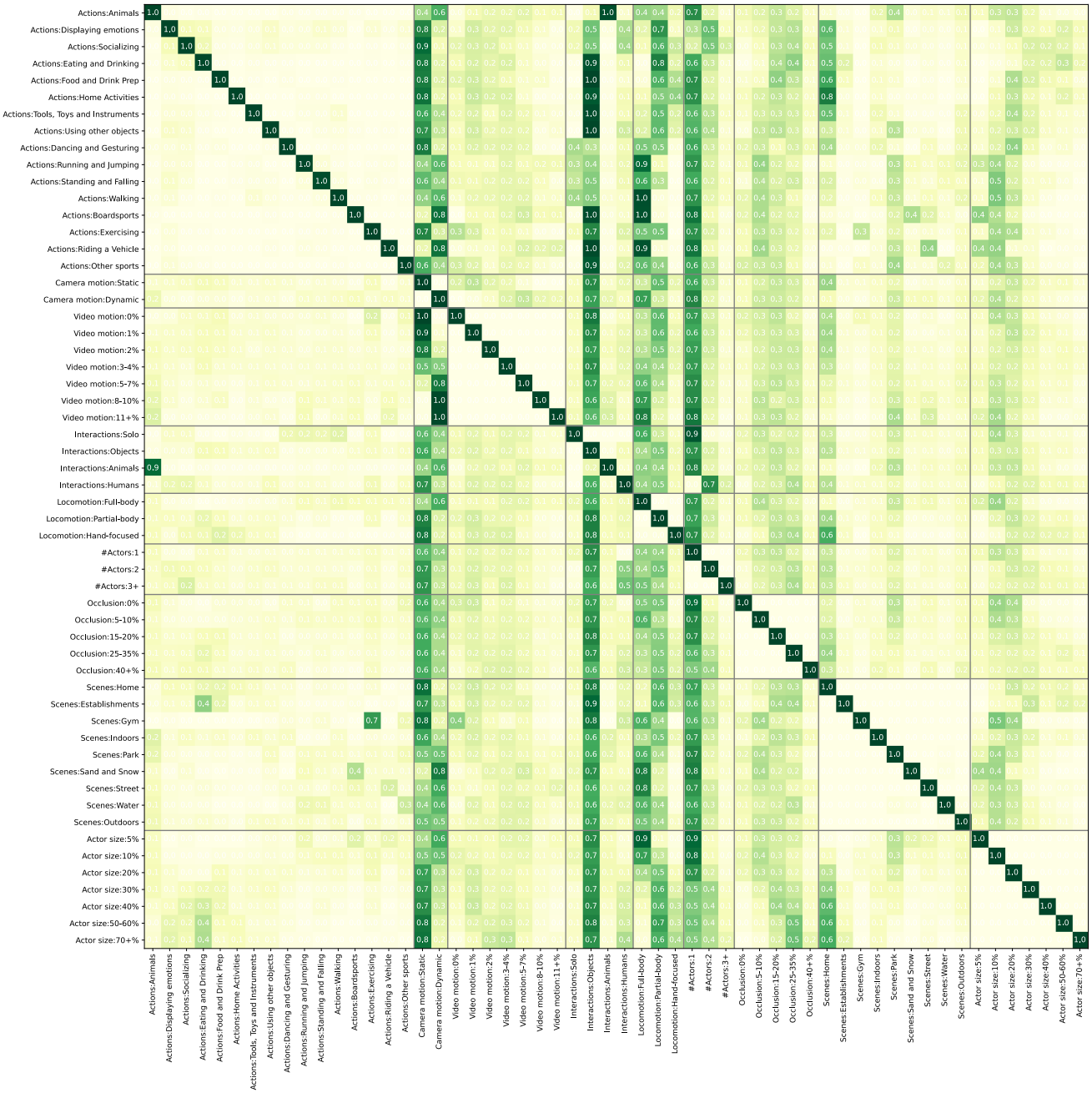

Figure 41. **Percentage of videos overlapping between two categories.** For a given row, computes how many of its videos (in percentage) are also available in another category (column). Best viewed on a screen due to its large number of entries.

C. Additional results

In this section, we report complementary results from our experiments.

Fig. 42 reports performance for our depth- and edge-conditioned models, while Fig. 43 shows the difference in errors when adding captions as an additional source of guidance to these models.

Figs. 44 to 46 report our investigations of pose-conditioned models w.r.t. (i) auto-encoder reconstruction capabilities, (ii) role of different pose detectors, and (iii) the need for human-labeled pose key-points.

Figs. 47 to 50 report category-level performance of our best models (MimicMotion, ControlNeXt and TF-T2V) according to sample-level metrics (ICD, OFE, pICD, pAPE).

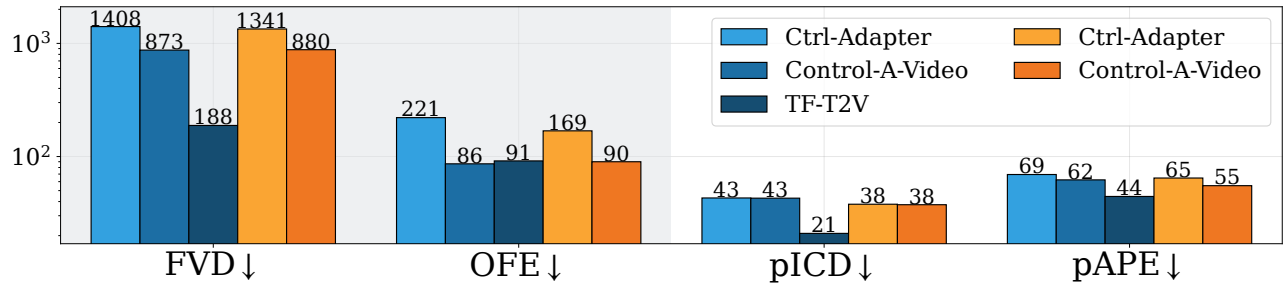

Figure 42. **Overall performance (left: video-level, right: human-level) of SOTA controllable depth- and edge-conditioned image-to-video models on WYD$_{16}$.** Depth models are shown in blue, while edge models in shades of orange. TF-T2V obtains the overall best performance across our models.

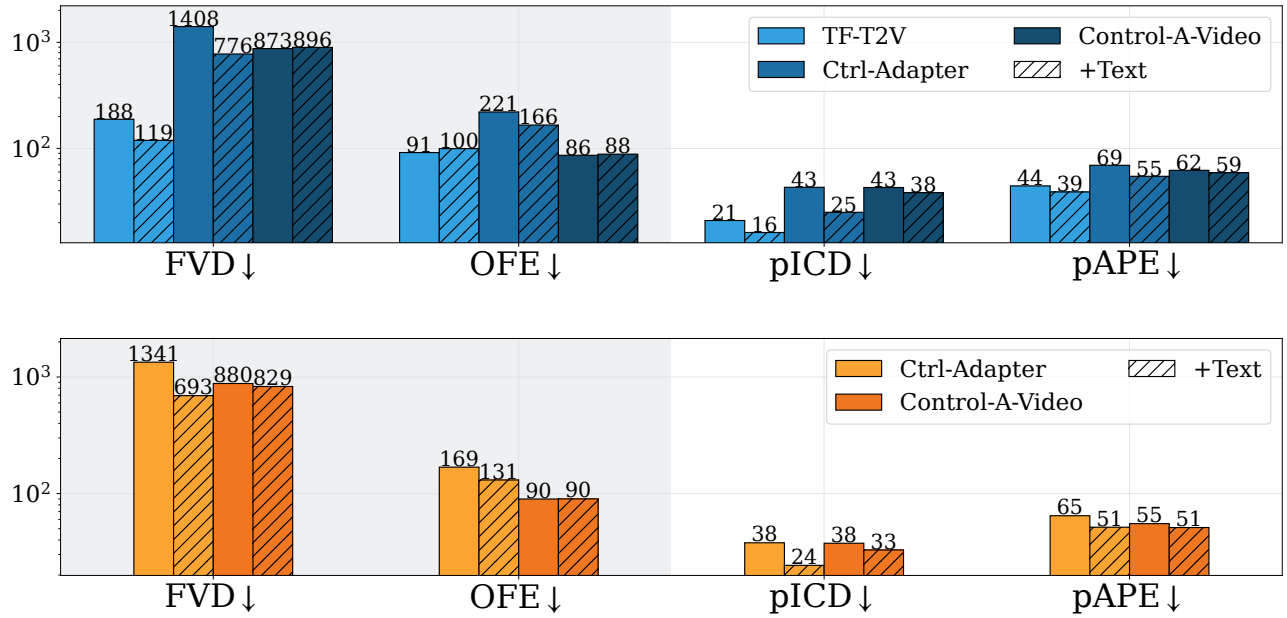

Figure 43. **Difference in errors in WYD after adding captions to depth- and edge-conditioned models.** Adding text guidance usually improves performance of depth-guided (top) and edge-guided (bottom) models.

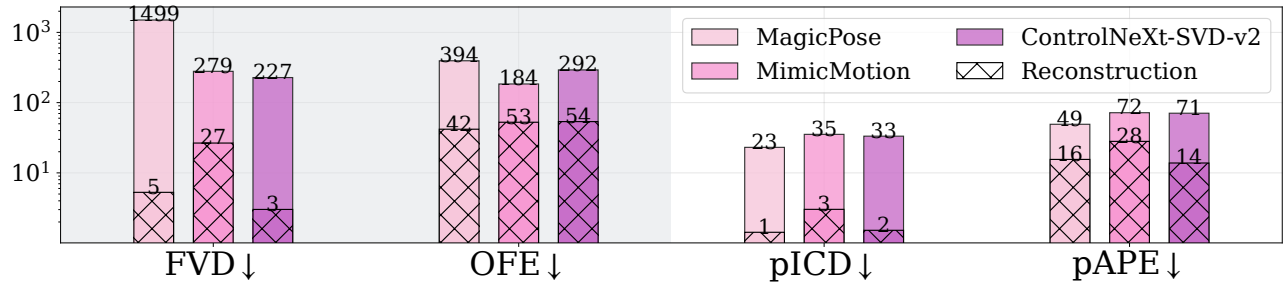

Figure 44. **Performance comparison between generation and auto-encoder reconstruction capabilities of pose-conditioned models on WYD$_{16}$.** We see a clear gap between generation and reconstruction across all metrics, showing that models are capable of generating the reference videos but struggle during the generation process.

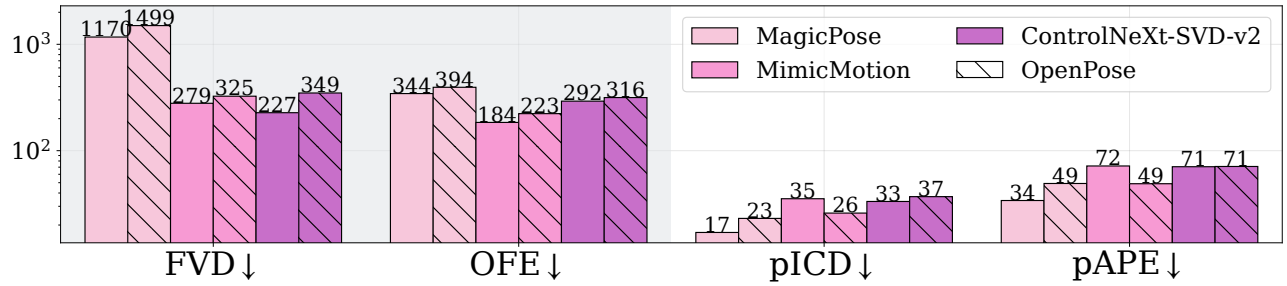

Figure 45. **Performance of pose-conditioned models when using the OpenPose detector rather than DWPose (default) on WYD_{16}.** We see that using DWPose gives typically better performance, even when applied to MagicPose, which was trained using OpenPose.

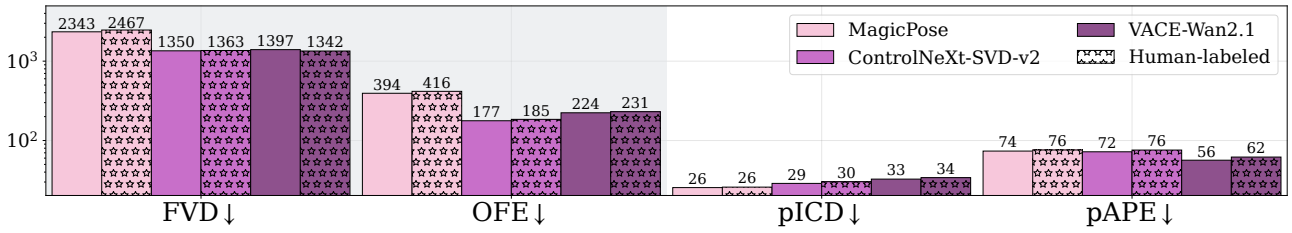

Figure 46. **Performance of pose-conditioned models when using the human-labeled 2D body key-points rather than DWPose (default) on a subset of 100 videos from WYD_{16}.** We see that models achieve similar performance in these two scenarios, verifying the correctness of our results based on poses detected with DWPose.

Figure 47. **Performance of best models w.r.t. '#Actors,' 'Actor size' and 'Actor occlusion.'** Animating multiple actors is harder than a single one. Small humans are also harder to generate precisely compared to when they cover a large portion of the frame. Performance also tends to degrade as the amount of occlusion increases.

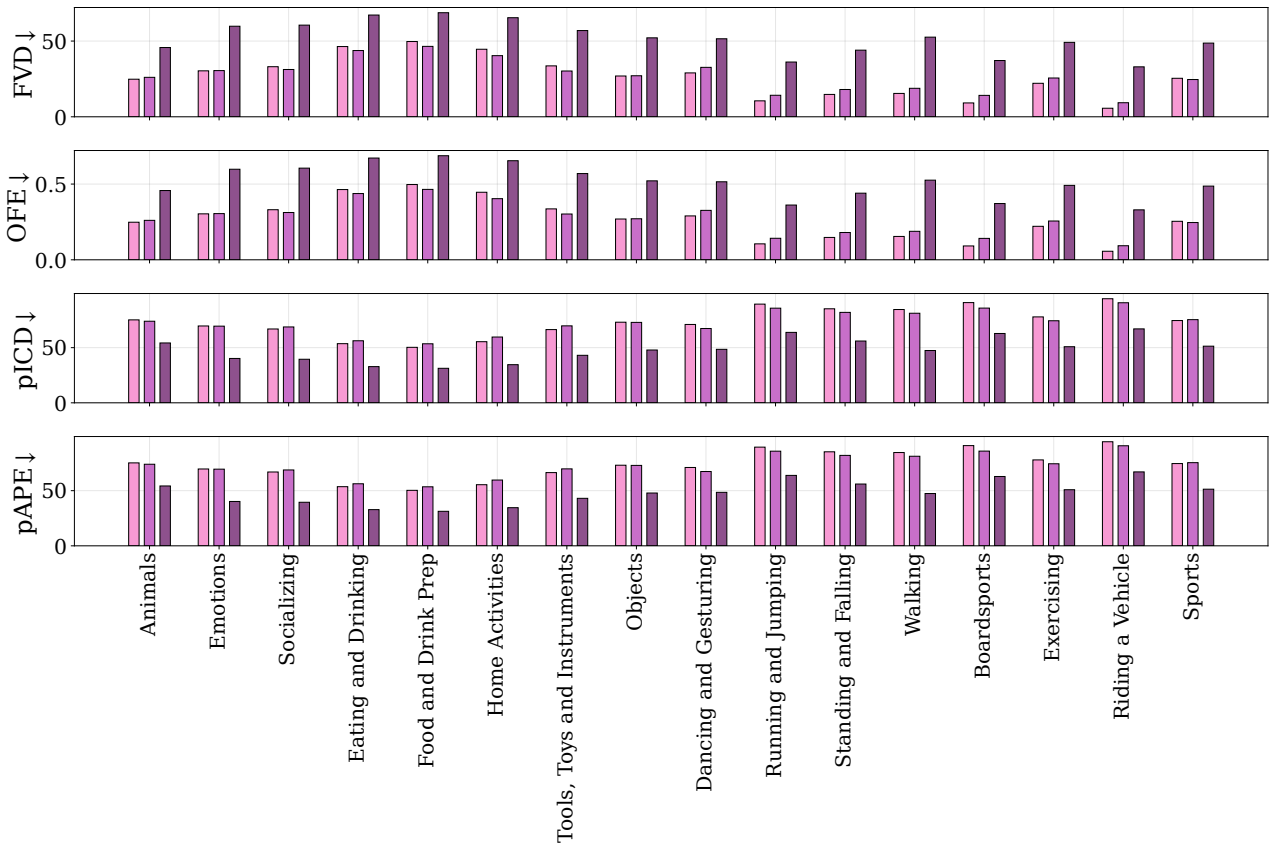

Figure 48. **Performance of best models w.r.t. 'Actions.'** Actions involving animals, riding a vehicle, running and jumping, and boardsports are challenging for SOTA models. Atypical movements, *e.g.*, standing up and falling down, are also hard.

Figure 49. **Performance of best models w.r.t. 'Locomotion,' 'Camera motion' and 'Video motion.'** Videos with full-body locomotion are more challenging to generate due to the larger changes required. Dynamic videos with high levels of motion are more challenging.

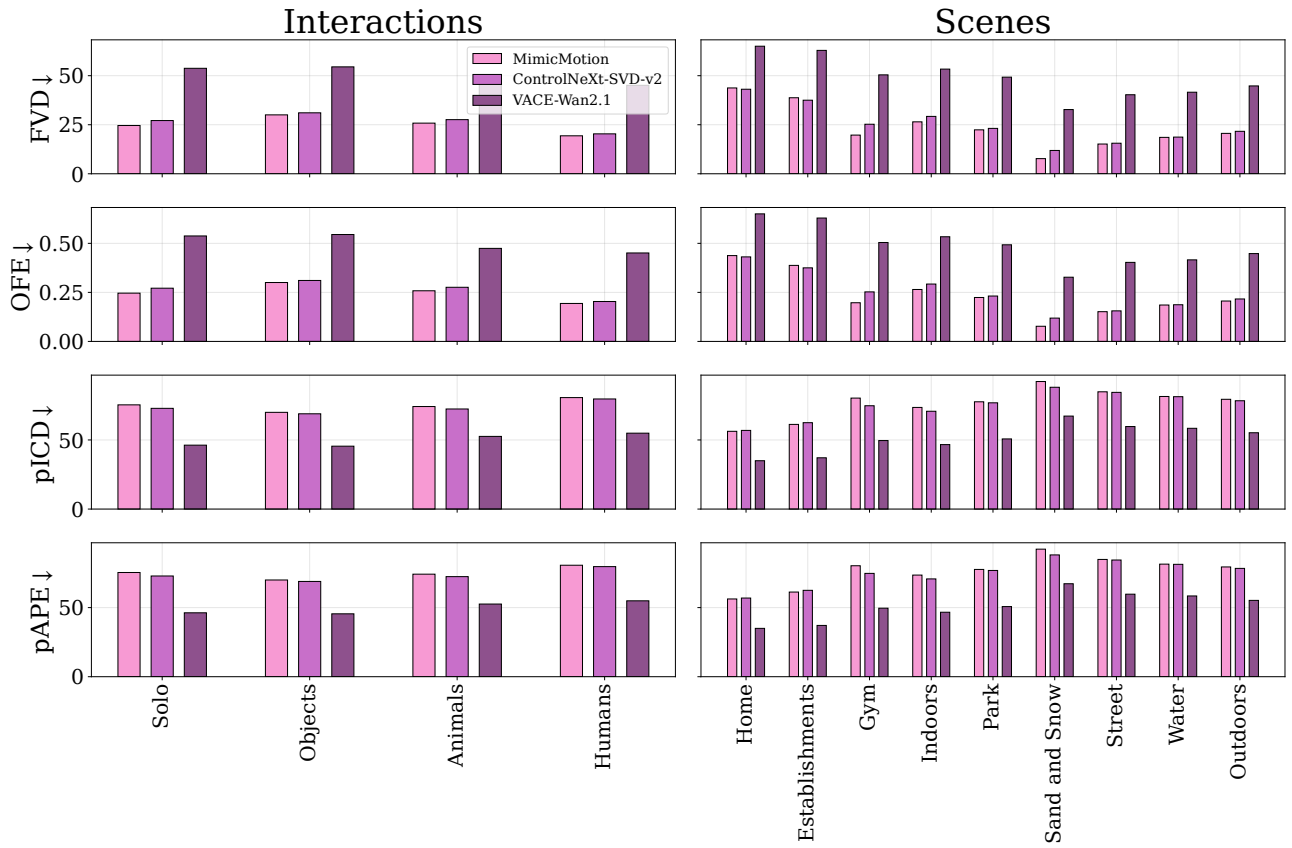

Figure 50. **Performance of best models w.r.t. 'Interactions' and 'Scenes.'** Generating videos of humans interacting with animals or other humans is more difficult than solo videos. Outdoors scenes (*e.g.*, on sand and snow, street, by the water) are also harder for SOTA models.

Figure 51. **UI used to collect human preferences in side-by-side evaluations for WYD.** We remove the actors' segmentation masks when comparing video motion similarity. No reference video is shown for video quality comparisons.

D. Evaluations

In this section, we report the performance of our evaluated models according to several automatic metrics and detail our human evaluation protocol.

Automatic metrics. Figs. 52 to 54 show how additional metrics score our seven models w.r.t. overall and human video quality, frame-by-frame similarity and video motion.

We note that, when computing automatic metrics, the pre-processing operations of a given video generation model are also applied to the reference videos. This is because current models were not initially designed to support all aspect ratios, and comparing their generations with the original reference videos would result in unfair evaluations. For instance, ControlNeXt generates portrait videos only. If we had not applied the same transformations to the reference videos, we would not have been able to compute pixel-level metrics. Moreover, this ensures that the results obtained with pAPE capture scale differences (despite model-specific resolutions) and correctly identifies the re-scale and re-center issue in MimicMotion and ControlNeXt described in Sec. 6. We encourage future work to develop models that can process different aspect ratios so as to ensure fully comparable results and benchmarking on WYD.

Side-by-side human evaluations setup. For side-by-side evaluation, we sample 100 random videos from WYD and ask four researchers familiar with the task. Each researcher annotates 25 comparisons for each model pair and across four evaluation setups: (i) video quality, (ii) video motion similarity, (iii) human quality w.r.t. reference, and (iv) human motion w.r.t. reference.

For video quality, we compare all 21 model pairs; while we only compare five model pairs (MimicMotion vs. ControlNeXt, MagicPose vs. ControlNeXt, MagicPose vs. Ctrl-Adapter, MimicMotion vs. TF-T2V, TF-T2V vs. Control-A-Video) on the other tasks due to the increased amount of time required for careful evaluations.

An example of our UI for side-by-side evaluations is shown in Fig. 51, where we remove the actors' segmentation masks for setup (ii), and only show the generated videos for setup (i). We used the following the templates for evaluation.

1. **Video quality:** Choose the video that you think is of highest quality (less defects, distortions, artifacts, excessive blur, etc.). If both have the same quality, choose the one that is more appealing to you (more interesting, better composition, etc.). If both videos are equally appealing, click on "Equally Good/Bad."
2. **Video motion similarity:** Choose the video that you think best matches the motion of the entire reference video (shown in the middle). Please try to ignore potential defects or bad quality of the videos. If both videos equally follow the motion of the reference video, click on "Equally Good/Bad."
3. **Human quality w.r.t. reference:** Choose the video that you think has the highest quality (less defects, distortions, artifacts, excessive blur, etc.) for the people highlighted in the reference video (shown in the middle). If the highlighted people in both videos are equally good/bad, click on "Equally Good/Bad."
4. **Human motion w.r.t. reference:** Choose the video that you think best matches the movements of the people in the reference video (shown in the middle). Please try to ignore potential defects or bad quality of the videos. If both videos equally follow the motion of the reference video, click on "Equally Good/Bad."

E. Ethics statement

The aim of *What Are You Doing?* (WYD) is to enable better evaluation of current and future controllable video generation models with respect to human characters and motion, which arguably are of particular importance to people. While this kind of models have great potential to assist and augment human creativity, there are broader societal issues that need to be considered when developing these models.

Video generative models may be misused to generate fake, hateful, explicit or harmful content. For example, they could be used to spread misinformation and portray false situations by synthesizing fake content (*i.e.*, deepfakes). To mitigate these harms, digital watermarks can be applied to generated videos [48] to identify whether a given video was produced by a particular model.

Generative models rely on massive amounts of data harvested from the Web, which reflect social stereotypes, oppressive viewpoints, and harmful associations to marginalized identity groups [3, 4, 52]. It is essential that generated content avoids perpetuating harmful stereotypes and respects cultural sensitivities. In fact, models primarily trained on samples with English data may reflect Western cultures [44, 58]. We acknowledge that our benchmark does not explicitly aim to encompass several cultures and populations, and it may perpetuate biases present in the datasets on which it is based. We encourage future work to develop training and evaluation setups that aim to widen the social and cultural representations of these technologies.

Moreover, we note that training video generative models is computationally expensive, both financially (*e.g.*, hardware and electricity) and environmentally, due to the carbon footprint of modern tensor processing hardware. We encourage future research that explores more efficient architectures.

Due to the impacts and limitations described above, we remark that WYD aims to measure progress in video generation research. By no means should our data be extended for use in sensitive domains. We believe that generative technologies, like the type of controllable image-to-video models that can be evaluated in WYD, can become useful tools to enhance human productivity and creativity.

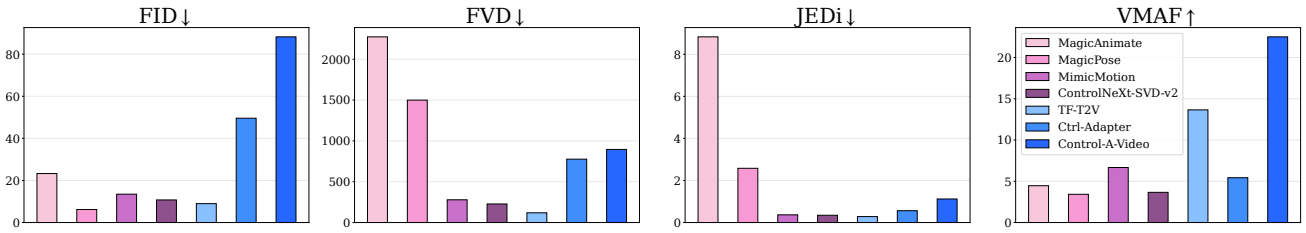

Figure 52. **Comparison of video quality metrics on WYD$_{16}$.** FID favors MagicPose's flickering videos, and VMAF ranks Control-A-Video's videos with distortions and artifacts first. FVD and JEDi rank video generations with high agreement to human judgments.

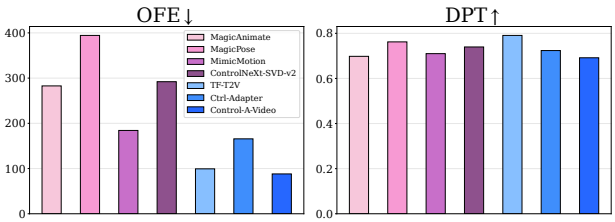

Figure 53. **Comparison of video motion similarity metrics on WYD$_{16}$.** We see that a depth-based metric (DPT) ranks the flickering generations from MagicPose as the second best ones, and those of pose-guided ControlNeXt better than those of depth-conditioned Ctrl-Adapter and Control-A-Video, which better generate videos with dynamic camera. OFE better agrees with the human rankings.

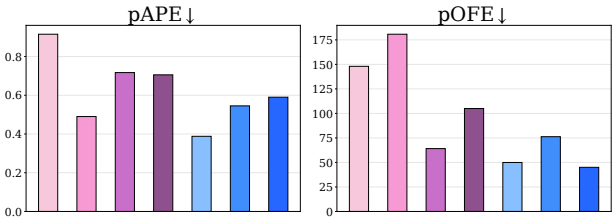

Figure 54. **Comparison of people motion metrics on WYD$_{16}$.** MimicMotion and ControlNeXt do not achieve good pAPE despite good visual quality (FVD). Analyzing further, we find that they always re-scale and re-center the generated humans (see Fig. 11)