ENABLING CONTROLLABLE, IDENTITY PRESERVING, NON-RIGID EDITS IN HUMAN-CENTRIC IMAGES

Nikolai Warner[1], *Jack Kolb*[1], *Meera Hahn*[2], *Jonathan Huang*[2], *Irfan Essa*[1,2], *Vighnesh Birodkar*[2]

[1]Georgia Institute of Technology, Atlanta, GA
[2]Google, Inc., Mountain View, CA

ABSTRACT

We approach the problem of inserting a person into a novel scene and controlling their pose via text guidance. Given an image of a person, a masked image of a scene, and a text description of the target pose, our model presents realistic images while being highly controllable. We validate the robustness of our model's true-to-text accuracy and identity preservation via a user study on in-the-wild images. In addition, we present a novel dataset containing pairs of frames from human-centric and action-rich videos, with text captions of the difference in human pose between frames. We also explore the challenges of controllable identity preservation for in-the-wild scenes and the failure modes of similar models. Our methods achieve a 10% increase in pose adherence (PCKt@0.5) over comparable methods without compromising visual fidelity, and show a clear qualitative improvement.

***Index Terms*—** Multimodal Conditioning, Non-Rigid Image Editing, Generative Image Models, Identity Preservation

1. INTRODUCTION

Generative diffusion models are an essential tool for creating high-quality images, videos, and 3D models. A core feature of generative models is "inpainting", or the ability to edit images in a user-controlled manner through textual, visual, or other inputs. This enables user-defined modifications without the need for detailed manual editing or full image re-generation. While the community has seen remarkable advancements in this space, virtually no works have approached editing human poses through text instructions or explicit pose keypoints while maintaining the subject's identity. In this paper we use generative diffusion to conduct complex, non-rigid pose edits that address this challenging class of modifications.

In addition, current editing methods do not generalize well to in-the-wild images and video. Such content often contains challenging backgrounds or blurry perspectives, presenting difficulties for optimizing to identity preservation and prompt adherence. We are interested in this domain for its potential in consumer applications.

Prior work has organized the complexity of human subject image editing into rigid and non-rigid edits [1]. Rigid edits preserve the subject's pose and alter their appearance (e.g., changing their shirt), while non-rigid edits modify the pose and may preserve appearance (e.g., generating yoga poses from a reference image). Non-rigid edits are complex as they involve deforming the subject in a natural and coherent manner. Furthermore, non-rigid edits can include object interactions, requiring that the model learns natural interactions across a diverse range of object classes.

To the author's knowledge, no public datasets contain image pairs with scene difference captions, and no related works support inpainting of text-guided pose while maintaining subject identity.

Fig. 1. **Our model enables controllable, identity-preserving edits to in-the-wild data**. Given a masked insertion scene and a reference image containing a person, our model inserts the person into the scene as controlled by a text caption. The results shown were made with held-out images using the captions below each row.

To address the literature gap, this paper contributes:

1. **Improved Pose Adherence with Photorealism:** By modeling a pose distribution aligned with the diffusion process using start and end poses, we obtain a pose adherence (PCKt@0.5) of .62 (text-guided) and .73 (keyframe-guided) to ground truth scenes.
2. **Text-Guided Non-Rigid Editing:** We enable complex, text-controlled pose edits. Non-rigid edits maintain subject and background appearance, changing their pose or location in a scene. We evaluate our qualitative performance via a user study.
3. **Reduced Hallucinations in Automated Scene Captions:** We curate a dataset of scene difference captions using a multimodal LLM, and reduce hallucinations by including pose estimation and prompt tuning.

We construct our dataset through automating scene-change captioning on large human-centric video datasets. We use pose estimators to extract useful caption pairs by identifying meaningful pose changes, and reduce hallucinations in captions through few-shot prompting and the inclusion of filtering criteria based on pose estimation. All code and data is available at (project webpage URL).

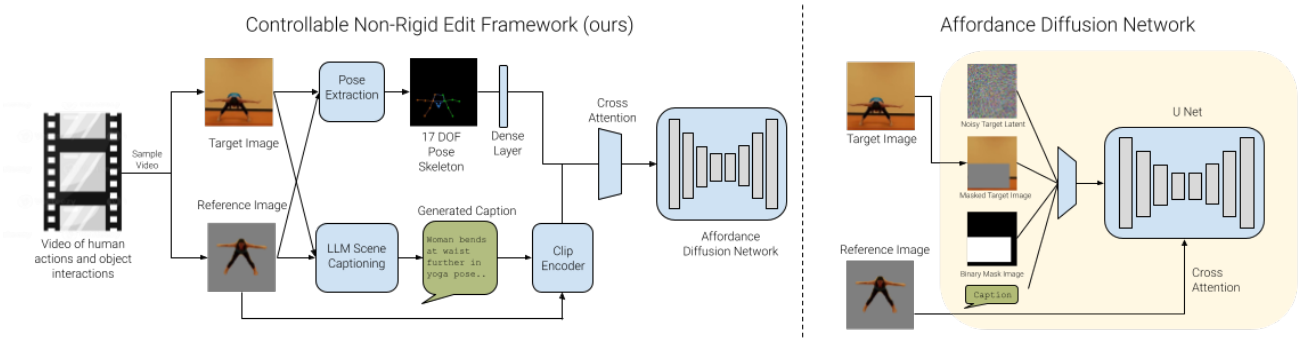

Fig. 2. **By feeding poses from the reference and target into a learned embedding, we force the network to balance photorealism and prompt adherence**. The system diagram shown illustrates the process of generating a desired edit using multiple inputs including noise target latent, binary mask, masked target latent, reference image, and a text caption. The pose embeddings are implicitly aligned with CLIP through our training process. The Affordance Diffusion Network builds upon the formulation proposed by [1], to which we introduce controllability via the framework detailed on the left. No related work effectively combines controllable non-rigid edits with identity preservation.

2. RELATED WORKS

Recent years have seen high research interest in improving the visual quality and controllability of images produced by generative models.

2.1. Text-Based Control & General Controllability

The original latent diffusion paper introduced conditional control through the use of cross-attention layers in the U-Net [2]. Combined with performing diffusion in the latent space with smaller spatial dimensions, this enabled a variety of control signals and auxiliary tasks including text-to-image, layout-to-image, inpainting, and super-resolution. Since then, works have greatly expanded upon the performance of image generation.

Prior works have approached text-based control of images. DreamBooth allows editing various image elements, such as the background, style, and accessories, while maintaining the image's primary subject [3]. The model's authors fine-tuned existing text-to-image diffusion models using a few-shot dataset of the subject to insert them into new domains. Imagic introduced a three-stage approach to enable non-rigid pose edits by performing linear interpolation between the optimized and target text embeddings [4].

ControlNet aims to address the issue of catastrophic forgetting [5]. The model uses a reference image and a text prompt, and is capable of non-rigid edits, however it does not preserve identity. MASACtrl enables text-driven nonrigid edits in a tuning-free manner via "masked mutual attention", however is sensitive to parameters chosen to control where in the diffusion process the masked mutual attention occurs [6]. While neither model has identity preservation, we use ControlNet and MASACtrl as baselines in this work.

2.2. Pose-Based Control

Very few works have approached non-rigid edits given a target pose. PIDM incorporates a noise prediction model and a texture encoder to maintain the subject's style given a target pose [7]. The model is trained on DeepFashion [8] with $\sim$52,000 images, however is brittle outside that domain. PCDM fine-tunes a pre-trained latent diffusion model [9]. The model aligns the source image and target pose through a three-layer trainable pose network to project the source and target poses into a latent pose embedding. We use PIDM as a baseline for this work despite its narrow scope.

2.3. Identity-Preserving Diffusion

Recent advancements in identity-preserving diffusion models have maintained subject identities while enabling detailed and flexible image editing. InstantID uses a novel IdentityNet module that integrates spatial conditioning with a diffusion model for identity preservation using a single facial image [10]. He et al. introduced regularization in the dataset generation, enhancing identity preservation across various text-to-image models [11]. Banerjee et al. proposed a technique for simulating aging and de-aging in face images using latent text-to-image diffusion models, maintaining biometric identity with high photorealism [12]. These methods maintain subject identities, however are designed for non-rigid edits (i.e., style transfer).

2.4. Non-Rigid Identity Preservation

Only one prior work has accomplished non-rigid pose changes with identity preservation on in-the-wild data. Kulal et al. showed that an inpainting diffusion framework can train a model to hallucinate a plausible and photorealistic insertion of a person from one scene into another [1]. For reference images, the model employs a nearest-neighbor approach where frames are sampled from a large meta-dataset of person and activity-centric videos [13]. However, the model lacks controllability, used proprietary data, and no model checkpoints are available.

As Kulal et al's objectives are similar to ours, we re-implement Kulal's methods on publicly available data. To enable controllability we train the model on our dataset of automated captions and extracted poses from frames, and use additional filtering criteria to improve the quality of resulting images.

3. METHODS

3.1. Conditional Diffusion Models

Latent diffusion models operate in a compressed, lower-dimensional latent space rather than the pixel space, making the generation process more efficient. By performing diffusion in this latent space, these models can handle higher resolution images and complex transformations with reduced computational cost. This approach leverages pre-trained autoencoders to map images to and from the latent space, maintaining high image fidelity while optimizing processing time.

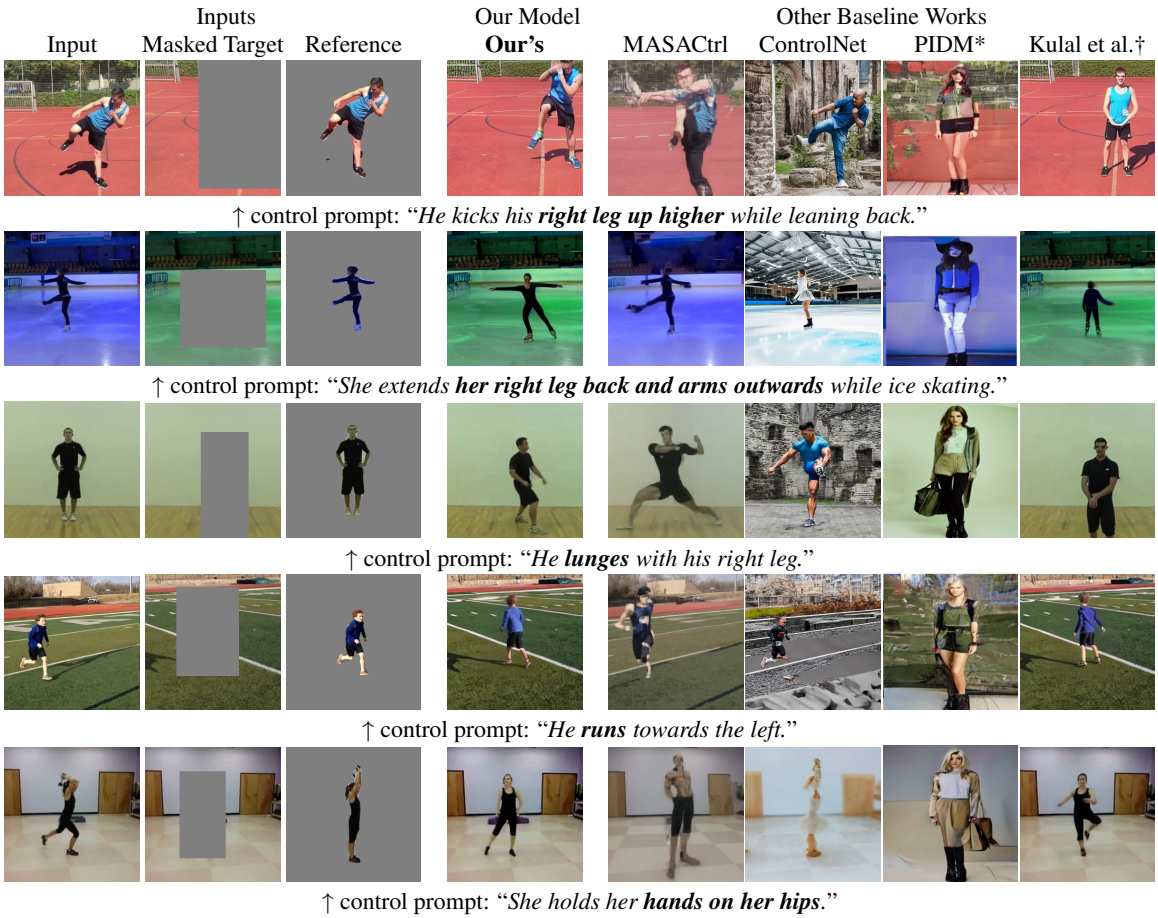

Fig. 3. Most baselines fail to preserve the subject identity on in-the-wild data under complex edits. No other baseline both preserves identity and performs controllable insertion. Comparison of our approach to baselines for identity preservation and controllability of in-the-wild images. The input image is controlled using the prompts below each row, and the human subject is transferred to new frames for relevant models (our's and Kulal et al.). Baselines either insert a person without controllability (Kulal et al.), or are controllable but fail to generalize to in-the-wild images (MASACtrl, PIDM). Our approach maintains similar photorealism to Kulal et al., with improved controllability. Note the differences in caption adherence bolded below. Zoom in for details.
* PIDM works well on its training dataset but is brittle in the wild, likely due to its fashion-related training dataset.
† Kulal et al. results are from the re-trained model with image conditioning only.

Conditional diffusion models enhance the basic diffusion process by incorporating additional information to guide the generation. This conditioning information can be in the form of text descriptions, class labels, or other metadata. The objective is to steer the diffusion process to generate images that meet specific criteria or exhibit particular characteristics defined by the conditioning input.

3.2. Inpainting Diffusion Formulation

We frame affordance diffusion learning as an inpainting diffusion fine-tuning problem. Person-centric video datasets are collected and processed such that only one person is present in each video and redundant frames with little motion are removed. We then fine-tune an inpainting diffusion checkpoint from Stable Diffusion on sampled pairs of images from these videos.

Each video must meet two pose detection criteria: a single pose skeleton must be present in each sampled scene, and the majority of joints must be visible. For each pair of sampled frames, one frame is used as the masked target where the person is masked out, and the other frame is cropped to include only the person. This cropped frame is then used to preserve the identity of the person via cross-attention mechanisms. By focusing on pairs of frames with significant motion, the model learns to inpaint the masked target image while maintaining the identity and pose of the person from the reference frame. For more details, refer to [1].

3.3. Enabling Controllability from Noisy Supervision with Scene Difference Captions

Previous literature has trained models to hallucinate a plausible way to insert a person into a scene [1, 13]. Given a scene with a masked area to inpaint, and a segmented person to insert, the model learns plausible ways to insert a person into a scene. However, this suffers from limited controllability. For example, there are multiple ways a child could interact with a slide in a playground.

We fine-tune the inpainting checkpoint to make image edits that respect textual prompts. From preprocessing the Kinetics-700 dataset [14] for pose detection and fidelity, we obtain 5,787 cap-

tioned videos. We also collect 7,700 annotated image pairs from videos in NTU-RGBD [15], and caption pairs from Charades [16] and Fit3D [17]. We use GPT-4V [18] in a 10-shot manner to caption the difference between scenes given two images. This is often an ill-defined problem, as depending on the degree of difference between two frames, conflicting captions are possible. For example, an image pair could show dropping an object or reaching to pick it up. Images were sent as side-by-side composite images instead of successive images, as we found it reduced GPT-4V's hallucination.

3.4. Reducing Hallucinations in Generated Captions

A key challenge was training on limited publicly available data to create photorealistic edits with identity-preservation and faithfulness to conditioning signals.

To facilitate learning complex non-rigid edits that generalize to in-the-wild data, we sample a handful of key frames per video instead of retaining all frames. Each keyframe is sampled according to several criteria. First, we use pose detection to remove frames that contain more than one person, or do not contain most of a pose skeleton present. We filtered out many frames from Kinetics-700 that contained close-up videos of partial pose skeletons, such as a person from the neck up or a zoomed in video of a hand. In addition, many videos contained multiple perspectives, where initially a single person was present but the zoom level changed.

We also specify a minimum pose distance between sampled frames equivalent to the length of the pose skeletons shoulder to head distance. We found that using a single-stage model such as OpenPose underperformed a two-stage *object detection + pose extraction model* such as RTMPose [19]. We use the pose information extracted during preprocessing to help guide the noisy captions during training for better person-object interactions. Image histogram similarity is used to set a minimum and maximum similarity between frames. Due to the small training dataset, frames must be similar enough that the background or context does not entirely change, but distinct enough to avoid redundant frames.

3.5. Improving Person-Object Interactions

To assist the model in learning a distribution of possible poses given an object, we add joint conditioning with pose data, as shown in Figure 2. At inference time, the model receives a masked scene to insert a person into, a reference person from the same video, a bounding box describing where to insert the person, and a text prompt describing how the person's posture or object interaction changes. This is a similar process to InstructPix2Pix [20], except the scene difference captions contain information about non-rigid deformations that do not respect the structure of the original person or photo.

We find that when we re-implement image-only conditioning as a baseline method, shown in Tables 2 and 3, identity preservation decreases when we focus on scenes that contain person-object interactions. This is because the model must focus on preserving the identity of the object. Because of our automated masking and segmentation pipelines, some object interactions may be described in the masked target image (e.g., half a bicycle), and some may only be described in the reference person crop (e.g., a kid holding a ball). When text prompts are generated they are relative to the unmasked target and reference images. Therefore, text and pose information are critical for successful modeling of person-object interactions.

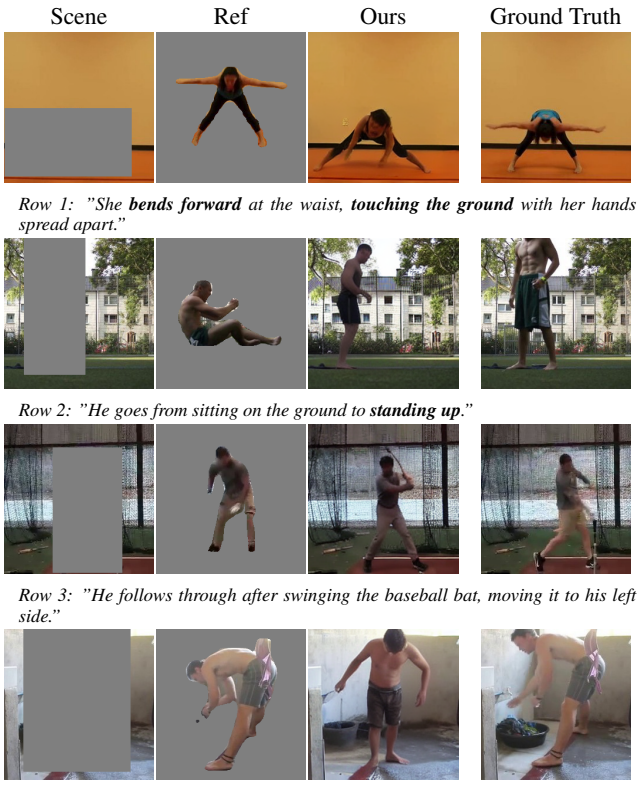

Fig. 4. **We achieve complex, non-rigid edits that preserve the subject and background identity**, while changing the pose or interaction of the user with the scene.

3.6. Inference Implementation Details

At inference time, we modify Classifier-Free Guidance (CFG) [21] for each of our conditioning setups. For image-only conditioning, representing a re-implementation of the Kulal et al. methods on our publicly available training data, we encode the unconditional representation as a tensor of all zeros in the same shape as the reference image conditioning with CLIP. The mask is expanded to cover the entire target image.

For image-text conditioning, we represent the unconditional reference image conditioning signal as a zeros tensor with the same shape as the reference image, encoded with CLIP. Then, we encode the unconditional representation of the scene difference caption as a null caption, encoded with CLIP.

For models with pose conditioning (image-pose, image-pose-text), we define an unconditional pose representation. An image with a neutral posture from the dataset is selected (a person standing straight in the center of the frame, with their hands at their side) and passed through a linear projection layer to obtain the same embedding dimension as CLIP. The resulting unconditional pose representation is then concatenated with the other unconditional signals.

4. EXPERIMENTS

Our architecture is shown in Fig. 2. CLIP's text encoder is used to enable multimodal joint conditioning and for compatibility with Stable Diffusion's embedding space. Using ViT-L/14 [22], the hidden

Method	Identity	In-The-Wild	Pose Control	Text Control	↓ FID@1000	↑ PCKt@0.5	↑ CLIP Sim.
ControlNet			✓	✓	81.3	0.79	0.70
Instruct P2P	✓			✓	61.4	0.00	0.80
PIDM	✓		✓		115.0	0.48	0.74
MASACtrl	✓			✓	55.2	0.00	0.84
Kulal*	✓	✓			**45.3**	0.63	**0.87**
Ours (Img, Pose)	✓	✓	✓		46.5	**0.73**	**0.87**
Ours (Img, Text)	✓	✓		✓	46.6	0.62	**0.87**

Table 1. **Our pose-conditioned model achieves a 10% increase in PCKt@0.5 score compared to similar methods, and no other models successfully preserve the subject's identity and controlled pose.** Despite the numerical similarity between scores, small differences in the CLIP cosine similarity or FID result in the loss of identity preservation (see Figure 3). Other methods that respect pose (ControlNet, PIDM) and text control (Instruct P2P, MASACtrl) struggle with identity preservation. In addition, most models fail on in-the-wild action-centric data, such as scenes from Kinetics, which is reflected in poor identity preservation and high FID. *We reimplement Kulal et al. for fair comparison.

Method	↑Identity	↑Controllable
Kulal	61%	Not Appl.
Ours (Img, Text)	55%	39%
Ours (Img, Pose)	63.5%	**57.5%**
Ours (Img, Pose, Text)	**68.5%**	51%

Table 2. **Our text-conditioned model is the first to allow for controllable, nonrigid edits on in-the-wild data, respecting the target prompt up to 51% of the time, despite limited and noisy training data.** To obtain the metrics shown, we had eight participants rate whether images preserve the subject identity and adhere to the text caption. Adding text or pose improves identity preservation and helps respect the control signal.

state contains 257 channels for the image encoder and 77 channels for the text encoder. To facilitate identity preservation, we use the last hidden state of the image and text encoders instead of the final encoded vector, which maintains an associated channel dimension. A linear layer projects the image encoder's last hidden state from 1024 dimensions to 768 dimensions, to match the text encoder.

To enable pose conditioning we use 2D, 17-DOF skeletons. To obtain the poses, RTMPose was selected over OpenPose as it performs well on the blurry and complex scenes found in Kinetics. Each pose keypoint contains a predicted x, y coordinate as well as a confidence score. We concatenate and flatten the pose into a $(1, 51)$ vector and project it through a linear layer into a 768 dimension embedding space. The pose, image, and text embeddings are concatenated into a $(batchsize, 335, 768)$ tensor.

4.1. Evaluation

Our objective is to enable complex non-rigid pose edits driven by text prompts. In addition, we are interested in improving the quality and contextual awareness of person-object interactions.

To quantify the realism of the generated images, we first evaluate on traditional metrics including FID (photorealism) [23], PCKt@0.5 (pose adherence) [24], and CLIP cosine similarity across all four model configurations. The numerical results are shown in Table 1. However, the metrics are imperfect for measuring pose preservation across two images in a scene. While FID measures photorealism, the metric does not enforce identity preservation; and while PCKt measures pose adherence, our goal is to hallucinate plausible but controllable poses. Therefore, a model can learn an effective distribution for how to insert a person, but receive unremarkable PCKt scores. In absence of a metric to quantify identity preservation and pose believability, we conduct a user study to obtain human ratings for each configuration. Details are available in the Appendix.

4.2. Data Preprocessing and Training

In order to fine-tune the inpainting diffusion network, we first process video data into a series of self-supervised frame pairs. We found that for joint conditioning, the sampling criteria and procedure are important for obtaining high-quality data – after applying our filtering procedure we retained only a fraction of the original videos. We then condense clips down to 2-5 frames, depending on our minimum edit distance criteria. We also experimented with retaining additional clips from the NTU dataset and found that the model learned poor action distributions and would converge to the same identity pose. We attribute this to reduced variability in poses, as most videos feature people standing or sitting without performing actions. We follow the same masking and augmentation procedures as [1].

At training time, we fine-tune the Stable Diffusion checkpoint with four configurations: *baseline*, with *pose*, with *text*, and with *pose + text* conditioning. Each configuration was trained for approximately 600 epochs. Table 1 shows that all models maintained similar amounts of photorealism, however the image + pose model increases PCKt@0.5 by 0.10. Table 2 shows that human graders reported that the model trained with joint conditioning of the subject's identity, pose, and text held the best identity preservation at 68.5% accuracy, and that the accuracy of text controllability ranges from 39% (no pose conditioning) to 51% (with pose conditioning).

5. CONCLUSION

In this paper, we presented a novel method for controllable and complex, non-rigid image edits using multimodal conditioning and self-supervised learning from video datasets on in the wild data. Our approach demonstrates significant improvements in identity preservation and contextual accuracy in generated images, particularly in scenes involving human-object interactions. We explore the limitations of existing methods for non-rigid editing on in-the-wild data, particularly in identity preservation and pose adherence. We propose a novel approach that aligns pose distributions with the diffusion process, improving pose adherence without compromising photorealism. Additionally, we introduced a dataset with automated scene-difference captions and reduced hallucinations. Future works could explore how to better model person-object interactions in scenes.

6. REFERENCES

[1] Siddhesh Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jianwei Yang, Jiajun Lu, Alexei A. Efros, and Karsten Kreis Singh, "Putting people in their place: Affordance-aware human insertion into scenes," *arXiv preprint arXiv:2304.14406*, 2023.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," *arXiv preprint arXiv:1505.04597*, 2015.

[3] Lvmin Zhang, Qingyang Chen, Chengyi Chen, Yeping Huang, Zhe Gan, and Li Ma, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," *arXiv preprint arXiv:2208.12242*, 2022.

[4] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani, "Imagic: Text-based real image editing with diffusion models," *arXiv preprint arXiv:2210.09276*, 2022.

[5] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding conditional control to text-to-image diffusion models (controlnet)," *arXiv preprint arXiv:2302.05543*, 2023.

[6] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng, "Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing," *arXiv preprint arXiv:2304.06234*, 2023.

[7] A. K. Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan, "Person image synthesis via denoising diffusion model," *arXiv preprint arXiv:2304.14406*, 2023.

[8] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 1096–1104, IEEE.

[9] Fei Shen, Hu Ye, Jun Zhang, Cong Wang, Xiao Han, and Wei Yang, "Advancing pose-guided image synthesis with progressive conditional diffusion models," *arXiv preprint arXiv:2310.08563*, 2023.

[10] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu, "Instantid: Zero-shot identity-preserving generation in seconds," *arXiv preprint arXiv:2401.07519*, 2024.

[11] Xingzhe He, Zhiwen Cao, Nicholas Kolkin, Lantao Yu, Kun Wan, Helge Rhodin, and Ratheesh Kalarot, "A data perspective on enhanced identity preservation for diffusion personalization," *arXiv preprint arXiv:2401.07519*, 2024.

[12] Sudipta Banerjee, Govind Mittal, Ameya Joshi, Chinmay Hegde, and Nasir Memon, "Identity-preserving aging of face images via latent diffusion models," in *2023 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2023, pp. 1–8.

[13] Tim Brooks and Alexei A. Efros, "Hallucinating pose-compatible scenes," *arXiv preprint arXiv:2112.07933*, 2022.

[14] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.08103*, 2019.

[15] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] Jingran Zhang, Fumin Shen, Xing Xu, and Heng Tao Shen, "Temporal reasoning graph for activity recognition," in *arXiv preprint arXiv:1908.08648*, 2019, Presented at the Computer Vision and Pattern Recognition Conference (CVPR).

[17] Mihai Fieraru, Mihai Zanfir, Silviu-Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu, "Aifit: Automatic 3d human-interpretable feedback models for fitness training," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[18] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[19] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen, "Rtmpose: Real-time multi-person pose estimation based on mmpose," *arXiv preprint arXiv:2303.06244*, 2023.

[20] Tim Brooks, Aleksander Holynski, and Alexei A. Efros, "Instructpix2pix: Learning to follow image editing instructions," *arXiv preprint arXiv:2211.09800*, 2022.

[21] Jonathan Ho and Tim Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *arXiv preprint arXiv:1706.08500*, 2017.

[24] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele, "2d human pose estimation: New be nchmark and state of the art analysis," *arXiv preprint arXiv:1406.3863*, 2014.