MagicPose: Realistic Human Poses and Facial Expressions Retargeting with Identity-aware Diffusion

Di Chang[1,2] Yichun Shi[2] Quankai Gao[1] Jessica Fu[1] Hongyi Xu[2]
Guoxian Song[2] Qing Yan[2] Xiao Yang[2] Mohammad Soleymani[1]
[1] University of Southern California [2] ByteDance Inc.
https://boese0601.github.io/magicdance/
{dichang,quankaig,fujessic,msoleyma}@usc.edu
{yichun.shi,hongyi.xu,guoxian.song,qing.yan,xiao.yang}@bytedance.com

International Conference on Machine Learning (**ICML**) 2024

Presenter: Di Chang

Outline

- **Introduction & Problem Definition**
- **Recap Diffusion Model & ControlNet**
- **Motivation**
- **Method - Pipeline**
- **Results & User Study**
- **Conclusion & Future Work**

Outline

- **Introduction & Problem Definition**
- Recap Diffusion Model & ControlNet
- Motivation
- Method - Pipeline
- Results & User Study
- Conclusion & Future Work

Introduction

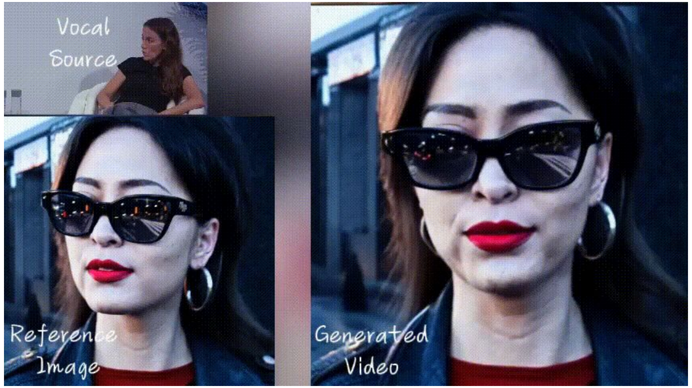
- AI Generated Content (AIGC) using **Diffusion Models** is the most popular topic in the recent Computer Vision Research Community.
- Some representative works include: **Stable Diffusion**(Text2Image), **Sora**(Text2Video), **EMO**(Audio2Video).
- Image & Video Generation is the most important application of Diffusion Model.

A stylish woman wears sunglasses and red lipstick.

Stable Diffusion - LMU Munich

A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage.

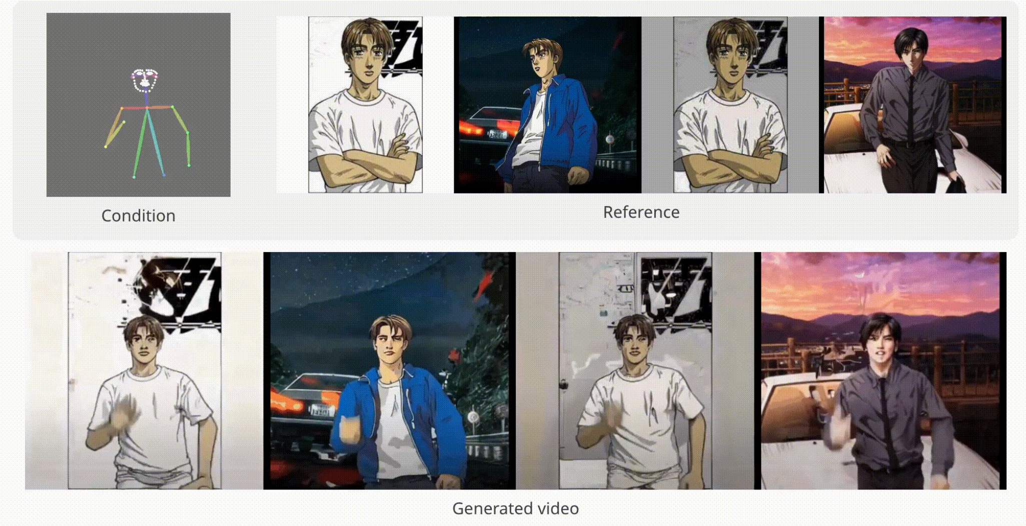
Sora - OpenAI



EMO - Alibaba

Problem Definition

- Human motion retargeting is the task of human image generation under the content control of a reference human subject **appearance** and the geometry control of **body motion and facial expression**.
- In this project, we propose **MagicPose**, an Image2Video model for motion retargeting with identity-aware diffusion.
- Such **Image2Image/Image2Video** technique can be applied to Virtual Reality, Creative Art Contents, Live Streaming, etc.



MagicPose - USC

Outline

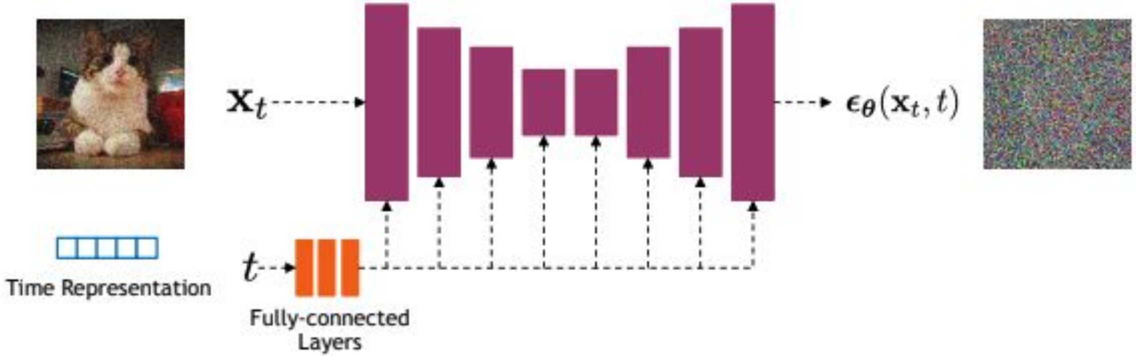
- Introduction & Problem Definition
- **Recap Diffusion Model & ControlNet**
- Motivation
- Method - Pipeline
- Results & User Study
- Conclusion & Future Work

Diffusion Model[1]

Diffusion(Adding Noise)

Denoise(Predicting Noise)

During Each Step t:

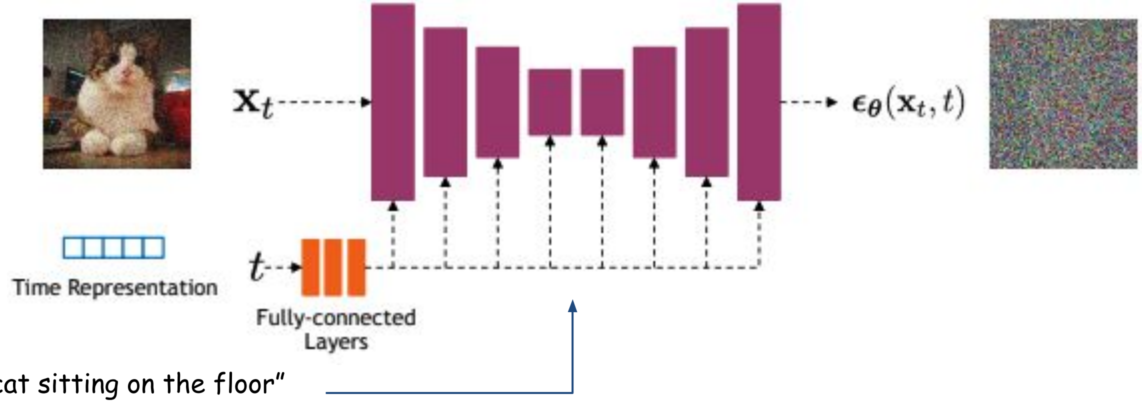


Stable Diffusion[2] (Latent Diffusion Model-LDM)

Diffusion(Adding Noise)

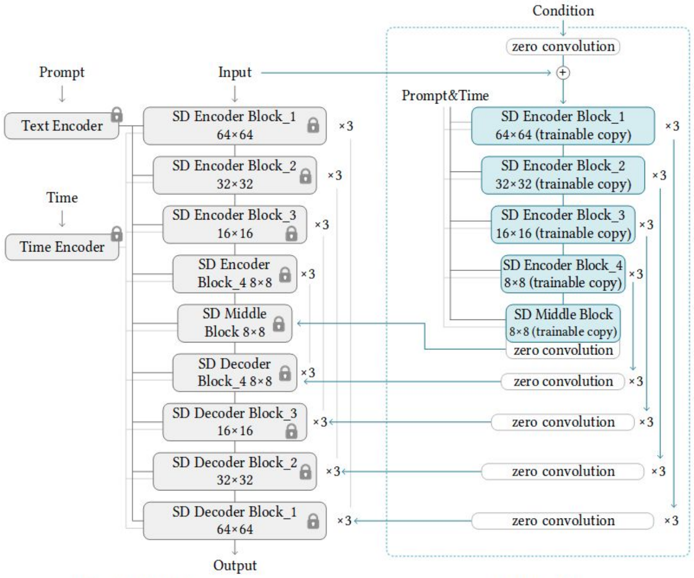
Denoise(Predicting Noise)

During Each Step t:

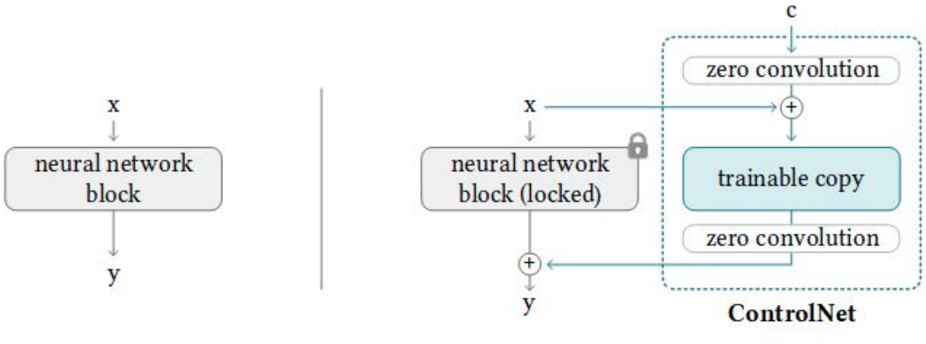


ControlNet[3]

- Text has already provided enough information for the content(appearance) of the object in the generation.
- Can we control the geometry of the object in the generation?
- Can we freeze the weights of pretrained Stable Diffusion Model (keep the model safe)?



(a) Stable Diffusion (b) ControlNet



(a) Before (b) After

Method - ControlNet[3]

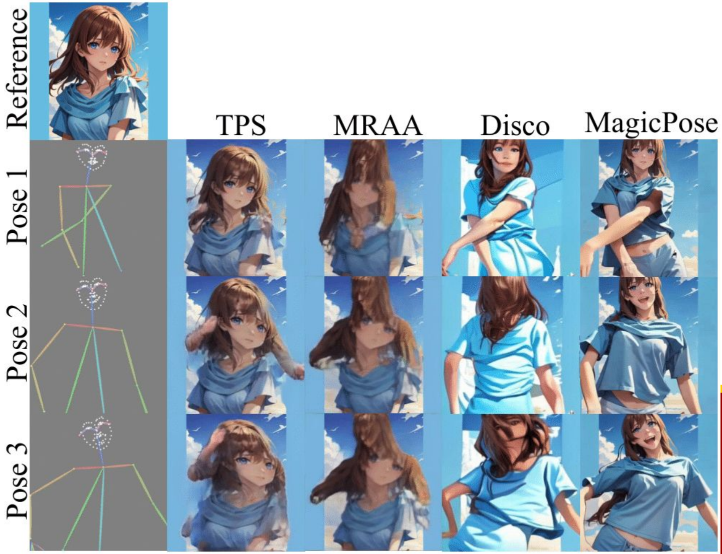
Outline

- Introduction & Problem Definition
- Recap Diffusion Model & ControlNet
- **Motivation**
- Method - Pipeline
- Results & User Study
- Conclusion & Future Work

Motivation & Key challenges and limitations

- Existing works based on diffusion model cannot provide satisfactory identity/appearance preserving ability for real-human image generation.

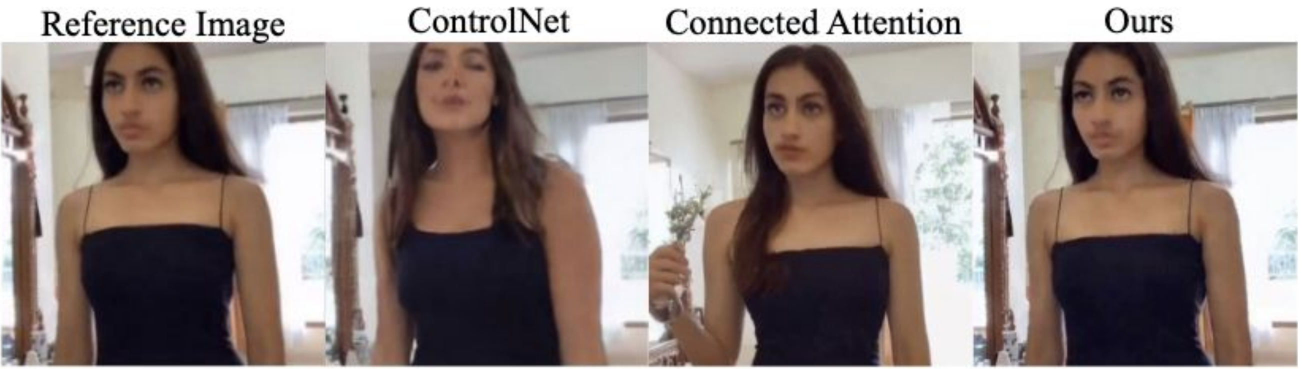
- Current works cannot generalize to unseen out-of-domain data after training on real-human data.

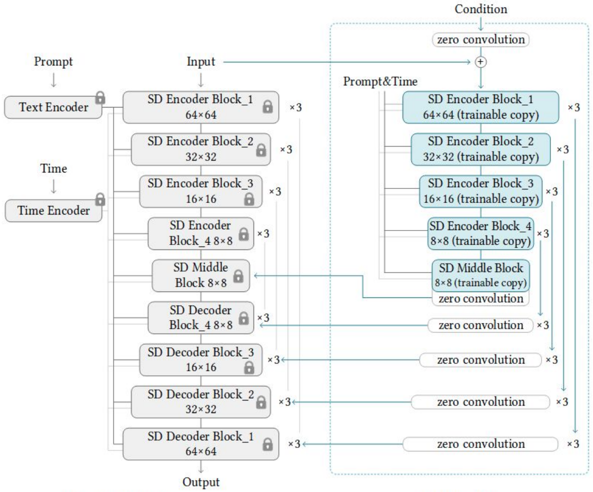


Outline

- Introduction & Problem Definition
- Recap Diffusion Model & ControlNet
- Motivation
- **Method - Pipeline**
- Results & User Study
- Conclusion & Future Work

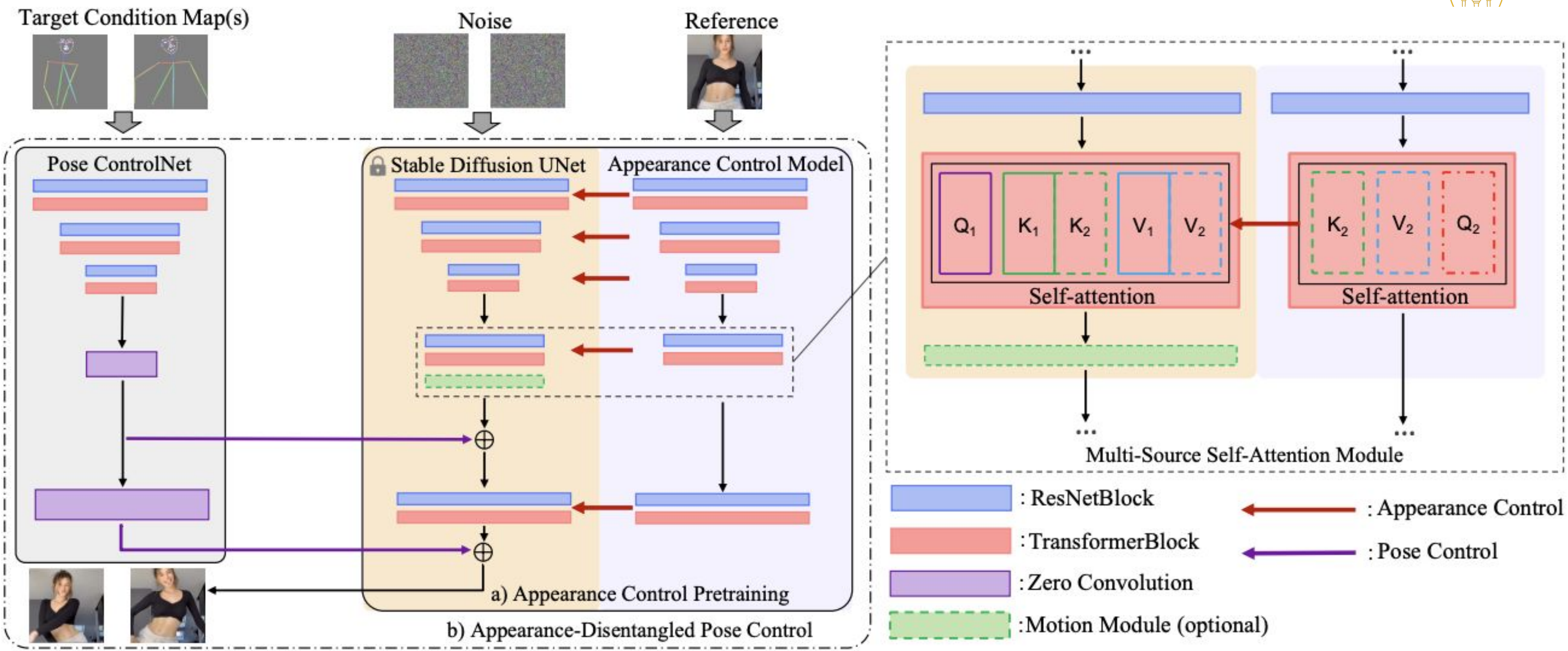
Method - Exploration of Appearance Control Mechanism





ControlNet

Method - Pipeline



Outline

- Introduction & Problem Definition
- Recap Diffusion Model & ControlNet
- Motivation
- Method - Pipeline
- **Results & User Study**
- Conclusion & Future Work

Experiment - Dataset

TikTok Dataset

- Consists of 350 single-person **dance videos** (with video length of 10-15 seconds). Most of these videos contain the face and **upper-body** of a human.

EverybodyDanceNow Dataset

- Consists of full-body videos of five subjects. Experiments on this dataset aim to **test** our method's generalization ability to in-the-wild, **full-body out of domain motions**.

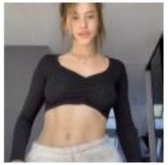
Self-collected Out-of-Domain Images

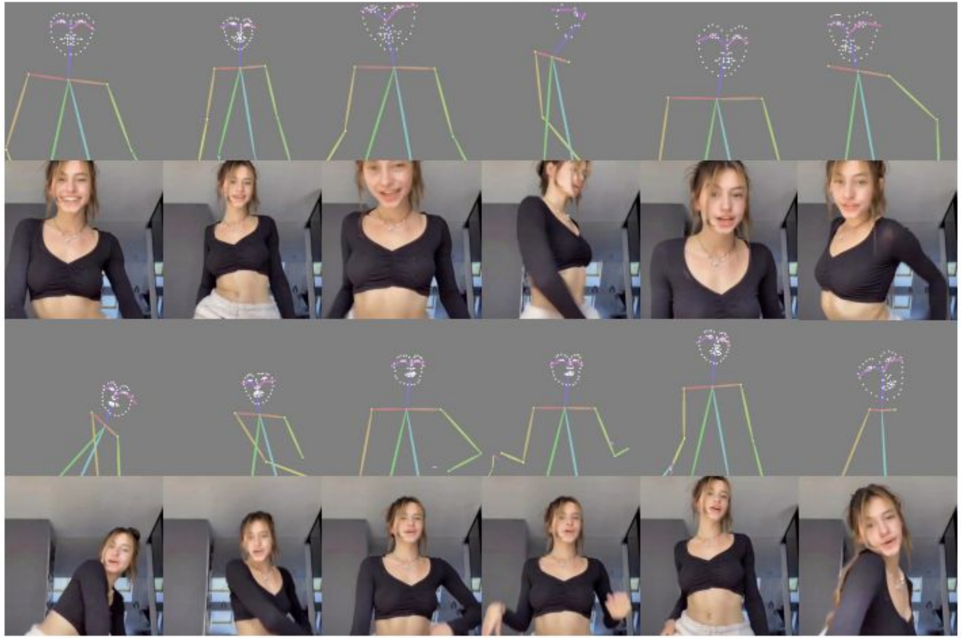
- Come from online resources. We use them to **test** our method's generalization ability to in-the-wild, **out of domain appearance.**

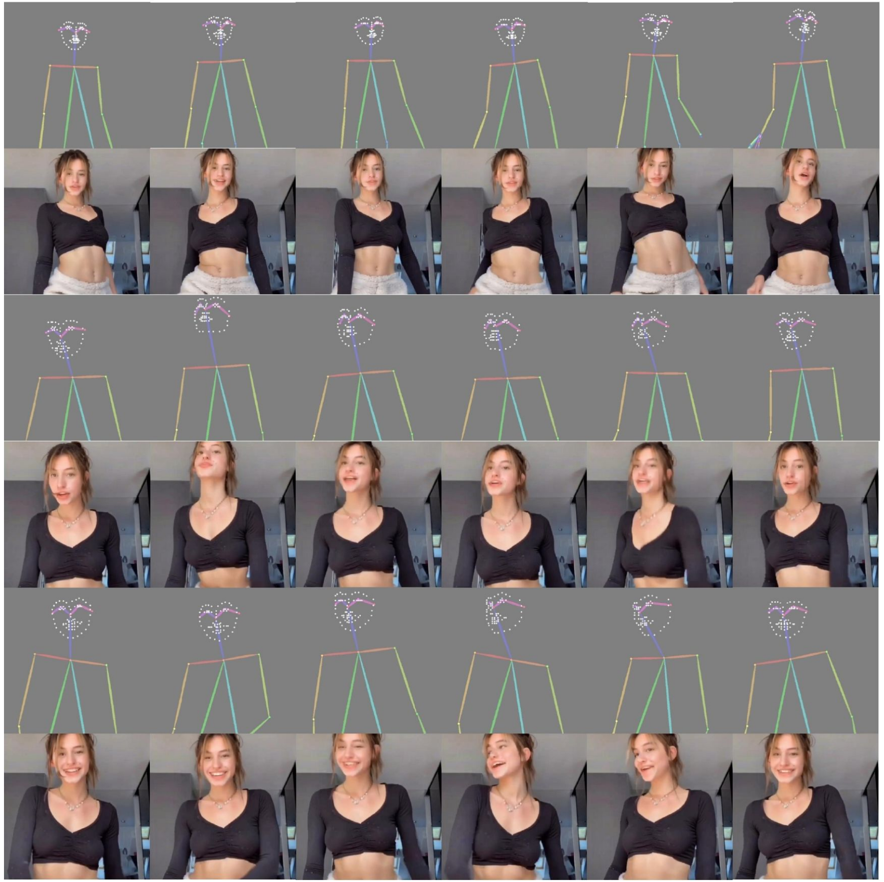
Experiment - TikTok

GT Pose TPS Disco MagicPose

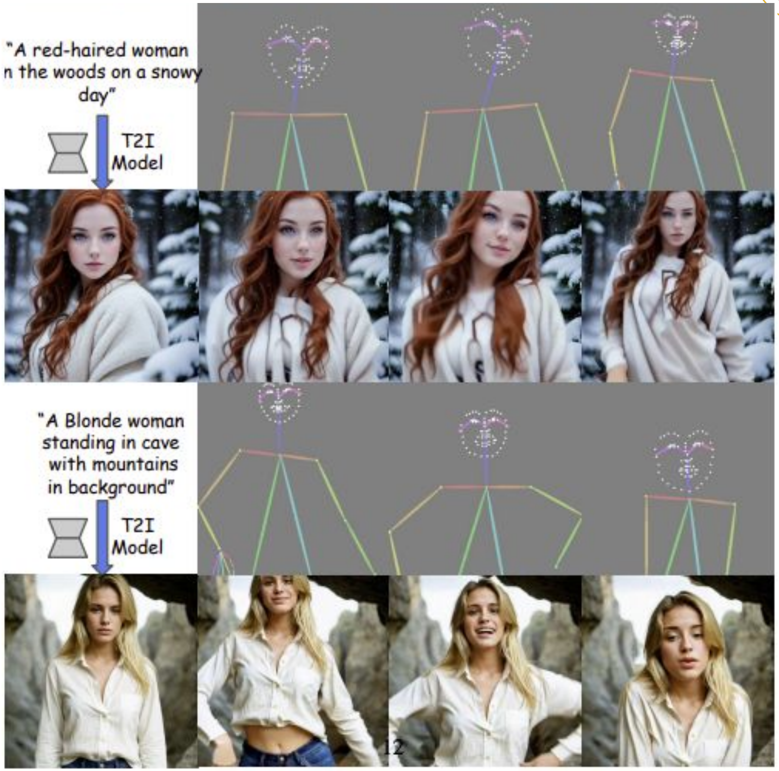
Experiment - TikTok



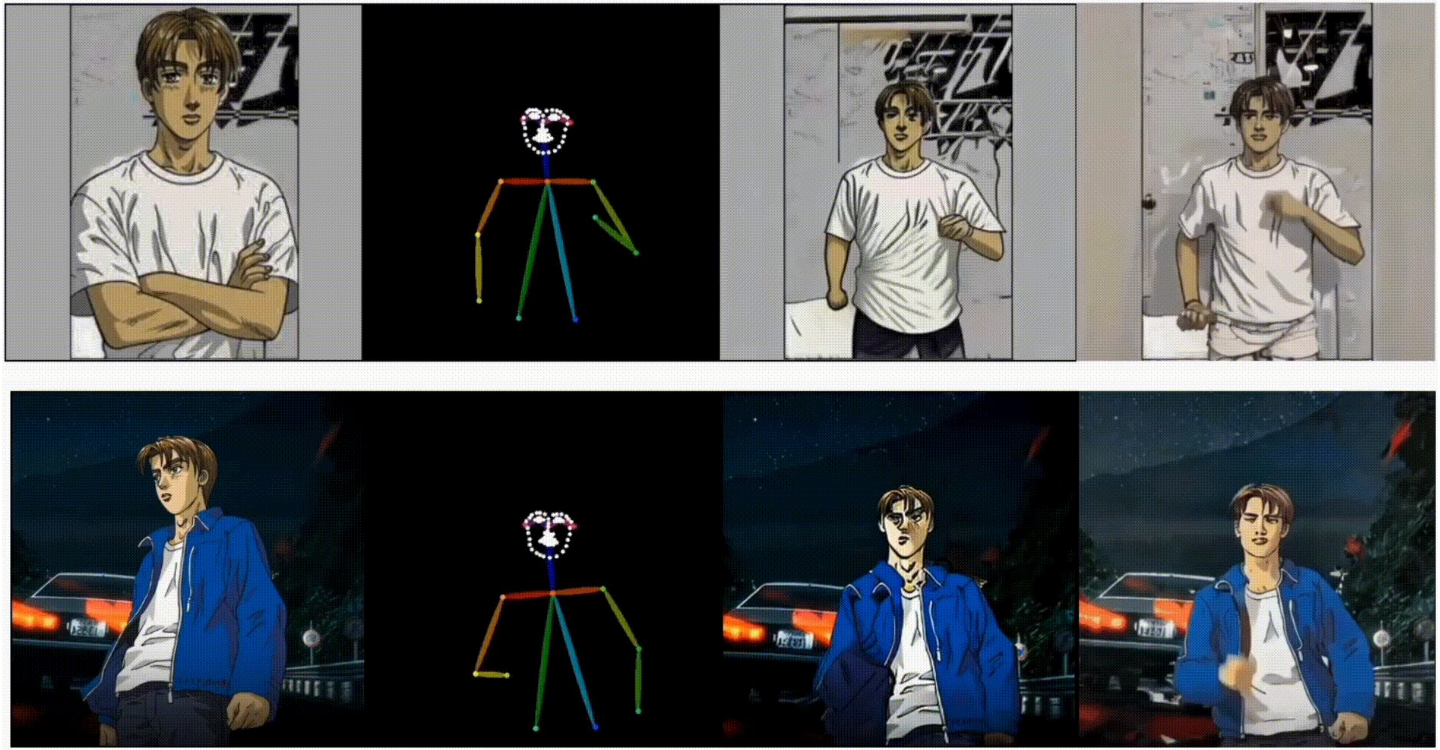




Experiment - Out of Domain Appearance

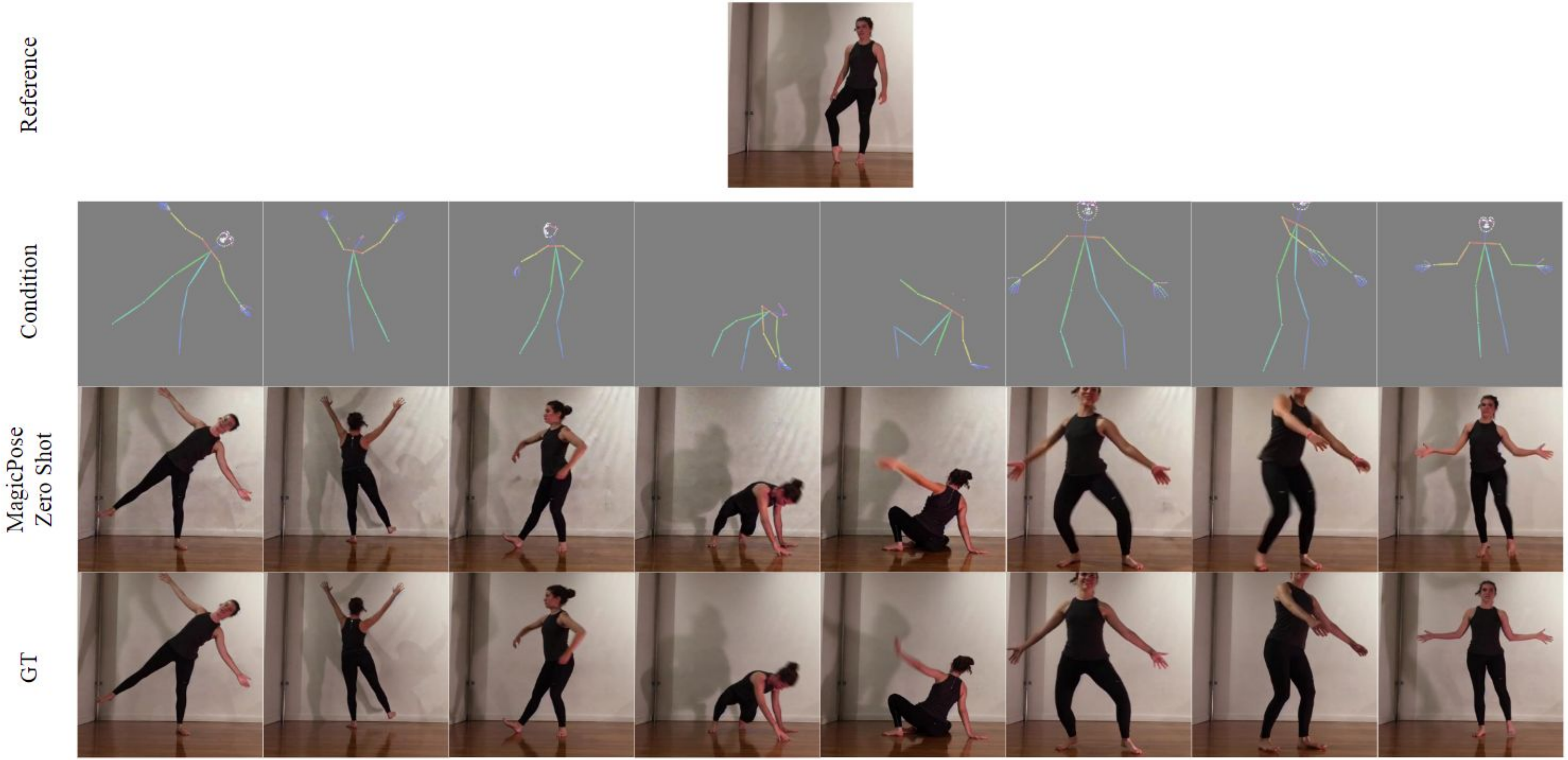


Comparison - Out of Domain Appearance



Reference | Pose | Animate Anyone | MagicPose

Experiment - Out of Domain Motions



Experiment - Quantitative Result

Method	Image						Video
	FID ↓	SSIM ↑	PSNR ↑	LPIPS ↓	L1 ↓	Face-Cos ↑	FID-VID ↓
FOMM* (Siarohin et al., 2019a)	85.03	0.648	29.01	0.335	3.61E-04	0.190	90.09
MRAA* (Siarohin et al., 2021)	54.47	0.672	29.39	0.296	3.21E-04	0.337	66.36
TPS* (Zhao & Zhang, 2022)	53.78	0.673	29.18	0.299	3.23E-04	0.280	72.55
DisCo (Wang et al., 2023)	50.68	0.648	28.81	0.309	4.27E-04	-	69.68
DisCo† (Wang et al., 2023)	30.75	0.668	29.03	**0.292**	3.78E-04	0.166	59.90
MagicPose	**25.50**	**0.752**	**29.53**	**0.292**	**0.81E-04**	**0.426**	**46.30**

Note*: **Face-Cos** is a novel metric which represents the cosine similarity of the extracted feature by AdaFace (Kim et al., 2022) of face area between generation and ground truth image.

User Study: Survey

- Participants: We use **Prolific**, an online platform designed to connect researchers with participants for academic studies and market research for all our user studies. The participants are English-speaking lay persons verified by this platform **without** any specific expertise in computer vision. We recruited **100** participants for this user study, and the hourly rate for this study is 72/hr. (~3 min per study)

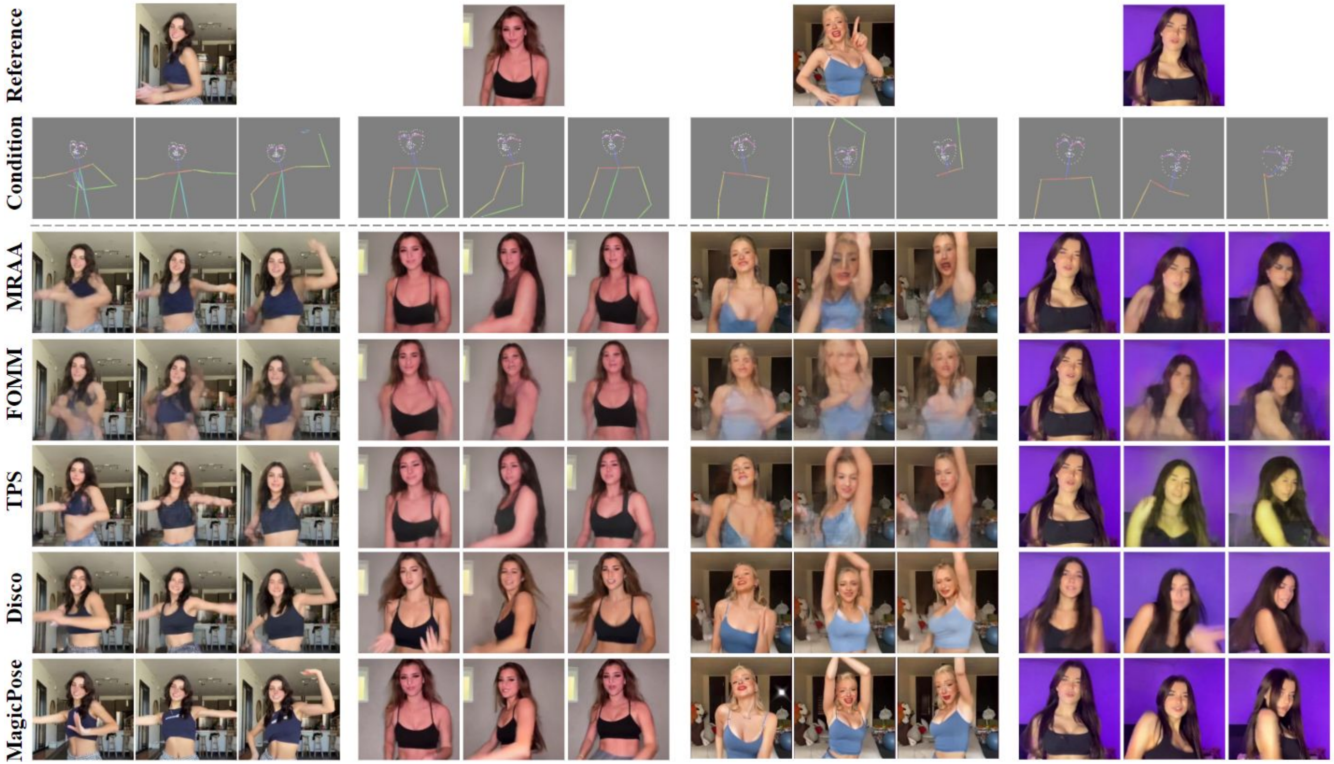
- Data: We visualize 8 video sequences from the TikTok dataset and compare the performance of MagicPose to prior works. Examples are shown in the next slides.

- Procedure:
    a. For each of the 8 video sequences, we visualize different human poses and facial expressions.
    b. The methods are anonymized as A, B, C, D, E, and the order of the generated image from the corresponding method is randomized.
    c. We ask the Participants to choose **only one** best method.

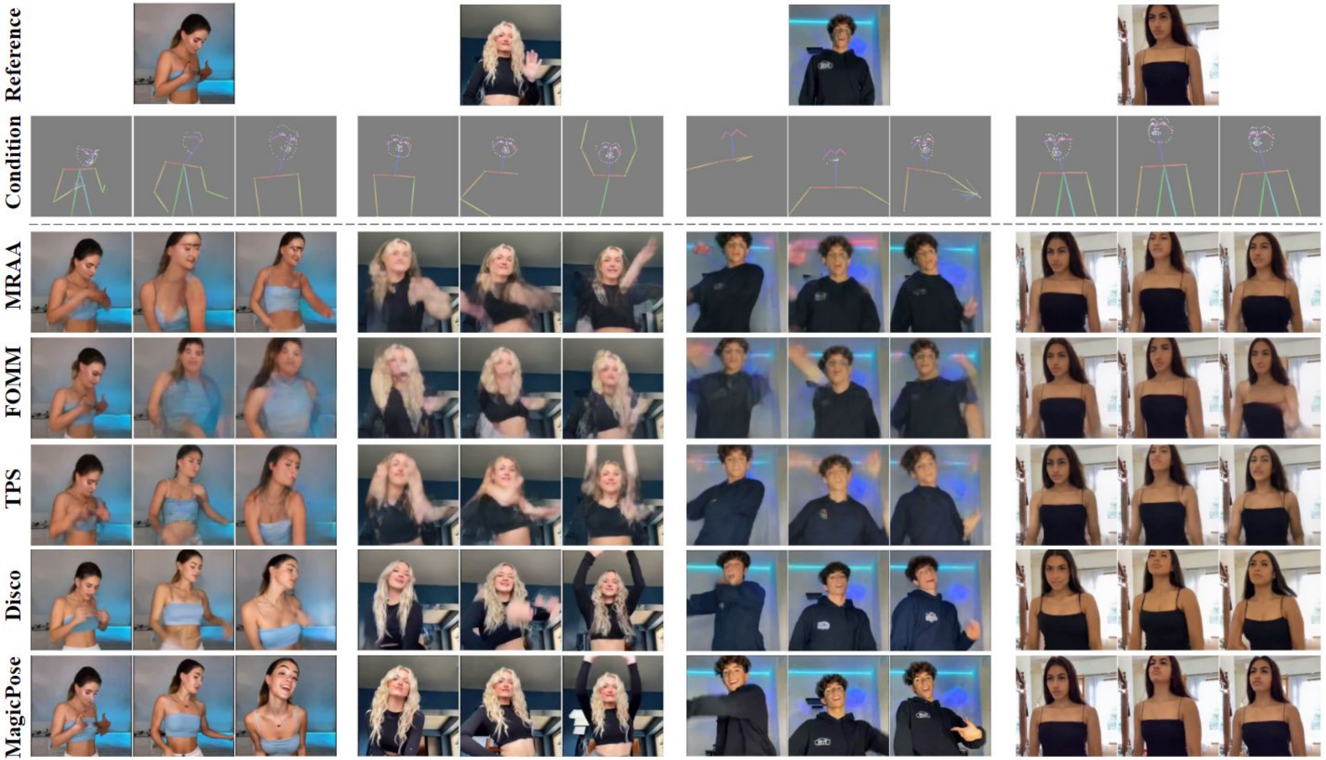
- Tutorial & Example: See here.

- Criteria for Judgment:
    a. 1) The appearance (Face, Clothes on the body, Background) of the generation should strictly match the given reference image input.
    b. 2) The motions and facial expressions of the generation should strictly match the given pose condition map input.

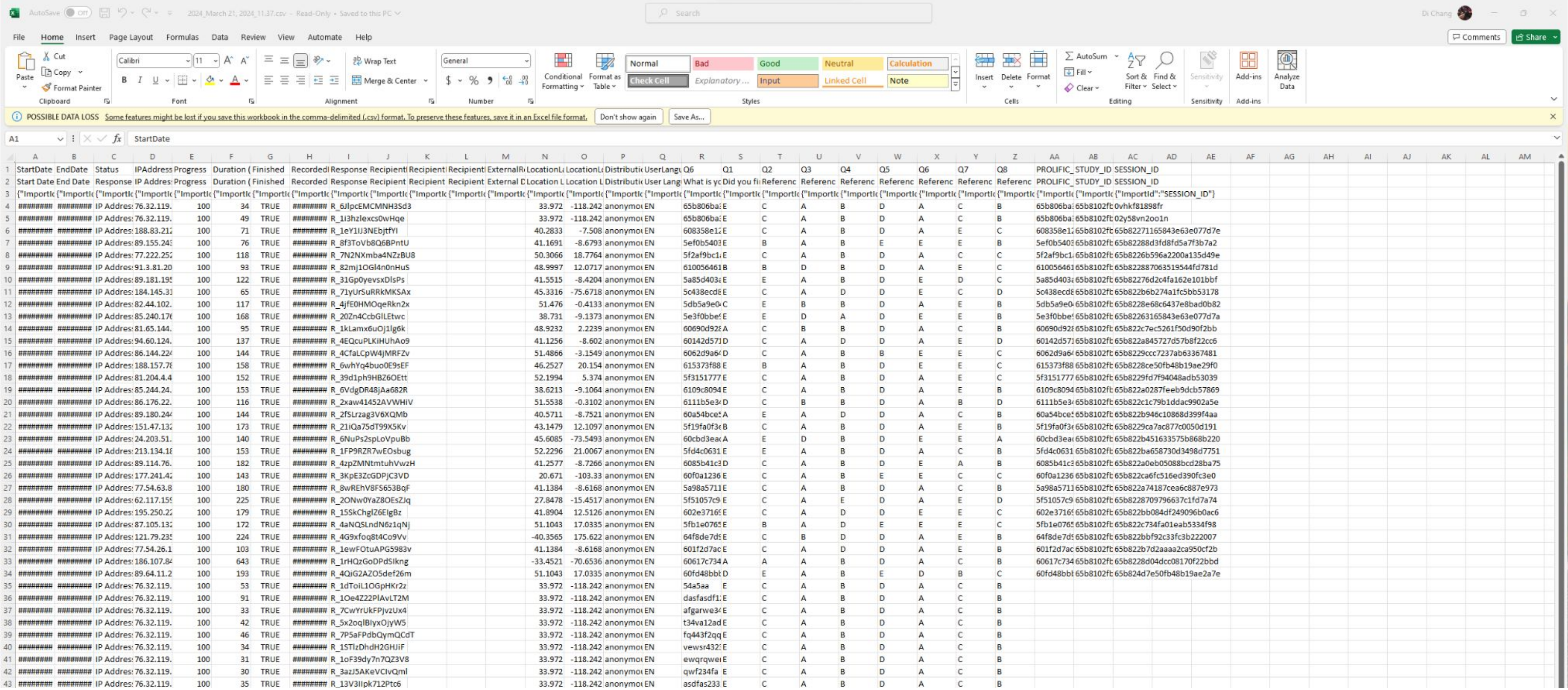
User Study



User Study



User Study - Initial Votes from Users



#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	StartDate	EndDate	Status	IPAddress	Progress	Duration (	Finished	RecordedI	Response	RecipientI	RecipientI	RecipientI	ExternalR	LocationLa	LocationLo	Distributio	UserLangu
2	Start Date	End Date	Response	IP Addres	Progress	Duration (	Finished	Recorded	Response	Recipient	Recipient	Recipient	External D	Location L	Location L	Distributio	User Lang
3	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI
4	########	########	IP Addres	76.32.119.	100	34	TRUE	########	R_6JlpcEMCMNH3Sd3					33.972	-118.242	anonymou	EN
5	########	########	IP Addres	76.32.119.	100	49	TRUE	########	R_1I3hzlexcs0wHqe					33.972	-118.242	anonymou	EN
6	########	########	IP Addres	188.83.212	100	71	TRUE	########	R_1eY1IJ3NEbjtfYI					40.2833	-7.508	anonymou	EN
7	########	########	IP Addres	89.155.243	100	76	TRUE	########	R_8f3ToVb8Q6BPntU					41.1691	-8.6793	anonymou	EN
8	########	########	IP Addres	77.222.252	100	118	TRUE	########	R_7N2NXmba4NZzBU8					50.3066	18.7764	anonymou	EN
9	########	########	IP Addres	91.3.81.20	100	93	TRUE	########	R_82mj1OGl4n0nHuS					48.9997	12.0717	anonymou	EN
10	########	########	IP Addres	89.181.195	100	122	TRUE	########	R_31Gp0yevsxDlsPs					41.5515	-8.4204	anonymou	EN
11	########	########	IP Addres	184.145.31	100	65	TRUE	########	R_71yUrSuRRkMKSAx					45.3316	-75.6718	anonymou	EN
12	########	########	IP Addres	82.44.102.	100	117	TRUE	########	R_4jfE0HMOqeRkn2x					51.476	-0.4133	anonymou	EN
13	########	########	IP Addres	85.240.176	100	168	TRUE	########	R_20Zn4CcbGlLEtwc					38.731	-9.1373	anonymou	EN
14	########	########	IP Addres	81.65.144.	100	95	TRUE	########	R_1kLamx6uOj1lg6k					48.9232	2.2239	anonymou	EN
15	########	########	IP Addres	94.60.124.	100	137	TRUE	########	R_4EQcuPLKIHUhAo9					41.1256	-8.602	anonymou	EN
16	########	########	IP Addres	86.144.224	100	144	TRUE	########	R_4CfaLCpW4jMRFZv					51.4866	-3.1549	anonymou	EN
17	########	########	IP Addres	188.157.78	100	158	TRUE	########	R_6whYq4buo0E9sEF					46.2527	20.154	anonymou	EN
18	########	########	IP Addres	81.204.4.4	100	152	TRUE	########	R_39d1ph9HBZ6OEtt					52.1994	5.374	anonymou	EN
19	########	########	IP Addres	85.244.24.	100	153	TRUE	########	R_6VdgDR48jAa682R					38.6213	-9.1064	anonymou	EN
20	########	########	IP Addres	86.176.22.	100	116	TRUE	########	R_2xaw41452AVWHiV					51.5538	-0.3102	anonymou	EN
21	########	########	IP Addres	89.180.244	100	144	TRUE	########	R_2fSLrzag3V6XQMb					40.5711	-8.7521	anonymou	EN
22	########	########	IP Addres	151.47.132	100	173	TRUE	########	R_21iQa75dT99X5Kv					43.1479	12.1097	anonymou	EN
23	########	########	IP Addres	24.203.51.	100	140	TRUE	########	R_6NuPs2spLoVpuBb					45.6085	-73.5493	anonymou	EN
24	########	########	IP Addres	213.134.18	100	153	TRUE	########	R_1FP9RZR7wEOsbug					52.2296	21.0067	anonymou	EN
25	########	########	IP Addres	89.114.76.	100	182	TRUE	########	R_4zpZMNtmtuhVwzH					41.2577	-8.7266	anonymou	EN
26	########	########	IP Addres	177.241.42	100	143	TRUE	########	R_3KpE3ZcGDPjC3VD					20.671	-103.33	anonymou	EN
27	########	########	IP Addres	77.54.63.8	100	180	TRUE	########	R_8wREhV8FS653BqF					41.1384	-8.6168	anonymou	EN
28	########	########	IP Addres	62.117.159	100	225	TRUE	########	R_2ONw0YaZ8OEsZJq					27.8478	-15.4517	anonymou	EN
29	########	########	IP Addres	195.250.22	100	179	TRUE	########	R_15SkChglZ6ElgBz					41.8904	12.5126	anonymou	EN
30	########	########	IP Addres	87.105.132	100	172	TRUE	########	R_4aNQSLndN6z1qNj					51.1043	17.0335	anonymou	EN
31	########	########	IP Addres	121.79.235	100	224	TRUE	########	R_4G9xfoq8t4Co9Vv					-40.3565	175.622	anonymou	EN
32	########	########	IP Addres	77.54.26.1	100	103	TRUE	########	R_1ewFOtuAPG5983v					41.1384	-8.6168	anonymou	EN
33	########	########	IP Addres	186.107.84	100	643	TRUE	########	R_1rHQzGoDPdSlkng					-33.4521	-70.6536	anonymou	EN
34	########	########	IP Addres	89.64.11.2	100	193	TRUE	########	R_4QiG2AZO5def26m					51.1043	17.0335	anonymou	EN
35	########	########	IP Addres	76.32.119.	100	53	TRUE	########	R_1dToiL1OGpHKr2z					33.972	-118.242	anonymou	EN
36	########	########	IP Addres	76.32.119.	100	91	TRUE	########	R_1Oe4Z22PlAvLT2M					33.972	-118.242	anonymou	EN
37	########	########	IP Addres	76.32.119.	100	33	TRUE	########	R_7CwYrUkFPJvzUx4					33.972	-118.242	anonymou	EN
38	########	########	IP Addres	76.32.119.	100	42	TRUE	########	R_5x2oqlBlyxOjyW5					33.972	-118.242	anonymou	EN
39	########	########	IP Addres	76.32.119.	100	46	TRUE	########	R_7P5aFPdbQymQCdT					33.972	-118.242	anonymou	EN
40	########	########	IP Addres	76.32.119.	100	34	TRUE	########	R_1STlzDhdH2GHJiF					33.972	-118.242	anonymou	EN
41	########	########	IP Addres	76.32.119.	100	31	TRUE	########	R_1oF39dy7n7QZ3V8					33.972	-118.242	anonymou	EN
42	########	########	IP Addres	76.32.119.	100	30	TRUE	########	R_3azJ5AKeVClvQml					33.972	-118.242	anonymou	EN
43	########	########	IP Addres	76.32.119.	100	35	TRUE	########	R_13V3Ilpk712Ptc6					33.972	-118.242	anonymou	EN

#	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	Q6	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	PROLIFIC_	STUDY_ID	SESSION_ID
2	What is yo	Did you fi	Referenc	Referenc	Referenc	Referenc	Referenc	Referenc	Referenc	PROLIFIC_	STUDY_ID	SESSION_ID
3	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportI	{"ImportId":"SESSION_ID"}
4	65b806ba	E	C	A	B	D	A	C	B	65b806ba	65b8102fb	0vhkf81898fr
5	65b806ba	E	C	A	B	D	A	C	B	65b806ba	65b8102fb	02y58vn2oo1n
6	608358e12	E	C	A	B	D	A	E	C	608358e12	65b8102fb	65b82271165843e63e077d7e
7	5ef0b5403	E	B	A	B	E	E	E	B	5ef0b5403	65b8102fb	65b82288d3fd8fd5a7f3b7a2
8	5f2af9bc1	E	C	A	B	D	A	C	C	5f2af9bc1	65b8102fb	65b8226b596a2200a135d49e
9	610056461	B	B	D	B	D	A	E	C	610056461	65b8102fb	65b822887063519544fd781d
10	5a85d403	E	E	A	B	D	E	D	C	5a85d403	65b8102fb	65b82276d2c4fa162e101bbf
11	5c438ecd8	E	C	A	D	D	E	C	D	5c438ecd8	65b8102fb	65b822b6b274a1fc5bb53178
12	5db5a9e0	C	E	B	B	D	A	E	B	5db5a9e0	65b8102fb	65b8228e68c6437e8bad0b82
13	5e3f0bbe	E	E	D	A	D	E	E	B	5e3f0bbe	65b8102fb	65b82263165843e63e077d7a
14	60690d928	A	C	B	B	D	A	C	B	60690d928	65b8102fb	65b822c7ec5261f50d90f2bb
15	60142d571	D	C	A	D	D	A	E	D	60142d571	65b8102fb	65b822a845727d57b8f22cc6
16	6062d9a64	D	C	A	B	B	E	E	C	6062d9a64	65b8102fb	65b8229ccc7237ab63367481
17	615373f88	E	B	A	B	D	E	E	C	615373f88	65b8102fb	65b8228ce50fb48b19ae29f0
18	5f3151777	E	C	A	B	D	A	E	C	5f3151777	65b8102fb	65b8229fd7f94048adb53039
19	6109c8094	E	C	A	B	D	A	E	B	6109c8094	65b8102fb	65b822a0287feeb9dcb57869
20	6111b5e34	D	C	B	B	D	A	B	D	6111b5e34	65b8102fb	65b822c1c79b1ddac9902a5e
21	60a54bce5	A	E	A	D	D	A	C	B	60a54bce5	65b8102fb	65b822b946c10868d399f4aa
22	5f19fa0f3e	B	C	A	B	D	A	E	B	5f19fa0f3e	65b8102fb	65b8229ca7ac877c0050d191
23	60cbd3ea	A	E	D	B	D	E	E	A	60cbd3ea	65b8102fb	65b822b451633575b868b220
24	5fd4c0631	E	E	A	B	D	A	C	B	5fd4c0631	65b8102fb	65b822ba658730d3498d7751
25	6085b41c3	D	C	A	B	D	E	A	B	6085b41c3	65b8102fb	65b822a0eb05088bcd28ba75
26	60f0a1236	E	C	A	B	E	E	C	C	60f0a1236	65b8102fb	65b822ca6fc516ed390fc3e0
27	5a98a5711	E	C	A	B	D	A	C	B	5a98a5711	65b8102fb	65b822a74187cea6c887e973
28	5f51057c9	E	C	A	E	D	A	E	D	5f51057c9	65b8102fb	65b8228709796637c1fd7a74
29	602e3716	E	C	A	D	D	E	E	C	602e3716	65b8102fb	65b822bb084df249096b0ac6
30	5fb1e0765	E	B	A	D	E	E	E	C	5fb1e0765	65b8102fb	65b822c734fa01eab5334f98
31	64f8de7d	E	C	B	D	D	A	E	B	64f8de7d	65b8102fb	65b822bbf92c33fc3b222007
32	601f2d7ac	E	C	A	D	D	A	E	B	601f2d7ac	65b8102fb	65b822b7d2aaaa2ca950cf2b
33	60617c734	A	A	A	B	D	A	C	B	60617c734	65b8102fb	65b8228d04dcc08170f22bbd
34	60fd48bb	D	E	A	B	E	D	B	C	60fd48bb	65b8102fb	65b824d7e50fb48b19ae2a7e
35	54a5aa	E	C	A	B	D	A	C	B			
36	dasfasdf1	E	C	A	B	D	A	C	B			
37	afgarwe34	E	C	A	B	D	A	C	B			
38	t34va12ad	E	C	A	B	D	A	C	B			
39	fq443f2qq	E	C	A	B	D	A	C	B			
40	vewsr432	E	C	A	B	D	A	C	B			
41	ewqrqwe	E	C	A	B	D	A	C	B			
42	qwf234fa	E	C	A	B	D	A	C	B			
43	asdfas233	E	C	A	B	D	A	C	B			

User Studv - Table of Votes

Table 1. A large-scale user study with 100 participants. We collect the number of votes for eight video subjects from test set by five methods and report the percentage. Our MagicPose preserves the best identity information in pose and facial expression retargeting on all subjects.

Method	Subject1	Subject2	Subject3	Subject4	Subject5	Subject6	Subject7	Subject8	Average
MRAA (Siarohin et al., 2019)	8%	6%	0%	5%	2%	2%	8%	4%	4%
FOMO (Siarohin et al., 2021)	3%	1%	3%	1%	1%	0%	5%	8%	3%
TPS (Zhao & Zhang, 2022)	4%	16%	0%	4%	2%	3%	4%	2%	4%
Disco (Wang et al., 2023)	12%	3%	9%	18%	5%	20%	33%	27%	16%
MagicPose	**73%**	**74%**	**88%**	**72%**	**90%**	**75%**	**50%**	**59%**	**73%**

Statistical Analysis - Chi Square Test

1. State the Null Hypothesis: There is no association between the video subjects and the choice of method. The distribution of votes for each method is the same across all groups, meaning any observed differences are due to chance.
2. Chi-square statistic: 116.02. p-value: 1.11 × 10−12. Degrees of freedom: 28.
3. Conclusion and Discussion: Given the extremely small p-value (much less than 0.05), we can reject the Null Hypothesis. The differences in vote distribution are unlikely to have occurred by chance. The participants indeed prefer MagicPose more than other methods.

Outline

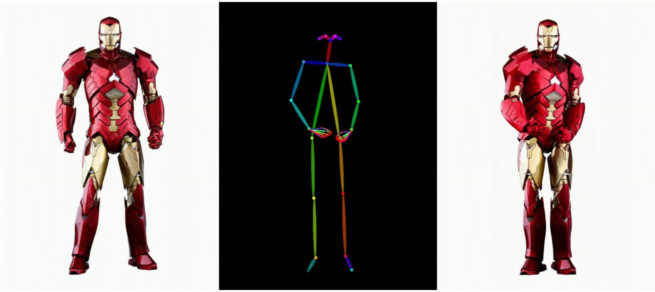
- Introduction & Problem Definition
- Recap Diffusion Model & ControlNet
- Motivation
- Method - Pipeline
- Results & User Study
- **Conclusion & Future Work**

Conclusion

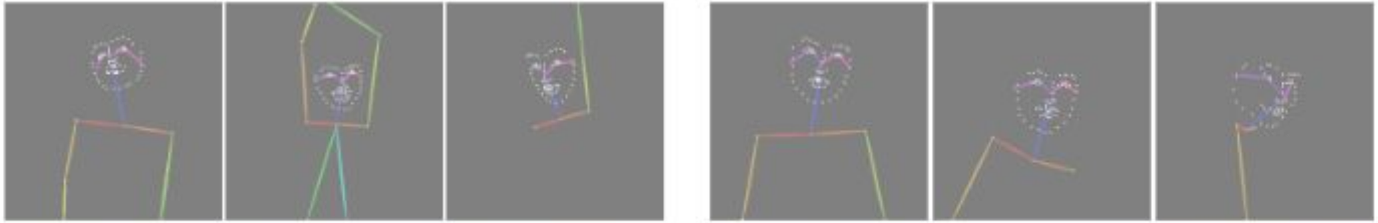
- We propose an effective method - **MagicPose**, for human pose and expression retargeting.
- The proposed model can be used as a **plug-in** extension for Stable Diffusion.
- We introduce **Multi-Source Attention Module** that offers detailed appearance guidance.
- Experiment on out-of-domain data demonstrating strong **generalizability** of our model to diverse image styles and human motions.

Future Work

- Improve temporal consistency for video quality.



- Adopt more advanced pose detector for better motion representation. Current openpose detector is not stable and fails to detect complete skeleton.



Social Impact

- Positive - Improving communication in digital or virtual environments
- Positive - Enhancing interactions in virtual meetings, online classrooms, and social networking platform
- Positive - Entertainment and media production, allowing for the creation of more lifelike and expressive characters in movies, video games, and animations

- Negative - Making fake animated videos of people which could be used in frauds
- Potential Solution - Digital watermarking and detection algorithms, enact and enforce stringent legal measures. Enhance public awareness and education on media literacy. Establish ethical guidelines within the tech industry.

Reference

[1] Dhariwal, Prafulla, et al. "Diffusion Models Beat GANs on Image Synthesis." *Arxiv Report*. 2021
[2] Rombach, Robin, et al. "High-Resolution Image Synthesis with Latent Diffusion Models." *Proceedings of the IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR).* 2022
[3] Zhang, Lvmin, et al. "Adding Conditional Control to Text-to-Image Diffusion Models." *Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision*. 2023
[4] Wang, Tan, et al. "DisCo: Disentangled Control for Realistic Human Dance Generation." *Proceedings of the IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR).* 2024

Thanks for listening!

Please scan the following QR-Code to check our project and more demos :))

Website:

Code: