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

Supplementary Material

1. Video Results

We present the video version of all the data and results we show in the paper, along with additional results, to demonstrate the generalizability of our model. Please refer to the video folder for the results. You can also click on `video_results.html` link to open it with your favorite browser (loading faster in Chrome than Safari!) to see everything all at once. Specifically, we present results of the following kinds:

- Single-person insertion results.
- Two-person insertion results.
- Multi-prompt interaction results.
- Comparison with image-to-video baselines.

We hope those real video results can showcase the quality of our generative model. Note that we tried to not do aggressive cherry picking on those results. All of the shown videos are generated in one pass without tweaking the random seed, and picked out of around one hundred validation samples to cover a diverse range of interesting behavior.

2. Data Processing Details

2.1. Data Filtering

We get the raw human-related dataset following the practice of video personalization in [5]. Specifically, we first get human videos by selecting videos with human-related concepts in their captions. We extract frames at one-second intervals and apply a face detector to keep videos that contain a single face and where the ArcFace cosine similarity score [2] between consecutive frames exceeds 0.5. This processing provides us with around one million text-video pairs where a single person appears, with duration from 4s to 16s. We additionally apply OpenPose [1] to only keep those with at least knee joints in the frame to avoid extreme close-ups. At the top of Fig. 1 we show some cases that we discard during the filtering process.

Note that interestingly, as we apply all the detection on middle frame, some earlier and later frames might not satisfy our requirements of full bodies. We choose to not specifically tackle these edge cases as they tend to have rich interactive contents with large-scale motions.

2.2. Human Removal

To process the data, we take the first and last frames of a video for human removal to get the scene image.

Human segmentation. We apply GroundingDINO [4] with the keyword `human` to get bounding boxes for each human in the image. We apply SAM 2.1 with the bounding box as guidance to segment out the binary human mask.

Inpainting. We apply the SDXL diffusion inpainting model. To avoid fuzzy segmentation boundary, we use OpenCV to dilate each binary mask by 50 pixels so that it's guaranteed to cover the whole human area. The positive prompt we use is "natural, photorealistic, empty, environment, blank, background, bg", and the negative prompt is "person, human, text". For two people videos, we separate the two person masks, and does inpainting with each mask separately. At the bottom of Fig. 1 we show a few additional data samples, including mask and detected poses.

2.3. Prompt Post-processing.

We split the prompt by sentences. For each sentence, we ask the LLaMA model [3] whether it describes the person or the background. If it's defined as a background prompt, we remove it from the caption. We additionally remove all sentences with the concept of camera in it, as we are not explicitly modeling any human-camera interaction.

3. Implementation Details

3.1. Base Model

We explain some training details of our base model below. Refer to [5] for more illustration. Note that while the training scheme and datasets are the same, we use a much smaller counterpart than the publicly announced Movie Gen model due to resource limitation.

We perform generation in a learned latent space representation of the video. This latent code is of shape $T \times C \times H \times W$. To prepare inputs for the Transformer backbone, the video latent code is 'patchified' using a 3D convolutional layer and then flattened to yield a 1D sequence. The 3D convolutional layer uses a kernel size of $k_{\mathrm{t}} \times k_{\mathrm{h}} \times k_w$ with a stride equal to the kernel size and projects it into the same dimensions as needed by the Transformer backbone. Thus, the total number of tokens input to the Transformer backbone is $THW/(k_t k_h k_w)$. We use $k_t = 1$ and $k_h = k_w = 2$, i.e., we produce 2×2 spatial patches.

We use a factorized learnable positional embedding to enable arbitrary size, aspect ratio, and video length. Absolute embeddings of D dimensions can be denoted as a mapping $\phi(i) : [0, \text{maxLen}\,] \rightarrow \mathbb{R}^D$ where i denotes the absolute index of the patch. We convert the 'patchified' tokens into separate embeddings ϕ_h, ϕ_w and ϕ_t of spatial h, w, and temporal t coordinates. We define $H_{\max}, W_{\max}$, and $T_{\max}$ as the maximum sequence length for each dimension, which corresponds to the maximum spatial size and video length

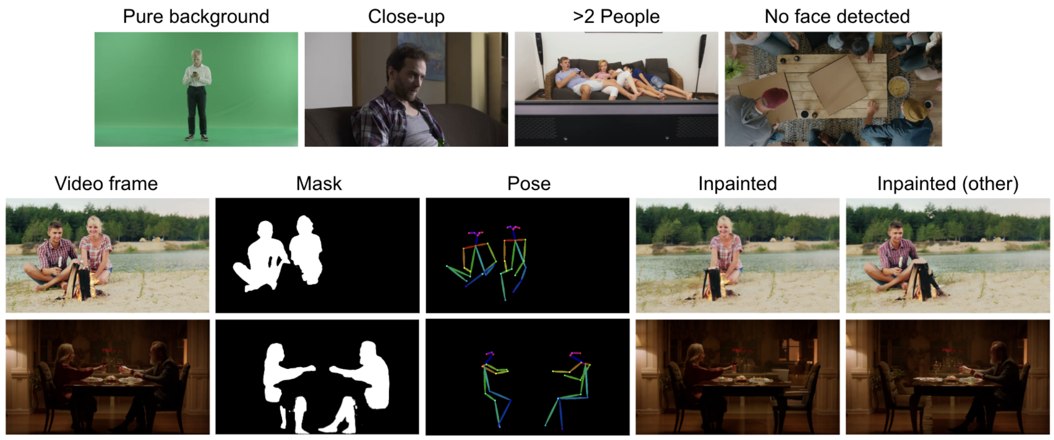

Figure 1. Additional illustration of our data processing pipeline. We include discarded data samples on top, and intermediate outputs of detection and filtering on bottom.

of the patchified inputs. We calculate the final positional embeddings by adding all the factorized positional embeddings together, and finally adding them to the input for all the Transformer layers.

3.2. Conditioning Branch

We build our cross attention conditioning branch by concatenating the text and image features. Specifically, we apply 2 layers of text enhancer self attention, 2 layers of image enhancer deformable attention, then 6 layers of cross-attention with image as key/value and 6 layers of cross-attention with text as key/value. We combine the enhanced image feature with the pre-trained text feature for cross-attention with Transformer layer outputs.

4. Evaluation Details

4.1. Baseline Details

T2I Inpainting. We deploy a pre-trained text-to-image inpainting model on the given scene frame. We use the ground truth human bounding boxes from GroundingDINO's prediction as a guidance mask for inpainting. Because the baseline's text encoder is not designed for long prompts, we only take the first two sentences in our caption as the positive inpainting prompt. In practice, they are able to describe the human action and appearance adequately. Note that this is not an exactly fair comparison, as we give the model a ground truth bounding box. We are able to show that, however, our model is able to generate more natural interaction even without a pre-defined position signal.

InstructPix2Pix and AnyV2V. Both of them are based on InstructPix2Pix, except that the second one is an extension into video after editing the first frame. We use LLaMa [3] to rewrite our prompts so that it falls into the instruction distribution. Instead of describing "the video shows a man", we rewrite the prompt as "adding a man". Similarly, due to the limit number of tokens the text encoder can take in, we only rewrite the first two sentences. We use the same prompt for both stages of AnyV2V.

Note that our baselines are mostly trained with squared images. Even though our model is exclusively trained with landscape videos, our Transformer architecture essentially enables generation of arbitrary aspect ratio. To accommodate the baselines, we use squared images for comparison in the main paper. We additionally provide some non-squared comparisons with the two image-based models in the next section.

4.2. Evaluation Metrics

FVD. FVD calculates the feature distance between two sets of videos. (the I3D features). We take the evaluation code and checkpoints from [6]. Specifically, the metric is computed by

$$\mathrm{FVD} = \|\mu_X - \mu_Y\|^2 + \mathrm{Tr}\left(\Sigma_X + \Sigma_Y - 2\left(\Sigma_X \Sigma_Y\right)^{1/2}\right)$$

where μ_X, μ_Y are the mean vectors and Σ_X, Σ_Y are the covariance vectors.

CLIP. We compute the CLIP similarity between generated visual contents and the text prompts. For videos, the distance is computed every one second, and averaged across the whole video.

Action Score. We design this metric to eliminate the influence of human appearance and solely evaluate whether the inserted human is doing the correct action. We ask LLaVA-Next [7] what the human is doing in a video, and provide samples of our action prompts as examples. We then compare the CLIP similarity between our prompt and the output. For the static images, we repeat the single static frame to make a video sequence. We notice that, as LLaVA is only

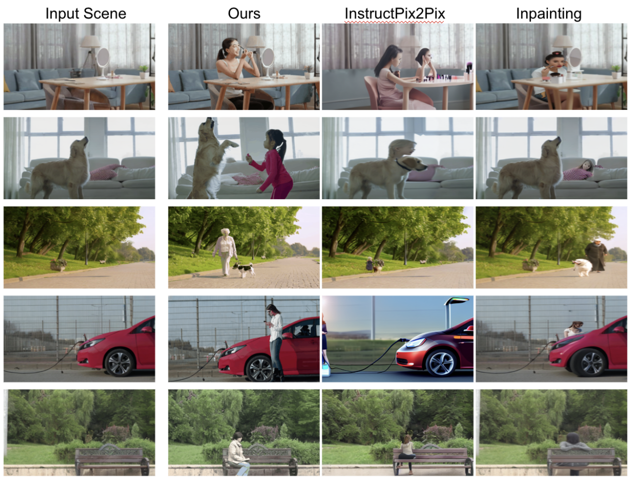

Figure 2. Additional comparison with baselines on non-square image inputs.

taking a few key frames to answer the question, repeating the static frames is a reasonable way to decide human actions in an image.

4.3. Human Evaluation Details

We run a user study to recruit thirty-seven people evaluating the results of our model. We randomly shuffle the results of ours versus the three baselines and the three types of ablations. Among the users, fourteen fill out the small questionnaire with 10 groups of randomly selected results, and twenty-three of them fill out the complete questionnaire with 80 groups. People are asked to select their preference of the results based on four dimensions as described in the main paper.

5. Additional Image Baseline Comparison

In Fig 2, we show additional frame-wise comparisons with the image-editing baselines to demonstrate our model's superior ability. Note from the results how our model is able to keep the scene consistent instead of generating something semantically similar, and also able to insert a human without a mask.

6. Ablation Visualizations

As shown in Fig. 3, our dual stream conditioning approach with both latent concatenation and feature enhanced cross-attention proves to be the best way of conditioning a T2V model on the scene image. Without latent concatenation, the model generates something semantically similar but not pixel-wise the same. Without fused cross-attention modules, the model is prone to generating distorted, unreasonable motions.

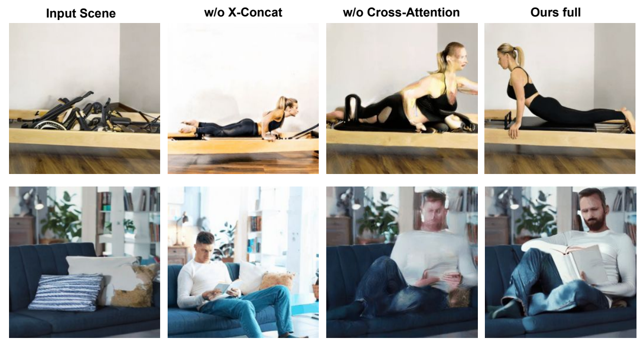

Figure 3. Comparison with alternative designs of our model.

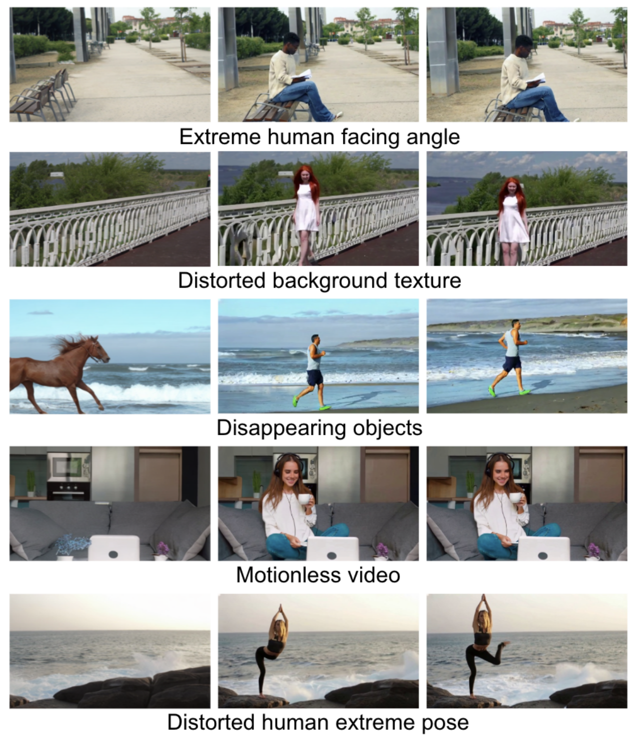

Figure 4. Limitation and failure cases of our model.

7. Limitations

We discuss a few key limitations and failure cases we noticed in our current method. Note that most of them are due to the base text-to-video model's limited capability, especially as we are basing our work on a smaller, lower resolution version. Overall, our method's quality greatly depends on the base model, and could be further improved with better model and more computing resources.

Videos with limited motions. Our model suffers from the common issue of generating videos with limited amount of motions (i.e. static videos). Specifically, we observe that some of our generated results have natural camera movements and environmental changes, while having the central character almost static. This is due to the data distribution which we use to train and fine-tune the model, and can likely be eliminated by providing higher quality fine-tuning dataset, or include motion guidance as an explicit condition to the model. Notably, we notice that our model is able to exhibit fair amount of motion with "action" prompts, like "running", "walking", "riking bike" whose underlying semantic requires great movements. And results are more static with "status" prompts like "sitting", "lying", which merely describes an existing state. Regardless of the amount of motion, our model is always able to insert the person into the correct place with reasonable interaction.

Human body distortion. Similar to other text-to-video models, our model is not perfect in generating human movements, especially in examples with extreme human motion like doing sports. Specifically, we observe artifacts in limbs and hands when the model expects to generate fine-grained, large-scaled movements. We consider this a common issue of current text-to-video model, and could be improved by using better base model.

Background texture distortion. We notice that our model fails to keep scene consistent if there is complex geometry or texture in the input image. For example, architectures with repetitive structures, or periodic textures with fine details. This is also an on-going issue of state-of-the-art text-to-video models awaiting solution.

Inpainting artifacts and object disappearing. Our human removal inpainting algorithms fail on a few edge cases, where it removes the human but replaces it with an additional object. Training on these data teaches the model to sometimes "remove" existing objects in a scene and replacing it by a person, even if it shouldn't disappear in first place. We believe this is a relatively minor data quality issue and could be mitigated by using better inpainting off-the-shelf method, or add an additional round of data filtering.

Extreme human facing angles. We model is not able to generate back-facing human. This is due to how we filter the data: we detect faces and only keep those with the same face across the whole video, which in nature eliminates back facing videos. In cases where the inserted human is expected to face an extreme angle such that most of the faces are unseen from the camera, our model tends to insert person in a wrong direction.

8. Reproducibility and Benchmark Release

While we are not able to release codebase or dataset due to copyright restrictions, we believe that with detailed descriptions of the base models in [5] and the extensive explanation of implementation details in this paper could provide the audience with a clear idea of our model's architecture and training. Moreover, as stated earlier, our goal is not to train the best model, but to explore how pretrained T2V models can perceive affordance from visual signals. We believe that our conditioning mechanism and cross-attention analysis can be applied to any such open-sourced models as well. We demonstrate results as a proof-of-concept, and hopefully would inspire more explorations in this field. We will release upon acceptance the benchmark dataset that we collected for evaluation to allow fair comparison for follow-up works.

References

[1] 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. 1

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[3] 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 Bhosale, 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,

ICCV #12877

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. 1, 2

[4] 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. 1

[5] 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, 4

[6] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. In *(NeurIPS) Advances in Neural Information Processing Systems*, 2022. 2

[7] 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. 2