LVM-Lite: Training Large Vision Models with Efficient Sequential Modeling

1. Appendix / supplemental material

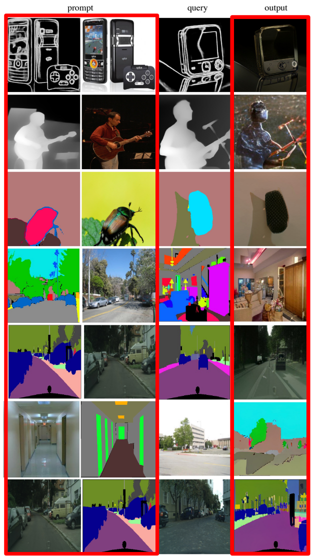

Figure 1. Example of prompts in our metric-evaluation. We use a single prompt containing one image pair to indicate a task.

1.1. Implementation details

Model configuration. In our experiments, we systematically explore four models whose configurations are listed in Table 1. These models are based on a decoder-only architecture, specifically leveraging the Llama-2 framework [29], chosen for its efficiency and adaptability to our framework. Due to limited computation resources, our largest 3B model adopted an advanced block-parallel transformer[17] to reduce memory requirements further. All of our experiments are conducted on a 256-core TPU-v3. Our implementation is based on JAX[3]

Table 1. Model architecture

model size	hidden dim	MLP dim	heads	layers
300M	1024	2688	8	22
600M	1536	4096	16	22
1B	2048	5504	16	22
3B	3200	8640	32	26

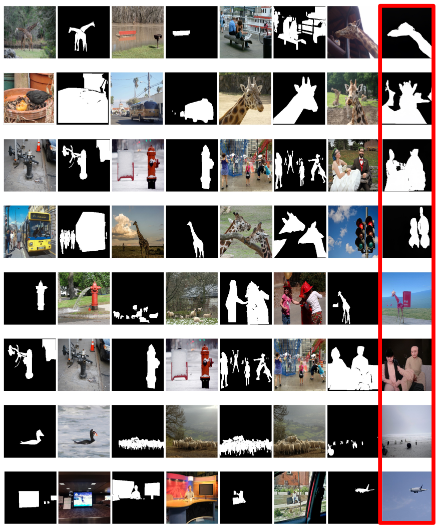

Figure 2. COCO[16] evaluation.Red: the generated results. First four rows: segmentation task. Second four rows: segmentation to images. For all examples, we use three prompts that contain 6 images in total to indicate the task and one query mask/image.

Table 2. Hyperparameters for pre-training and fine-tuning.

hyperparameter	single-image pre-training	sequence fine-tuning
learning rate schedule	linear warmup and cosine decay	
weight decay	0.1	
optimizer	AdamW[18]	
optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.95$	
base learning rate	1.5e-4	1.5e-5
final learning rate	1.5e-5	1.5e-6
warmup steps	2000	0
total training steps	125112	15639
batch size	8192	512
context length	256	4096

1.2. Training and evaluation.

We also provide detailed pre-training and fine-tuning hyperparameters in Table 2. We use training hyperparameters based on [2]. To enhance efficiency, we ensure that the total number of processed tokens per iteration remains constant, increasing the pre-training batch size by $\times 16$. For our evaluation, we utilize prompts to specify tasks in line with [2]. Instead of employing seven pairs of images, we discovered that a single pair is adequate for task indication. These

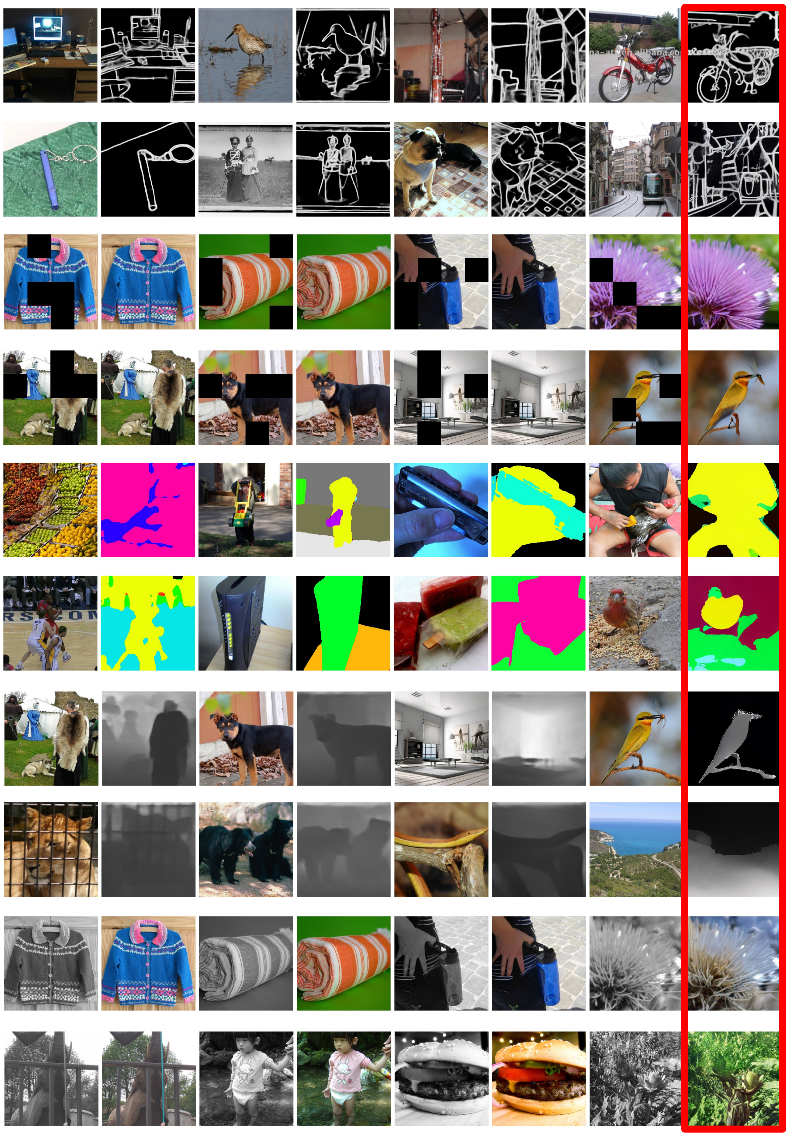

Figure 3. ImageNet-1K[9] qualitative evaluation on validation set. Red: the generated results. For all examples, we use three prompts that contain 6 images in total to indicate the task and one query. We show edge detection, inpainting, semantic segmentation, relative depth estimation, and colorization tasks.

prompts are illustrated in Figure 1. Our approach allows us to assess qualitative and quantitative results across different tasks. We set the Top-K as 100 and the temperature as 1. For the video prediction task. we sample 16 frames from the original video to be the ground truth and use the first 5 frames as prompt. We ask the model to generate the rest 11 frames. For ImageNet-1K[9] generation, we evaluate our model on the validation set. We randomly sample one of the training set prompts and use the same prompt for all validation images. We use [28] to infer the edges map, mask2former[6] to generate semantic masks, and depth-anything[32] to predict depth map. For ADE20K[34] and Cityscapes[7] generation task, we use ground truth segmentation map to be the condition. For each task, we first adjust the image size to 256 × 256 using bilinear interpolation and apply the nearest neighbor interpolation method to resize the masks.

1.3. Datasets

Here, we present in detail how we construct our datasets. For training, we mainly follow LVM[2] to construct our datasets; we pre-process most datasets listed in [2] in the same manner. We list all datasets used in our experiments in Table 3. We always keep 16 images for different tasks as the length of image sequences. Thus, for video generation task, we sample 16 frames from the original video. For image-

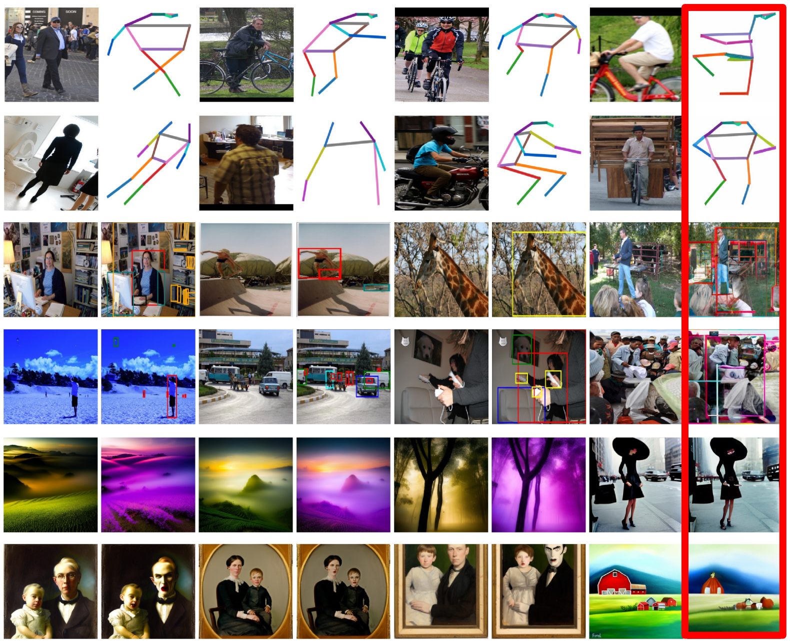

Figure 4. Other tasks' qualitative evaluation set. Red: the generated results. For all examples, we use three prompts that contain 6 images in total to indicate the task and one query. We show human pose estimation, object detection and style transfer tasks.

based tasks, we use 8 image pairs to form a single image sequence. For evaluation, we show the details of datasets we used in Table 4.

1.4. Additional Results

Our additional quality evaluation spans tasks on COCO[16], ImageNet-1K[9], and VIPSeg[20] datasets shown in Figure 2,3,4 and 11. We demonstrate our model's capability on COCO in human pose estimation and object detection. We utilize [4] for style transfer to showcase our approach's adaptability. On ImageNet-1K, we cover edge detection, inpainting, semantic segmentation, relative depth estimation, and colorization, illustrating the model's versatility across different image processing challenges. Additionally, VIPSeg's qualitative results are included, where the task involves generating frames from 8 object masks, highlighting our model's proficiency in image synthesis.

Long-video Generation. We enhance LVM-Lite's capability to generate longer videos by fine-tuning it with the SS-V2 dataset to process 64 frames after 1500 iterations. We highlight the proficiency of LVM-Lite to generate high-quality, extended sequences sequences in Figure 12.

1.5. Reproducing Evaluation of LVM's Official Checkpoint

We conducted a thorough evaluation of the LVM official checkpoint [1] to compare its performance against our LVM-Lite model. The results highlight the limitations of LVM in discriminative tasks. For example, LVM-7B achieved only 1.9 mIoU on ADE20K [34] and 0.1 mIoU on Cityscapes [7]. In contrast, our LVM-Lite-300M model achieved 2.3 mIoU on ADE20K and 10.1 mIoU on Cityscapes, showcasing improved segmentation capabilities with a much smaller model.

Upon closer examination of LVM's results as shown in Figure 5, we observed that while the segmentation outputs appear visually similar to the ground truth due to close color matches, the generated classes are often incorrect. This suggests that while the segmentation process itself may be adequate, the model struggles significantly with object recognition, resulting in mismatches between the predicted and actual classes.

Second, the choice of tokenizer plays a critical role in determining the oracle's performance. A more effective tokenizer could potentially improve the model's ability to distinguish between classes and enhance segmentation quality.

Our work focuses on improving training efficiency and analyzing LVM's behavior. By decoupling the training process into pre-training and fine-tuning stages, we can study

[1] https://huggingface.co/Emma02/LVM_ckpts

Table 3. Full training dataset. We follow LVM[2] to construct datasets but divide them into generative and discriminative tasks.

dataset	task type	annotation source
random image sequence		
DataComp-1B[10]	inpainting	ground truth
natural sequence		
UCF101 [27]	video generation	ground truth
HMDB [14]		
Moments in Time [21]		
Multi-moments in Time [22]		
Co3D [24]		
Charades v1 [25]		
Something-something v2 [12]		
Kinetics 700 [5]		
Jester [19]		
MultiSports [15]		
CharadesEgo [26]		
AVA [23]		
Ego4D [13]		
Objaverse [8] Rendered Multiviews		
generative sequence		
ImageNet-1K [9]	image to image	ground truth
	segmentation map[6] to image	
	depth map [32] to image	
	edge map [28] to image	
	inpainting	
	colorization	
COCO [16]	instance segmentation to image	
ADE 20K [35], Cityscapes [7]	segmentation map to image	
Subset of InstructPix2Pix [11]	style transfer	
Charades V1 [26]	segmentation map[6] to video	
VIPSeg [20]	panoptic segmentation to video	
Co3D [24]	object mask to video	
Co3D [24]	depth to video	
discriminative sequence		
COCO [16]	object detection	-
ADE20K [35], Cityscapes [7]	semantic segmentation	ground truth
ImageNet-1K [9]	semantic segmentation	Mask2Former [6]
COCO [16]	human pose	ground truth
COCO [16], ImageNet-1K [9]	depth map image	Depth-anything [32]
COCO [16], ImageNet-1K [9]	edge detection	DexiNed [28]
SIDD [1]	denoised image	ground truth
LOL[30]	light-enhanced image	ground truth
VIPSeg [20]	video panoptic segmentation	ground truth
VOS [31]	video object segmentation	ground truth
Co3D [24]	video object segmentation	ground truth
Co3D [24]	video object segmentation	ground truth

LVM's capabilities in a cost-effective and meaningful manner. Our findings indicate that while LVM performs well in generative tasks and demonstrates scalability in video and conditional image generation, further development is required to achieve reasonable performance in discriminative tasks such as semantic segmentation.

1.6. Comparison on More Discriminative Tasks

Through previous analysis, we find the quantitative results are not comparable. We show a qualitative comparison with the official LVM, such as edge detection and depth estimation on ImageNet-1K (IN-1K); our LVM-Lite model demonstrates competitive performance as shown in Figure 6,7,9,8 and 10, especially considering its lightweight architecture. In LVM's instruction tuning setting, all labels on ImageNet-1K are generated, making quantitative analysis for this task challenging. Therefore, we compare the quality of the two models based on their default outputs.

1.7. Comparison on More Generative Tasks

For generative tasks, such as video prediction (UCF-101) and image synthesis (ADE20K-G and Cityscapes-G), our LVM-Lite model achieves competitive FID and FVD scores compared to LVM, especially with significantly reduced model size.

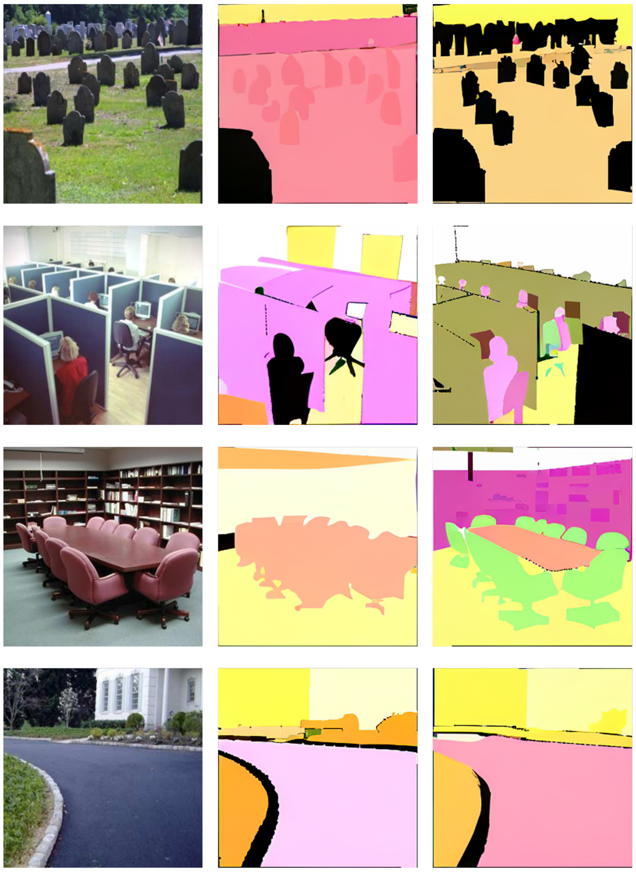

Figure 5. Evaluation of LVM [2] on ADE20K [34]. Left: query image. Middle: model prediction. Right: ground truth. Although the qualitative segmentation mask seems reasonable, compared with ground truth, class-mismatching and pixel shift problems resulting in the mIoU are pretty low.

Table 4. Evaluation datasets and metrics used for comparison for Table 4.

dataset	split&number of samples	metric
frame prediction		
UCF101 [27]	test & 3783	FVD&IS
Something-something v2 [12]	validation & 24777	FVD
Kinetics 600 [5]	validation & 31593	FVD
image synthesis		
ImageNet-1K [9]	validation & 50000	FID
ADE20K [35]	validation & 2000	FID
Cityscapes [7]	validation & 500	FID
semantic segmentation		
ADE20K [35]	validation & 2000	mIOU&FID
Cityscapes [7]	validation & 500	mIOU&FID

Video generation For video generation, we follow the common practice [33] to use the first five frames as a condition (prompt) to ask the model to generate the remaining 11 frames. The generation speed is LVM-7B is 3 mins per video on a single A5000 GPU, while for a 3B model, the speed can be 0.5 mins per video. We compare both quantitatively and qualitatively. Our 3B model achieves nearly twice the FID score of the official LVM-7B model, demonstrating superior generative quality.

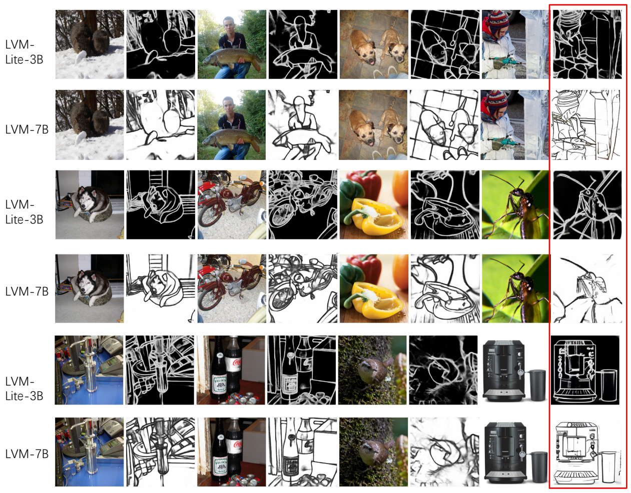

Figure 6. Comparison on edge detection. We noticed that our model uses different data pre-processing strategies to generate edge maps. Red rectangle: the generated results.

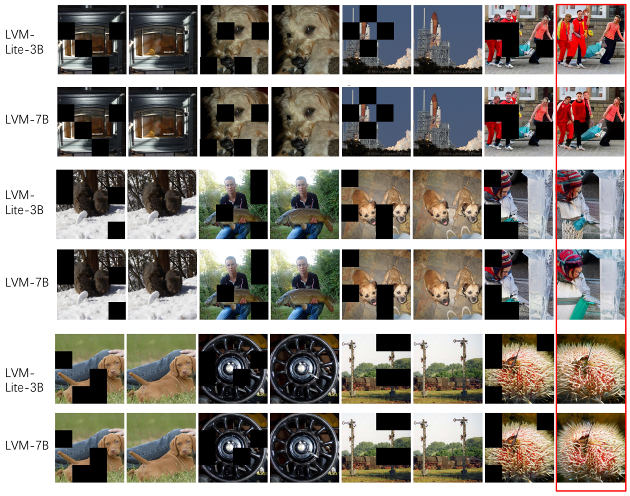

Figure 7. Comparison on inpainting task. Follow LVM [2] we use partially masked image-to-image Red rectangle: the generated results.

Upon closer inspection of the generated videos, as shown in Figure 13 and 14, we made a surprising observation: despite being smaller, our model is capable of generating fully consistent videos, whereas the LVM-7B model struggles with maintaining temporal consistency. Severe hallucination issues are evident in LVM-7B, with only the initial few frames adhering to the provided instructions, while subsequent frames deviate significantly.

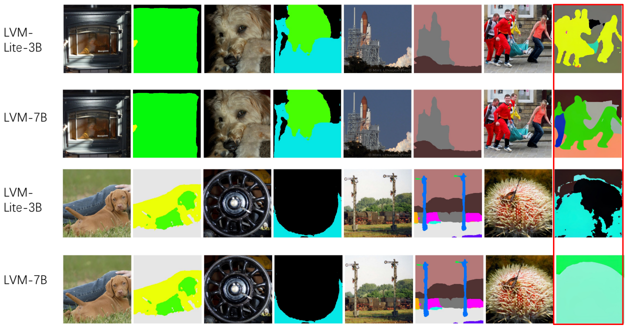

Figure 8. Comparison on segmentation task on IN-1K. Red rectangle: the generated results.

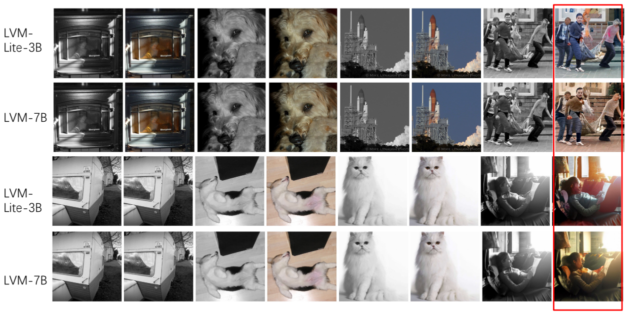

Figure 9. Comparison on colorization task on IN-1K. Red rectangle: the generated results.

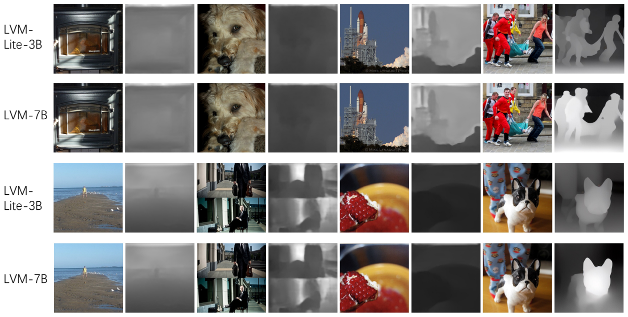

Figure 10. Comparison on depth estimation task on IN-1K. Red rectangle: the generated results.

Segmentation-mask to Image . Given that our image sequences include both segmentation masks and natural image masks and are inspired by the scalable generative capabilities of LVM-Lite, we evaluated the model's conditional image generation ability. This was done by providing a pair of prompts—a segmentation mask and its corresponding image—along with a query mask and asking the model to generate the corresponding image.

Quantitative results are presented in Table 2 in the main text, clearly demonstrating that our LVM-Lite-3B model achieves superior FID scores, highlighting its strong generative capabilities. Quantitative results are presented in Table 2 in the main text, clearly demonstrating that our LVM-Lite-3B model achieves superior FID scores, highlighting its strong generative capabilities. We suspect that the poor generative performance of the LVM-7B model stems from the absence of task-specific instruction data in its training dataset. In contrast, our two-stage training approach allows us to efficiently construct diverse visual instruction datasets with minimal training effort, enabling significantly improved generative performance. Qualitative results shown in Figure 15 reveal that on ADE20K, our model also struggles to infer objects accurately from semantic segmentation masks. This observation aligns with our segmentation task findings, highlighting our model's current limitations in recognizing object classes.

1.8. Qualitative Analysis

Visual examples of segmentation and generation tasks demonstrate that single-image pre-training enhances the model's ability to generalize and adapt across various datasets and conditions.

- Improved segmentation results are visually apparent after single-image pre-training.
- Enhanced diversity and fidelity in generative outputs.

1.9. Limitation and Broader Impacts

As discussed in Section 3, while our model shows excellent scalability, high-quality generation capabilities, and general task awareness, its performance on discriminative tasks, such as semantic segmentation, remains significantly lower compared to current state-of-the-art in-domain models. This modest segmentation performance may be partly due to the noise introduced during the tokenization of segmentation labels—performance remains substantially lower even with reconstruction from ground truth tokens. Additionally, the lack of pixel-to-pixel supervision, commonly used in supervised specialist models, further compounds the issue, as it is not employed in next-token prediction within LVM. Addressing this issue is beyond our current scope, as our focus is on efficiency and providing a comprehensive study on training effective LVM. We plan to leave this as future work.

Since this paper focuses on democratizing the training burden of current large vision models, we believe that migrating the training difficulty can help researchers reduce their research cycles and dedicate more efforts to developing robust novel methods. However, this paper's potential negative social impact is that our generative model might produce content using harmful or privacy-concerning training data that may be overlooked. To mitigate this, we will rigorously test our model and consider implementing gated access for safety concerns.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone

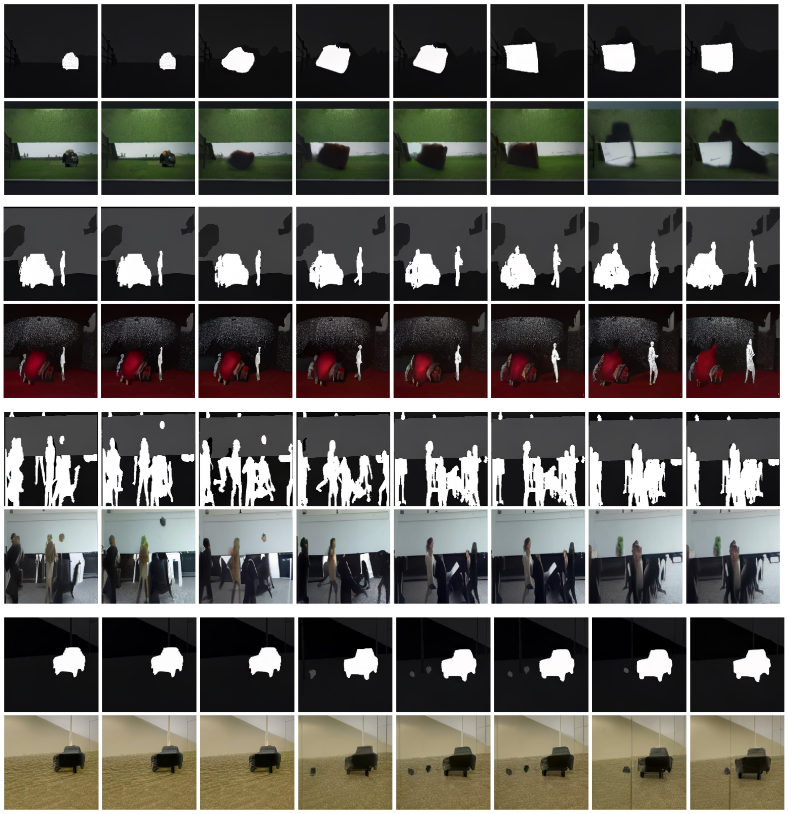

Figure 11. VIPSeg[20] qualitative results. Task: given 8 masks, generate the corresponding frames.

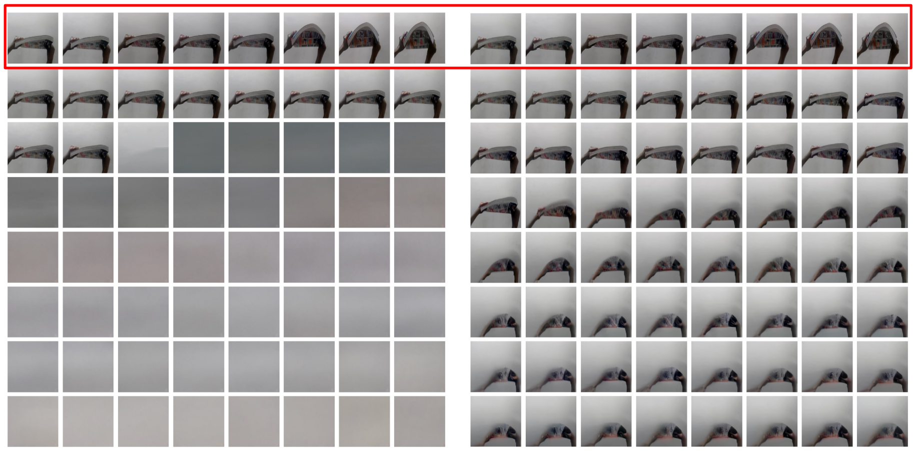

Figure 12. 64-frame video generation. Left: default train with 16-frame (4096 context length). Right: extended 64-frame (16K context length). Red rec.: a short action clip prompt. Task: predict the next 60 frames. Spatial resolution:256 × 256.

cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1692–1700, 2018. 4

[2] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *CVPR*, 2024. 1, 2, 4, 5

[3] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. 1

[4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 3

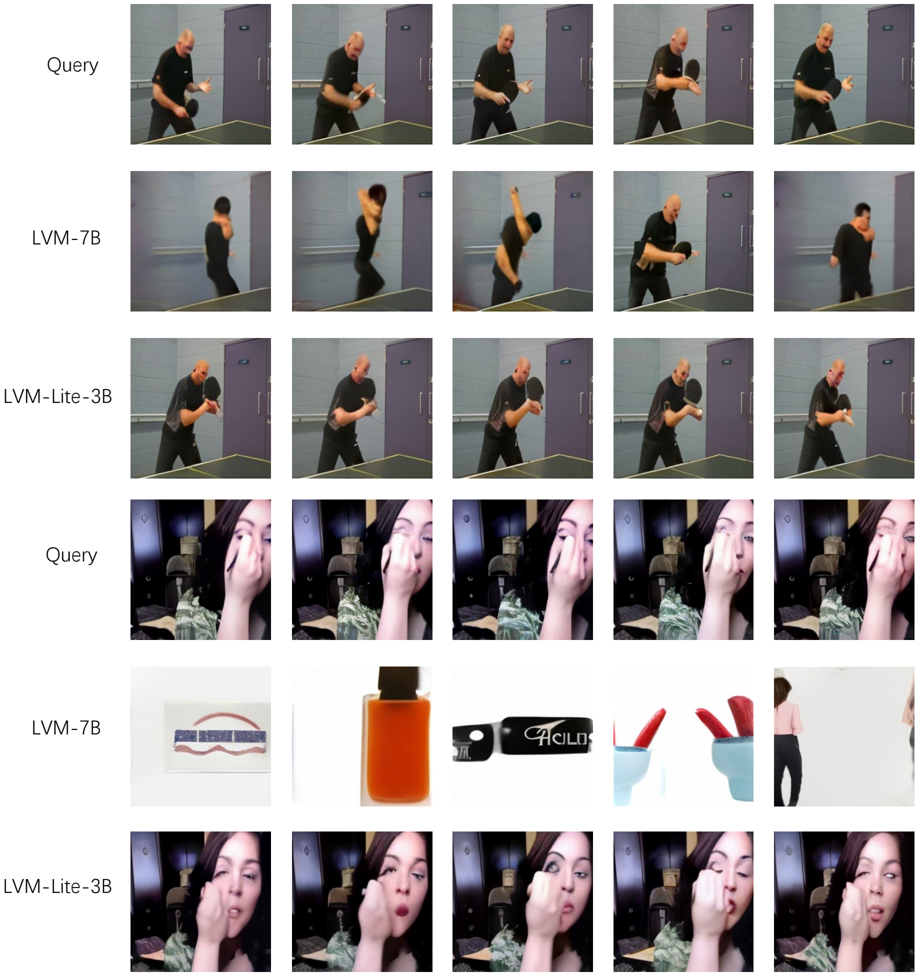

Figure 13. Frame comparison on UCF-101 [34]. We show two generated videos frame by frame.

[5] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 4, 5

[6] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2, 4

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 3, 4, 5

[8] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023. 4

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 3, 4, 5

[10] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. In *NeurIPS*, 2024. 4

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 4

[12] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 4, 5

[13] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 4

[14] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 4

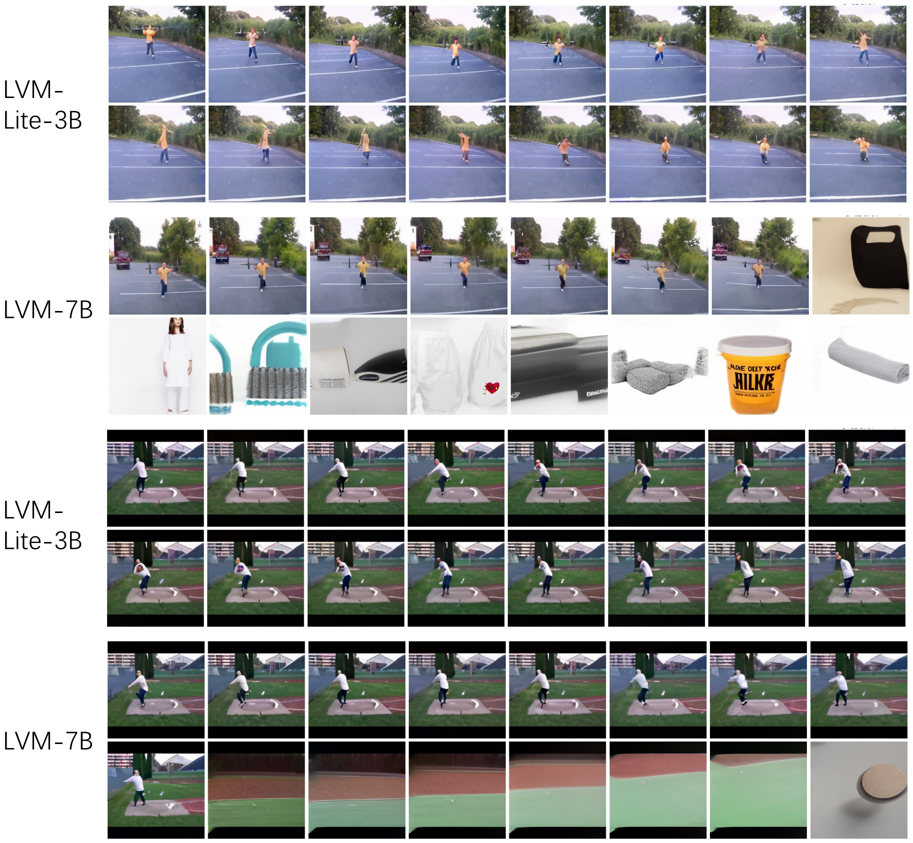

Figure 14. Video comparison on UCF-101 [34]. We show two whole-generated videos

[15] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *ICCV*, 2021. 4

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 3, 4

[17] Hao Liu and Pieter Abbeel. Blockwise parallel transformer for large context models. *Advances in neural information processing systems*, 2023. 1

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1

[19] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *ICCV workshops*, 2019. 4

[20] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *CVPR*, 2022. 3, 4, 7

[21] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019. 4

[22] Mathew Monfort, Bowen Pan, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Quanfu Fan, Dan Gutfreund, Rogério Schmidt Feris, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9434–9445, 2021. 4

[23] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012. 4

[24] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 4

[25] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016. 4

[26] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 4

[27] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah.

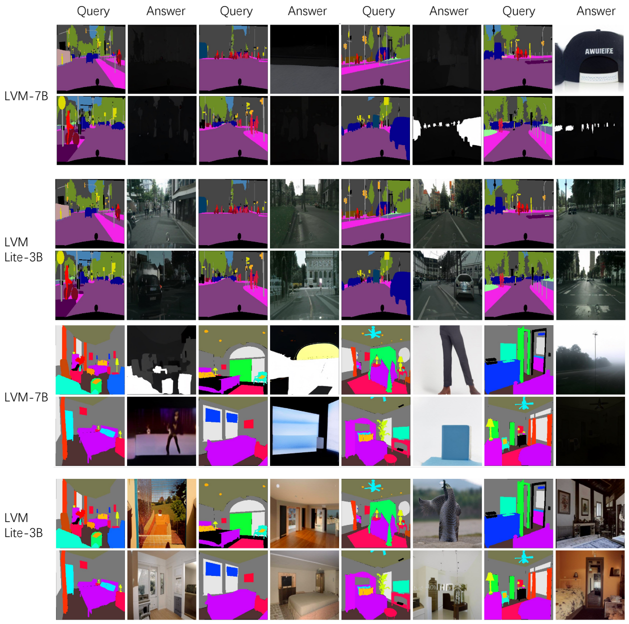

Figure 15. Conditional image generation.The first two rows are Cityscapes, and the last two rows are ADE20K.

Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 4, 5

[28] X. Soria, E. Riba, and A. Sappa. Dense extreme inception network: Towards a robust cnn model for edge detection. In *WACV*, 2020. 2, 4

[29] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, 2023. 1

[30] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 4

[31] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas S. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *CoRR*, abs/1809.03327, 2018. 4

[32] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 2, 4

[33] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. MAGVIT: Masked generative video transformer. In *CVPR*, 2023. 5

[34] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 2, 3, 5, 8, 9

[35] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 4, 5