Appendix Overview

This supplementary document complements the main manuscript by providing detailed insights and additional support. It is structured as follows:

Appendix A: Large Vision Models (LVMs) Detailed Overview – Explores the specifics of LVMs used in our study, including model sizes, architectural details, and optimization hyper-parameters.

Appendix B: Unified Vision Dataset (UVD) In-Depth Analysis – Provides a comprehensive examination of UVD, discussing its composition, data distribution, and more details.

Appendix C: Limitation – Shows the failure cases and limitations in this paper.

Appendix D: Additional Results – Offers extended results and visual evidence for our study, including supplementary figures and quantitative assessments.

A. Approach: Large Vision Models (LVMs)

A.1. Model Architectures.

As stated before, we use the Transformer variant of LLaMA [80] as our model architecture. To form different model sizes, we vary the hidden dimension, MLP intermediate dimension, number of heads and number of layers. We present the details in Table 3. For the rest of the hyperparameters, we keep them the same as the standard LLaMA model.

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

Table 3. Model architecture configurations of LVMs.

A.2. Training and optimizer details.

Folling the LLaMA [80] model, we use the AdamW optimizer to train our models. We use the same optimizer hyperparameters for all our models, and we present them in Table 4. All our models are trained on TPU-v3 pods on Google Cloud. Our largest model, LVM-3B, takes around 14 days to train on a v3-512 TPU pod.

Hyperparameter	Value
Learning rate schedule	linear warmup and cosine decay
Base learning rate	1.5e-4
Final learning rate	1.5e-5
Warmup steps	2000
Decay steps	144000
Weight decay	0.1
Optimizer	AdamW
Optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.95$
Batch size	2097152 tokens
Context length	4096 tokens

Table 4. Hyperparameters for pre-training LVM

B. Unified Vision Dataset (UVD) Details

B.1. Overview

The Unified Vision Dataset (UVD) represents an extensive compilation of visual data spanning a wide array of domains and annotation types. It integrates a diverse set of datasets, each contributing unique characteristics and annotations, thereby creating a rich resource for various vision-related tasks. The following Table 5 provides a detailed overview of UVD, categorizing the datasets into specific groups based on their content type and annotation features. This categorization includes unpaired image data, images with annotations, videos, videos with annotations, and synthetic 3D views. Each dataset within these categories is listed with its corresponding token count, annotation type, and annotation source, offering a comprehensive perspective of the UVD's structure and composition.

B.2. Summary of Dataset Distribution in UVD

The Unified Vision Dataset (UVD) encompasses a diverse array of visual data, aggregating over 430 billion tokens. The distribution of these tokens across various categories underscores the dataset's extensive coverage, see Figure 13:

Single Images (88.49%; 380.69 billion tokens) : This category, featuring datasets like LAION [70], is the largest, providing a vast collection of unannotated images suitable for a wide range of applications, particularly in unsupervised learning.

Images with Annotations (7.15%; 30.78 billion tokens) : Including prominent datasets such as ImageNet 1K [23] and COCO [52], this segment offers annotated images for image classification, object detection, semantic segmentation etc.

Videos (4.24%; 18.26 billion tokens) : Comprising datasets like UCF101 [78] and Moments in Time [57], this category provides unannotated video content, ideal for general video analysis and unsupervised learning in dynamic scenes.

Dataset	Tokens (Millions)	Annotation Type	Annotation Source
Unpaired Image Data			
LAION 5B [71] (1.5B images subset)	380690	-	-
Images with Annotations			
ImageNet 1K [23]	1317.40	Image Classification	Ground Truth
COCO [52]	363	Object Detection	MMDetection [14]
ADE 20K [100], Cityscapes [20]	66.88	Semantic Segmentation	Ground Truth
COCO [52], ImageNet 1K [23]	2078.06	Semantic Segmentation	Mask2Former [17]
COCO [52], lvmhp [49], mpii [3], Unite [47]	950.79	Human Pose	MMPose[19]
COCO [52], ImageNet 1K [23]	1623.85	Depth Map Image	DPT [65]
Subset of InstructPix2Pix [32]	415.46	Style Transfer	InstructPix2Pix [32]
COCO[52], ImageNet 1K[23]	1623.85	Surface Normal Image	NLL-AngMF [6]
COCO [52], ImageNet 1K [23]	1623.85	Edge Detection	DexiNed [79]
DID-MDN [98]	35.06	Rainy and Clean Image Pairs	Ground Truth
SIDD [2]	245.76	Denoised Image	Ground Truth
LOL[89]	0.458	Light Enhanced Image	Ground Truth
ImageNet 1K [23]	1321.07	Grayscale and Colorized Image Pairs	Ground Truth
ImageNet 1K [23]	1321.07	Inpainting	Ground Truth
Kitti [32]	9.21	Stereo	Ground Truth
Videos			
UCF101 [78]	109.11	-	-
DAVIS [63]	0.36	-	-
HMDB [46]	55.41	-	-
ActivityNet [11]	380.63	-	-
Moments in Time [57]	2979.00	-	-
Multi-moments in Time [58]	4124.04	-	-
Co3D [68]	228.75	-	-
Charades v1 [76]	241.53	-	-
Something-something v2 [35]	904.57	-	-
YouCook [21]	3.14	-	-
Kinetics 700 [12]	7092.04	-	-
MSR-VTT [92]	57.34	-	-
Youtube VOS [93]	63.70	-	-
jester [55]	606.47	-	-
diving48 [50]	150.73	-	-
MultiSports [51]	78.44	-	-
CharadesEgo [77]	193.06	-	-
AVA [59]	117.96	-	-
Ego4D [36]	1152.12	-	-
Videos with Annotations			
VIPSeg [56]	64.47	Video Panoptic Segmentation	Ground Truth
Hand14K [30]	1.96	Hand Segmentation	Ground Truth
AVA [59]	122.88	Video Detection	Ground Truth
JHMDB [41]	19.00	Optical Flow	Ground Truth
JHMDB [41]	37.92	Video Human Pose	Ground Truth
Synthetic 3D Views			
Objaverse [22] Rendered Multiviews	217.85	-	-

Table 5. Data sources of single images, images with annotations, videos and videos with annotations contained in UVDv1. In building the training data for LVM, we source annotations from a large number of datasets covering a diverge set of vision tasks. In addition to the ground truth annotations, we also leverage model-generated annotation to further broaden our diversity.

Videos with Annotations (0.06%; 0.25 billion tokens) : Though smaller in token count, this category is significant, with datasets like VIPSeg [56] and Hand14K [30] offering annotated videos for specific tasks like video segmentation and human pose estimation.

Synthetic 3D Views (0.05%; 0.22 billion tokens) : Datasets like Objaverse [22] in this category cater to advanced 3D vision tasks, providing synthetic 3D views for cutting-edge research.

Overall, UVDv1's rich composition, with its extensive token array, positions it as a comprehensive resource for var-

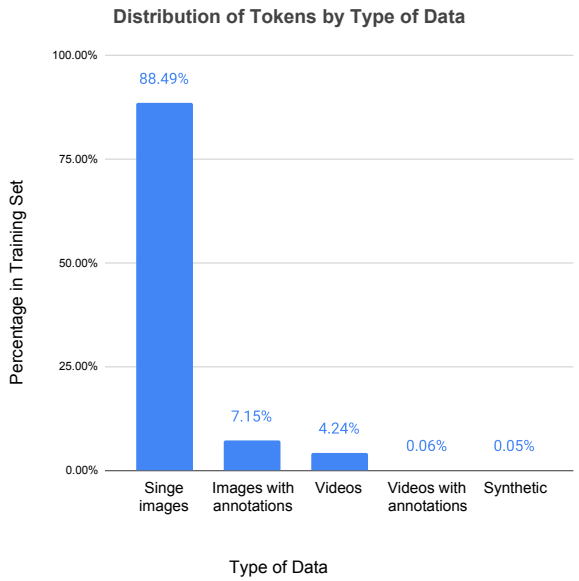

Figure 13. Tokens distribution of our training dataset. The majority of our training data comes from the single images of LAION, with the rest taking only 10%.

ious tasks in computer vision, from basic image processing to complex analyses in video and 3D data.

B.3. Details of Constructing Video Visual Sentences

We implemented specific tokenization strategies for each video dataset, taking into account their unique characteristics and contents. These tailored tokenization processes, inclusive of epoch details, ensure a comprehensive and diverse representation of each dataset's unique video content.

Something-something v2 [35]: Tokenized with strides of 4 and 7, capturing sequences of 16 frames. Random starting points were used for each of the 10 epochs to ensure diversity in human-object interactions.

CO3D [68]: Focused on 3D objects, tokenized with strides of 4 or 8 frames. Each sequence used 1 or 2 shots, with random starts in each epoch to capture object depth and detail.

Ego4D [36]: Strides of 12, 24, and 36 were employed, each sequence consisting of 16 frames. Randomization of starting points was implemented over 10 epochs to capture a range of egocentric activities.

Charades v1 [76]: Tokenized using strides of 10, 20, and 30 for 16-frame sequences. Random starting points across 2 epochs captured diverse narrative scenes.

Kinetics 700 [12]: Employed strides of 8 and 24, with each sequence capturing 16 frames. Random starts in each epoch over 10 epochs were used to represent a broad spectrum of human activities.

Diving48 [50]: Strides of 2 and 4 for tokenization, capturing 32-frame sequences to detail diving techniques. Random starting points were utilized across all epochs for comprehensive motion analysis.

AVA [59]: This dataset was tokenized with strides of 10 and 20, each sequence consisting of 16 frames. Random starts for sequences were used in each of the 50 epochs to capture varied human actions.

Jester [55]: Tokenized to capture the subtlety of hand gestures with 16-frame sequences. Randomization in the starting points was employed to enhance gesture diversity.

YouCook [21]: Tokenized with strides of 10, 20, and 30, each sequence comprising 16 frames. Random starting points over 4 epochs were used to capture a variety of cooking procedures.

CharadesEgo [77]: Focused on first-person narratives, tokenized using strides of 10, 20, and 30 for 16-frame sequences over 2 epochs.

YouTube VOS [93]: Tokenized using strides of 2, 4, and 8, focusing on detailed object movements within 16-frame sequences over 2 epochs.

MultiSports [51]: Captured sports actions with strides of 4, 8, and 12 for 16-frame sequences across 3 epochs.

ActivityNet [11]: Tokenized with strides of 5, 10, and 15, capturing 16 frames per sequence over 4 epochs to represent a wide range of activities.

Hand14K [30]: Focused on hand gesture recognition, tokenized with sequences of 16 frames, capturing detailed hand movements over multiple epochs.

Moments in Time [57]: Captured a wide array of activities and phenomena with a stride of 0, considering the short length of the videos, over multiple epochs.

Multi-Moments in Time [58]: An extension of Moments in Time, tokenized with strides of 0, 2, and 4 for different runs, each sequence comprising 16 frames to capture simultaneous actions over multiple epochs.

C. Limitations

Limited computing resources placed severe constraints that prevented us from exploring a range of intriguing problems, including the impact of different data sets and detailed ablation studies. It is important to note that, despite this being one of the biggest vision models to date, it is still rather small in comparison with modern Large Language Models. Therefore, the question of emergence and true generalization in Large Vision Models remains wide open and ripe for further study.

D. Additional Results

D.1. Sequential Prompting

Additional results for sequential prompting are presented, including:

Sketch Understanding: Figure 14 illustrates the model's capability in interpreting hand-drawn sketches from ImageNet-Sketch [86]. We construct visual sentence from a sequence of 15 images from ImageNet-Sketch [86] and then ask the model to predict the subsequent image. This method evaluates LVM's proficiency in interpreting and understanding hand-drawn sketches.

3D Rotation about arbitrary axes: In our evaluation set for Objaverse, we adopt a range of unseen objects to test LVM's ability to handle arbitrary axis rotation. The model predicts the next 4 images based on a visual sentence of 16 images. As illustrated in Figure 15, LVM demonstrates its capacity to reason about the direction of spatial rotation based on the context provided by the prompt, leading to reasonable predictions. For this tasks, LVM exhibits 11.8 as in perplexity.

Frames Prediction: Figures 16 to 21 demonstrate frame prediction using the evaluation set from Kinetics 700 dataset. The model predicts the next 4 frames based on a visual sentence of 16 frames. The Fréchet Inception Distance (FID) score for single-frame prediction conditioned on 15 frames is 21.018, indicating the LVM's proficiency in understanding spatial and temporal dynamics.

D.2. Analogy Prompting

Further results for analogy prompting in various contexts are provided, highlighting the model's adaptability and understanding in different scenarios.

Pose Estimation Analogy: In Figure 22, the pose estimation analogy is constructed using the visual sentence of "image-to-joint", where the model predicts poses from given images. This assesses the model's ability to interpret analogy pairs and understand human poses and joint relationships.

Depth Estimation Analogy: Figure 23 presents the "image-to-depth" analogy for depth estimation. The visualizations utilize the validation set from [23], whose annotations are generated by DPT [65], and re-normalised to [-1, 1] following [54].

Surface Normal Estimation Analogy: The "image-to-surface normal image" analogy is depicted in Figure 24. This analogy tests the model's depth of understanding of 3D structures from 2D data. Despite inaccuracies in some normal surface images from the prompts, our model shows notable robustness and generalization.

Semantic Segmentation Analogy: Results for the "image-to-segmentation" analogy are shown in Figure 25, emphasizing semantic segmentation. The visualizations are based on the validation set from ADE20K [100].

Edge Detection Analogy: Results for the "image-to-edge" analogy are shown in Figure 26, emphasizing edge detection. The visualizations are based on the validation set from [23], annotated using DexiNed [79].

Image Inpainting Analogy: In Figure 27, the "partially masked image-to-image" analogy is explored, demonstrating the model's capabilities in image inpainting. The model is challenged with different mask ratios, showing significant semantic understanding, as evidenced by a Mean Squared Error (MSE) of 0.106.

Image Colorization Analogy: Figure 28 shows the "grayscale image-to-image" analogy for image colorization. This test showcases the model's ability to handle complex image scenarios, with an MSE of 0.51.

Derain Analogy: Figure 29 shows the "rainy image-to-image" analogy for image deraining.

Figure 14. **Sketch understanding.** We construct visual sentences by sequences of sketches. LVM is asked to predict the next image.

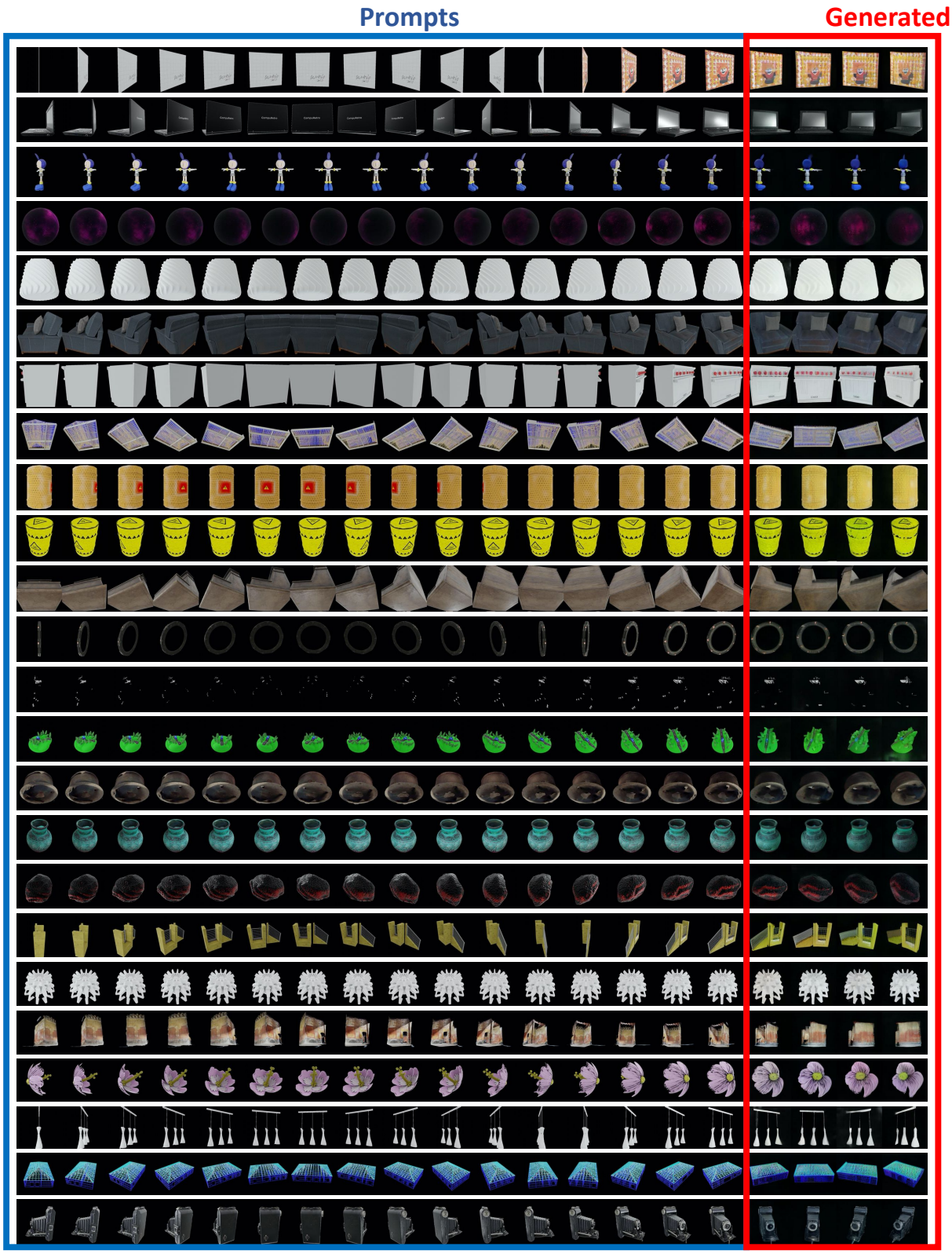

Figure 15. **3D Rotation about arbitrary axes.** We construct visual sentences by rotating images. LVM is asked to predict the next 4 views.

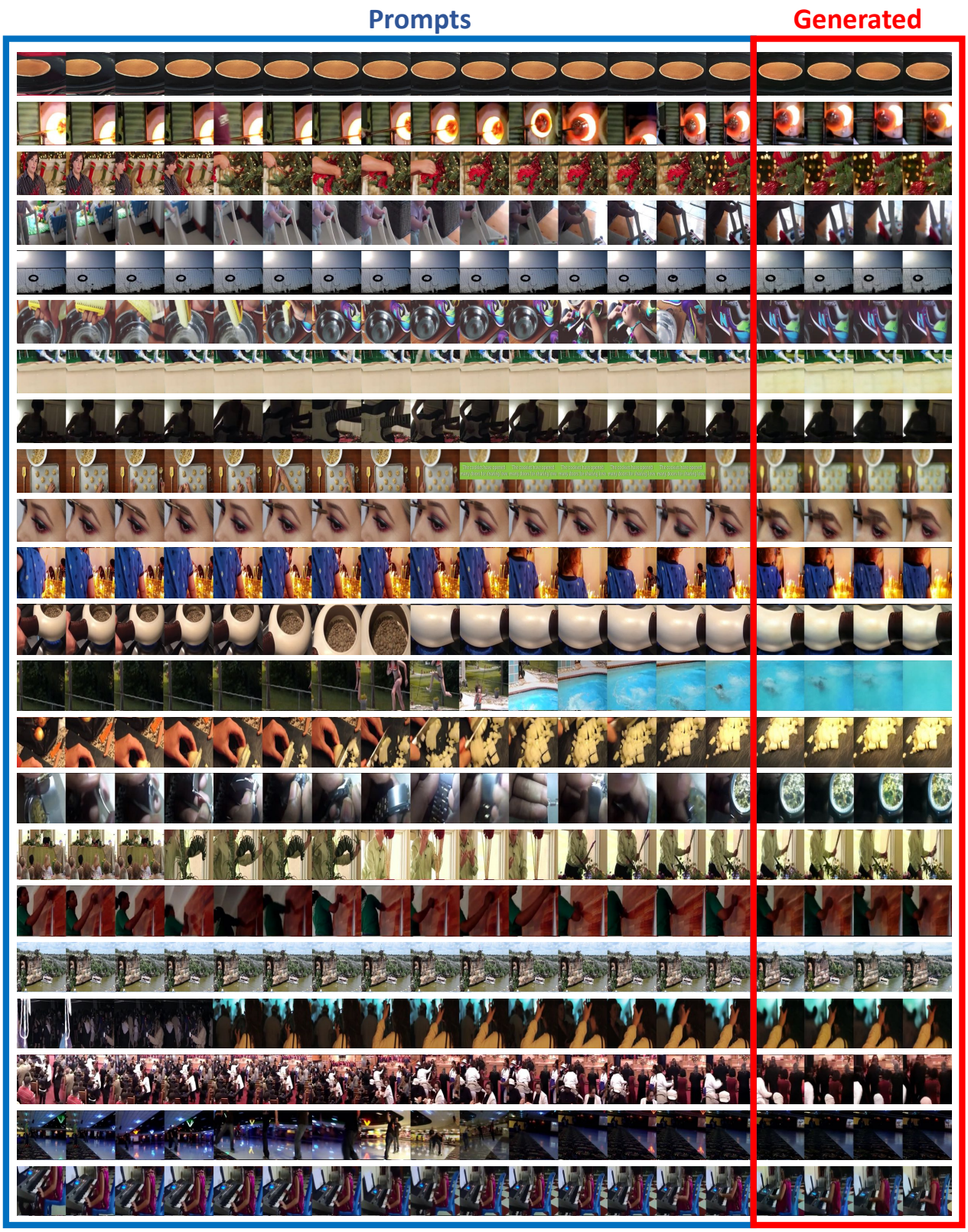

Figure 16. **Frames prediction.** We construct the visual sentence by sequences of frames. LVM is asked to predict the next 4 frames.

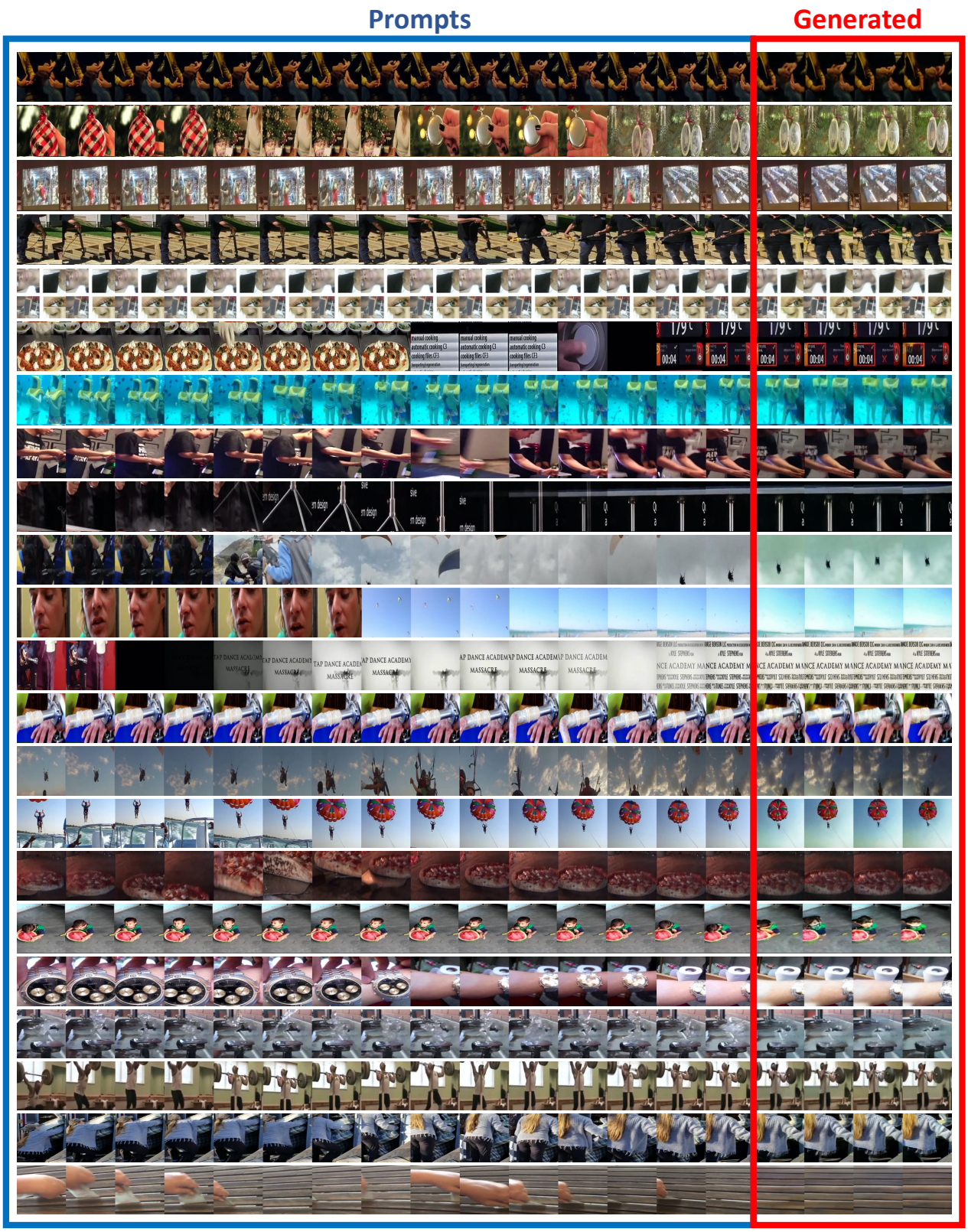

Figure 17. **Frames prediction.** We construct the visual sentence by sequences of frames. LVM is asked to predict the next 4 frames.

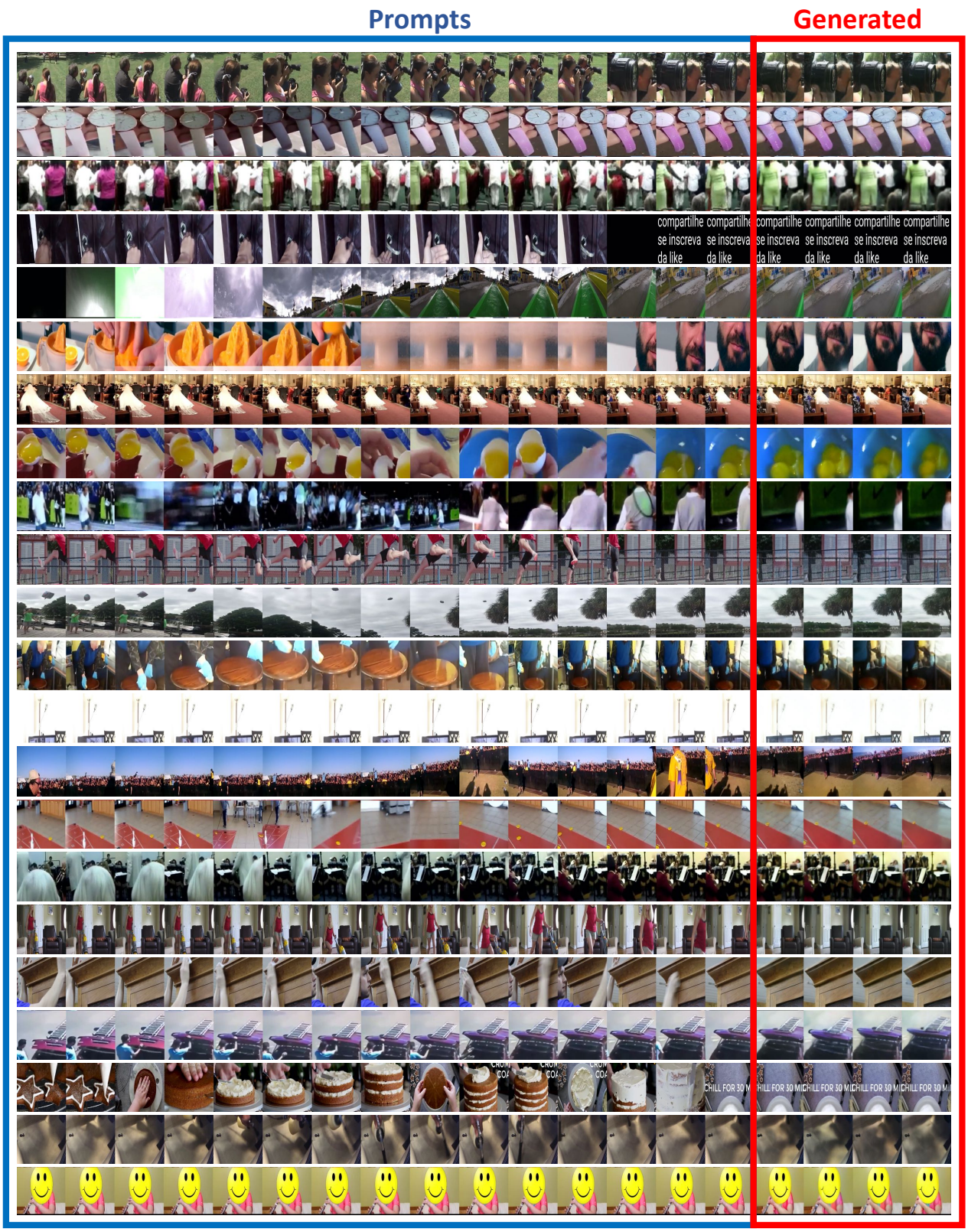

Figure 18. **Frames prediction.** We construct the visual sentence by sequences of frames. LVM is asked to predict the next 4 frames.

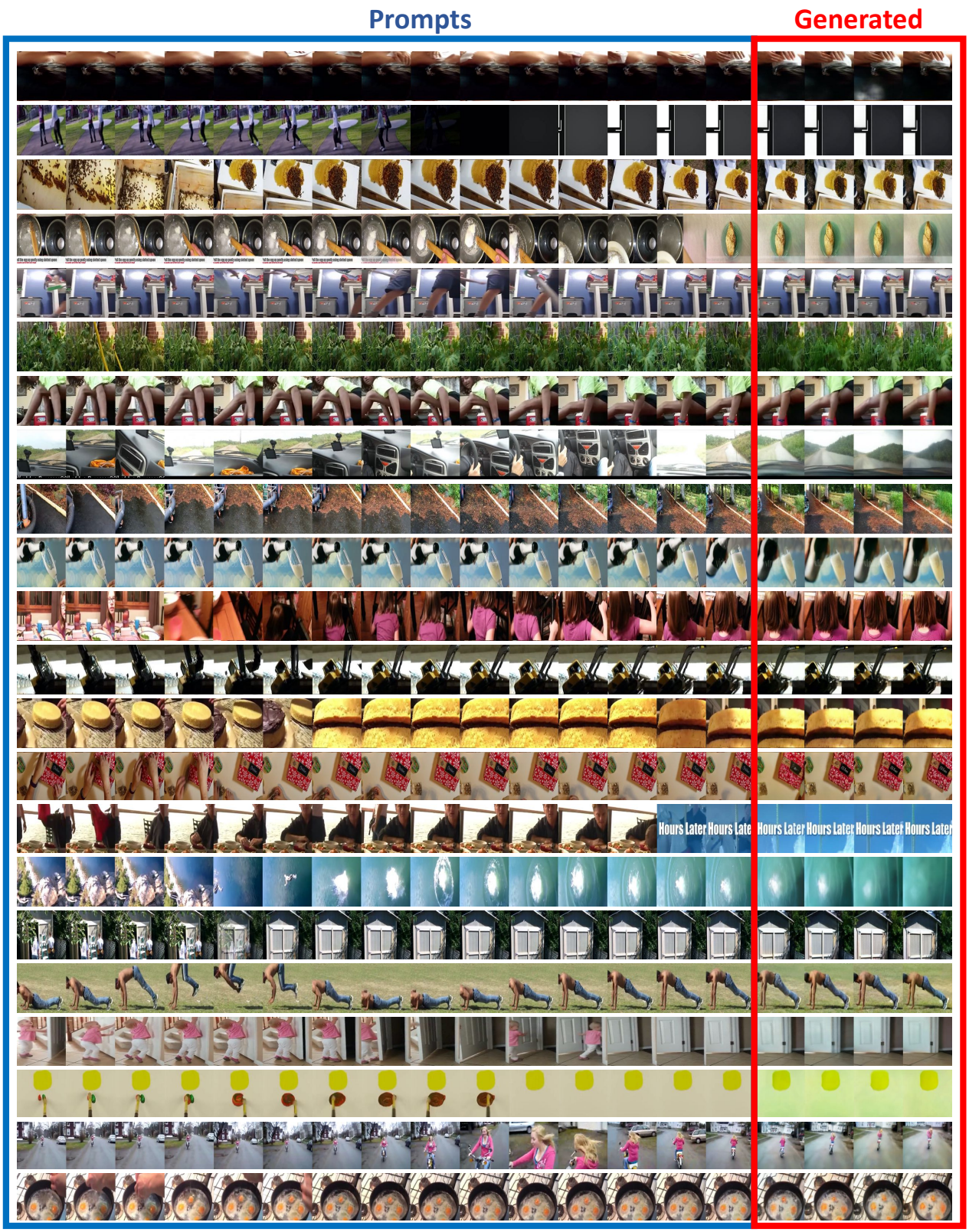

Figure 19. **Frames prediction.** We construct the visual sentence by sequences of frames. LVM is asked to predict the next 4 frames.

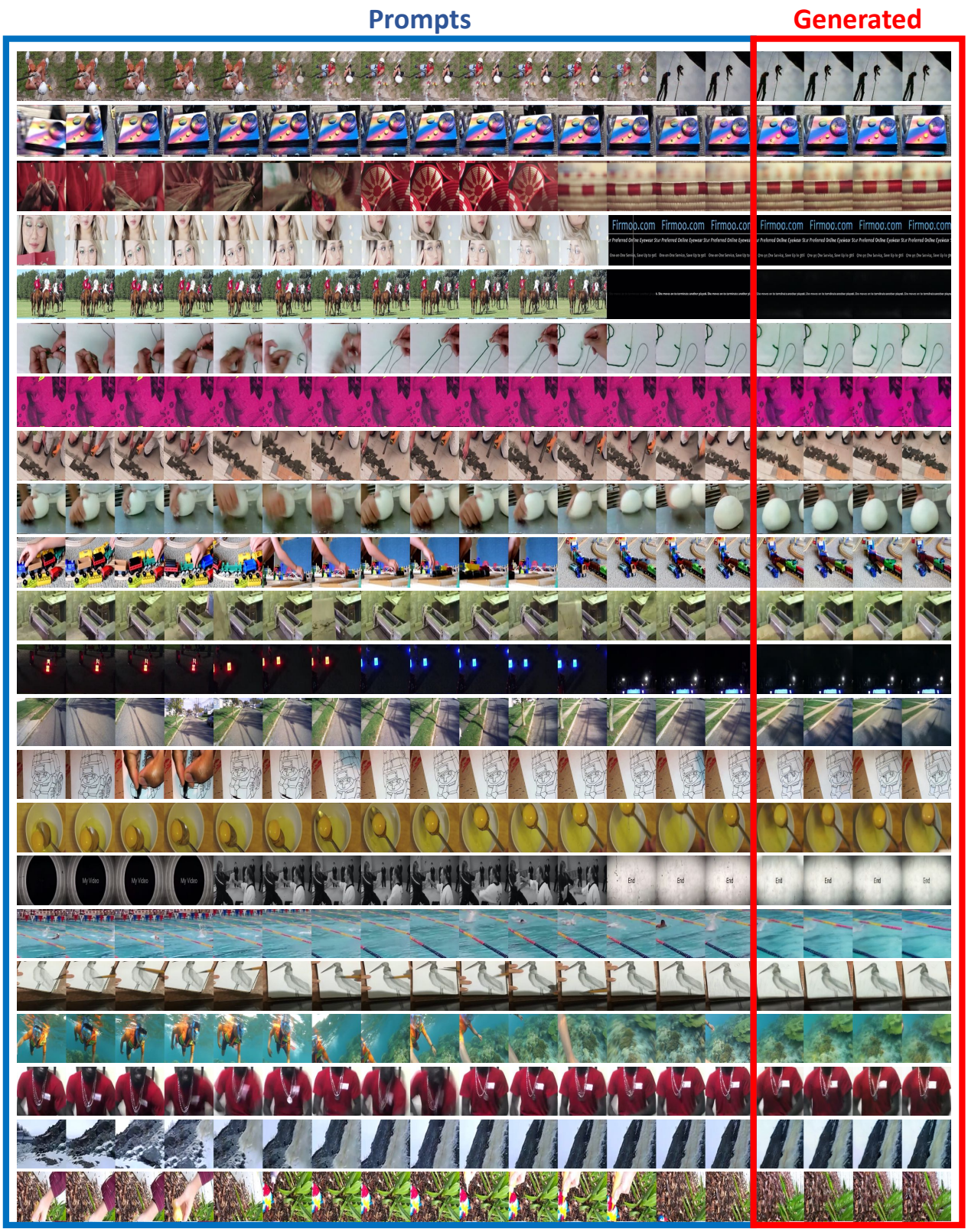

Figure 20. **Frames prediction.** We construct the visual sentence by sequences of frames. LVM is asked to predict the next 4 frames.

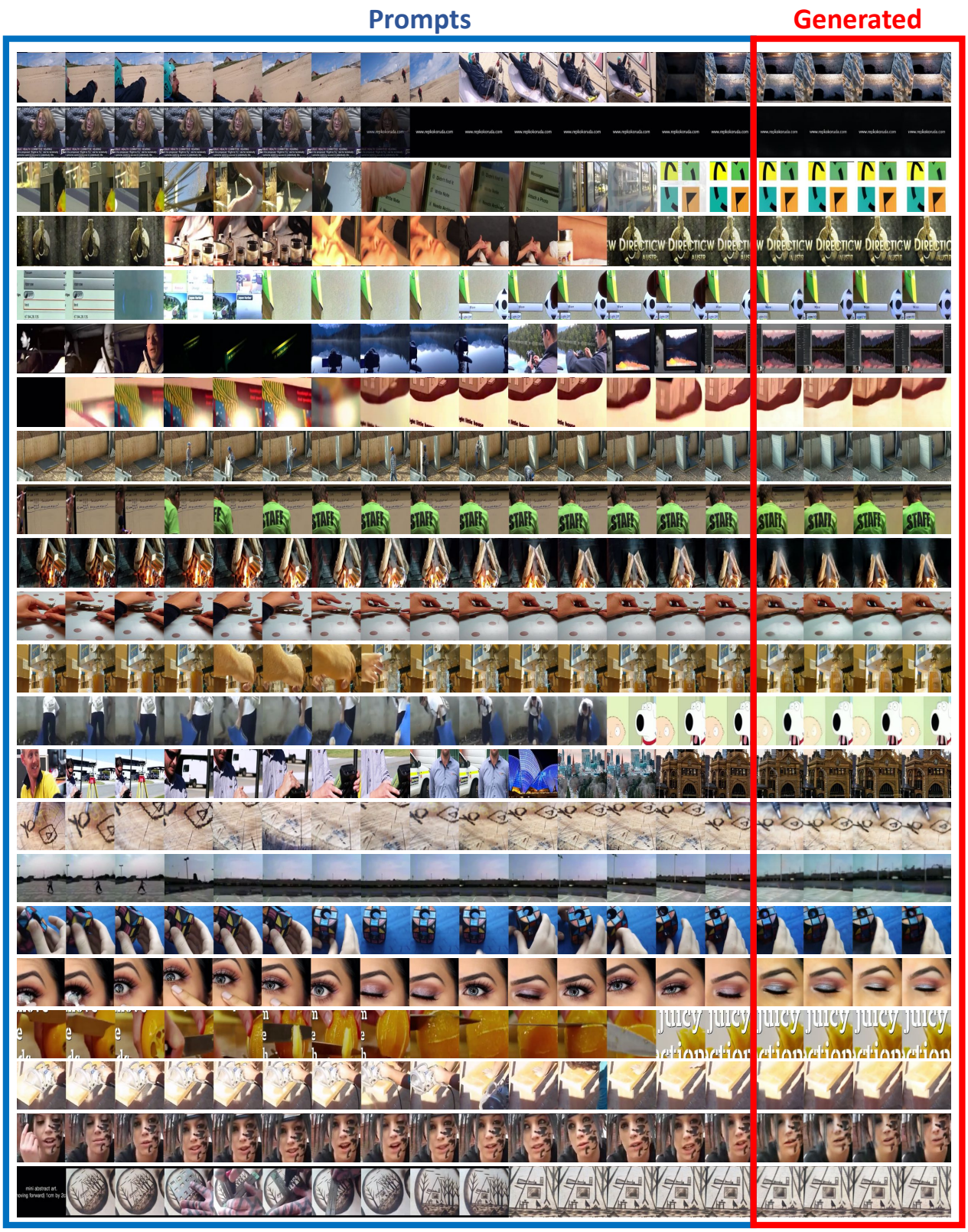

Figure 21. **Frames prediction.** We construct the visual sentence by sequences of frames. LVM is asked to predict the next 4 frames.

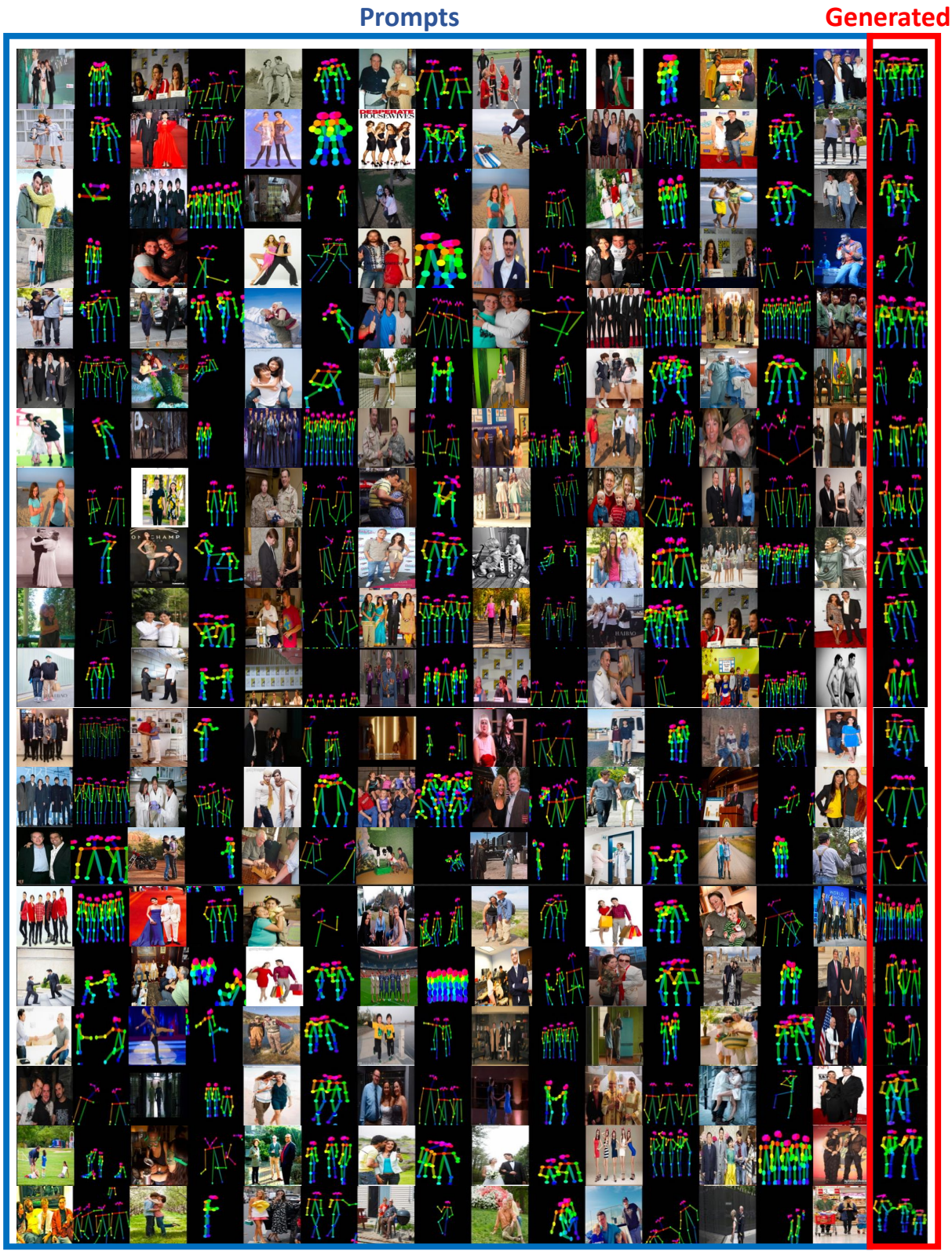

Figure 22. **Human keypoint detection.** We construct the visual sentence by "image-to-joints" analogy prompting from LVMHP [49] dataset. LVM is asked to predict the skeleton of all humans in the image.

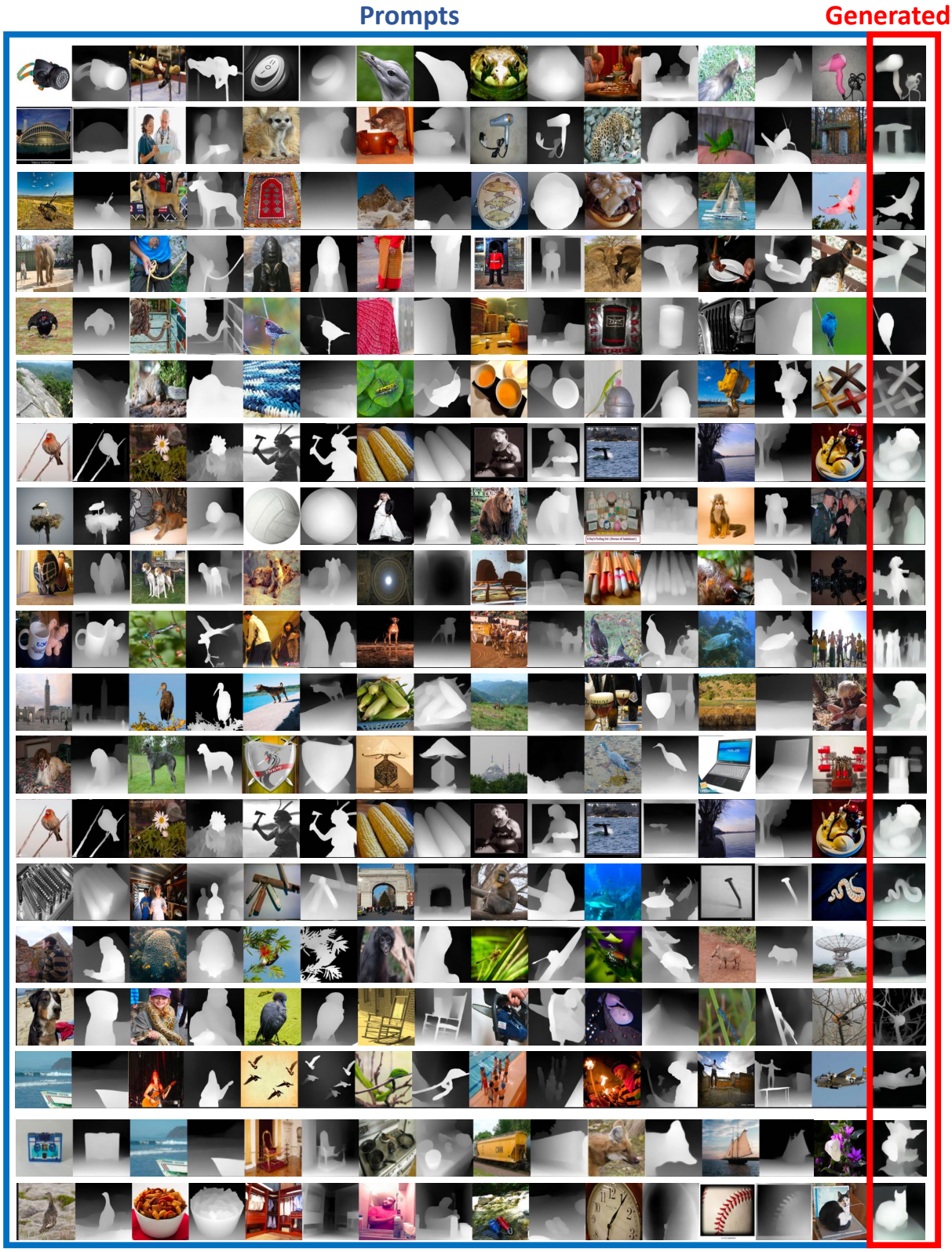

Figure 23. **Depth Estimation.** We construct the visual sentence by "image-to-depth image" analogy prompting from ImageNet validation set. LVM is asked to predict the depth map.

Figure 24. **Surface Normal Estimation.** We construct the visual sentence by "image-to-surface normal image" analogy prompting from ImageNet validation set. LVM is asked to predict the surface normal estimation map.

Figure 25. **Semantic Segmentation.** We construct the visual sentence by "image-to-segmentation" analogy prompting from ADE 20K validation set. LVM is asked to predict semantic segmentation color map.

Figure 26. **Edge Detection.** We construct the visual sentence by "image-to-edge" analogy prompting from ImageNet validation set. LVM is asked to predict the edge map given a new image.

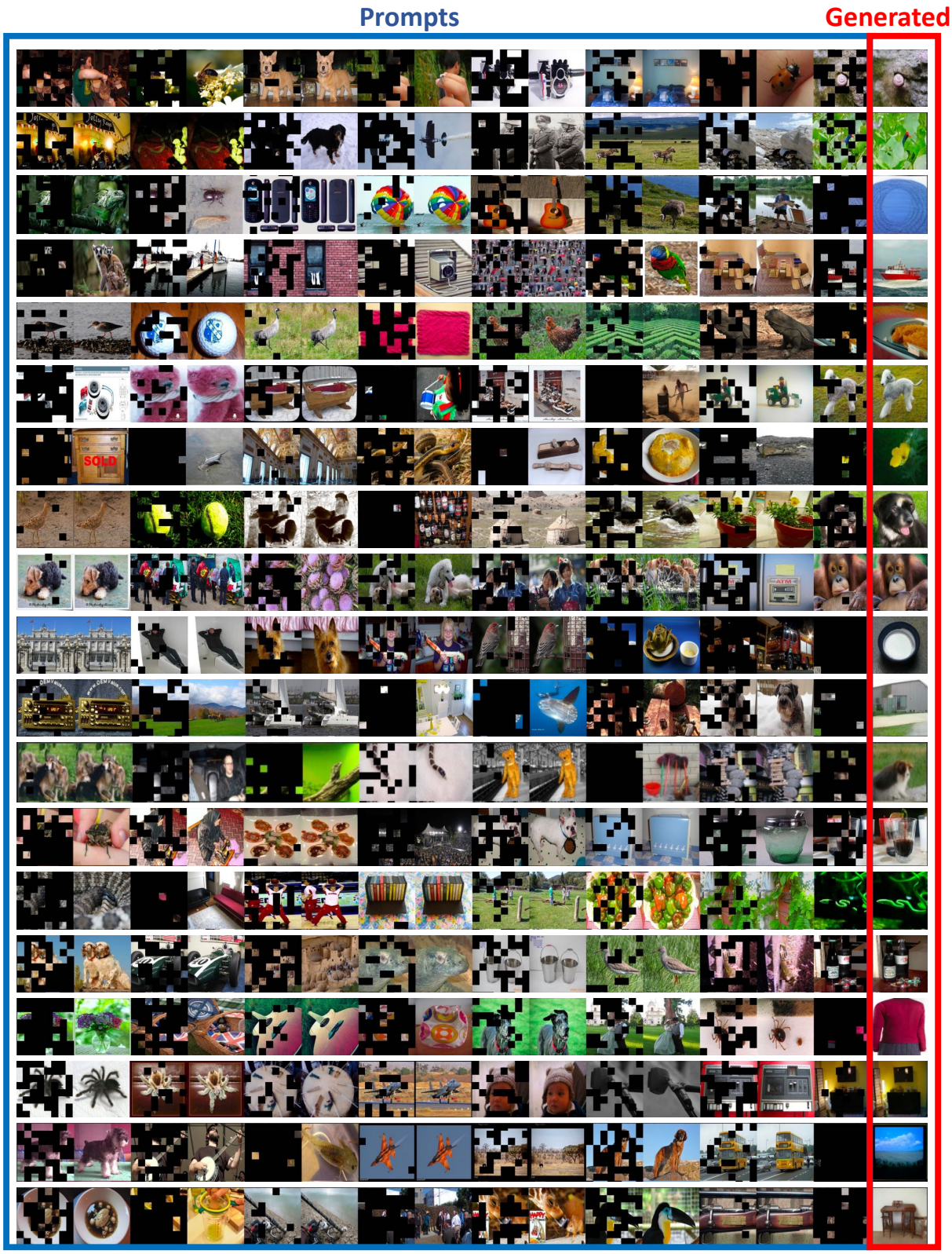

Figure 27. **Inpainting.** We construct the visual sentence by "partially masked image-to-image" analogy prompting from ImageNet validation set. LVM is asked to reconstruct the pixel of the masked region given a new partially masked image.

Figure 28. **Colorization.** We construct the visual sentence by "gray-scale image-to-image" analogy prompting from ImageNet validation set. LVM is asked to colorize the image given a new gray-scale image.

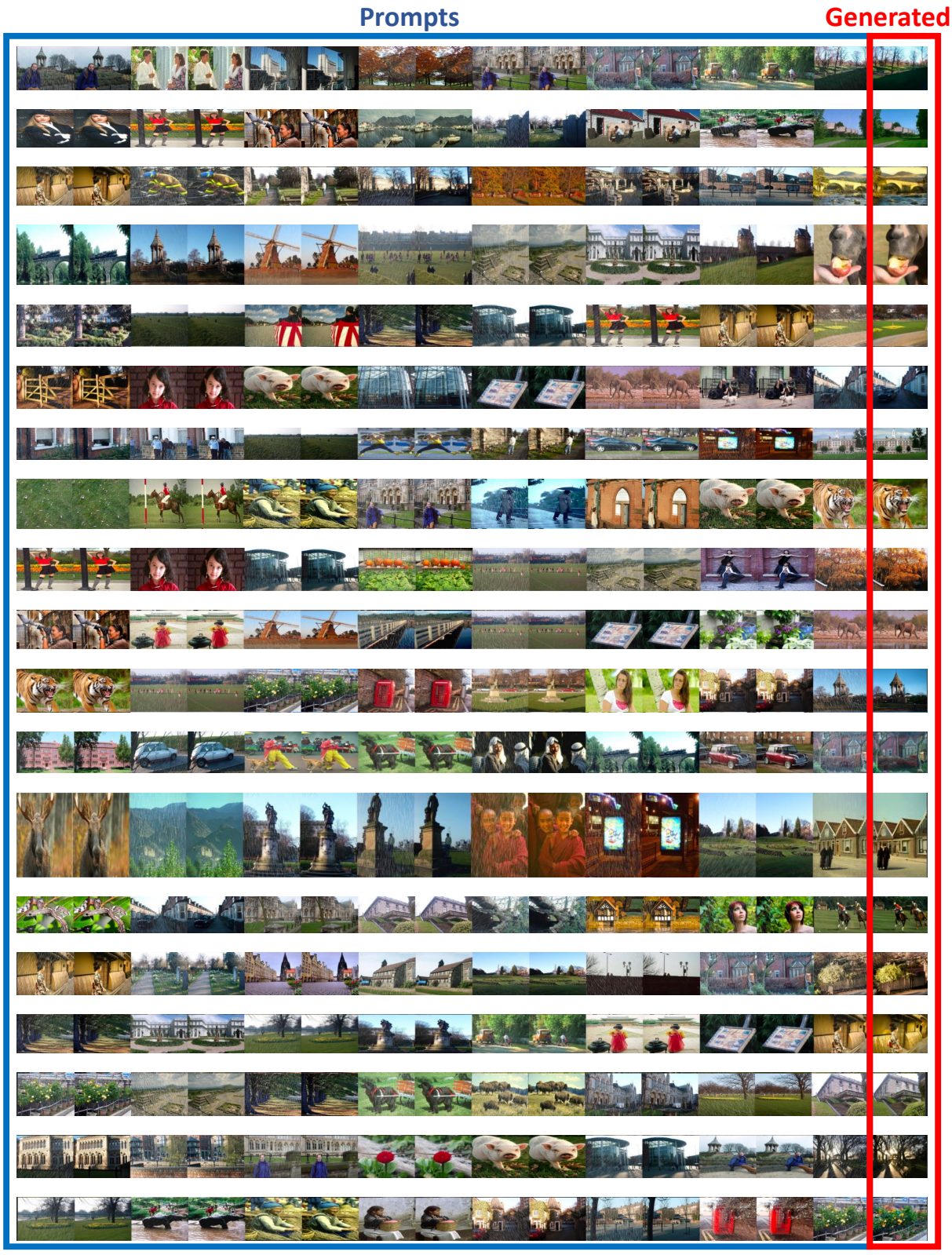

Figure 29. **Derain.** We construct the visual sentence by "rainy image-to-image" analogy prompting from DID-MDN [98] validation set. LVM is asked to derain the image.