SSGVS: Semantic Scene Graph-to-Video Synthesis Supplementary Material

Yuren Cong[1], Jinhui Yi[2], Bodo Rosenhahn[1], Michael Ying Yang[3]
[1]TNT/L3S, Leibniz University Hannover, [2]University of Bonn, [3]SUG, University of Twente

A. Transformer architecture

We adopt a GPT-like multi-layer Transformer in this paper. Each transformer layer consists of a classical multi-head attention module, a feed-forward network, and normalization layers as shown in Figure 1. We use the original full attention mechanism but not sparse attention in Transformers. The feed-forward network is a two-layer perceptron, while layer normalization is used in the Transformers for normalization.

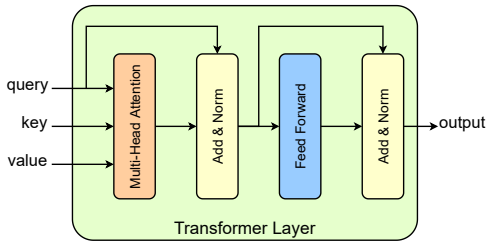

Figure 1. Architecture of the Transformer layer, which contain a multi-head attention module, a feed-forward network, and two normalization layers.

B. Dataset details

We split a sub-dataset from Action Genome [3], which is built upon Charades [6]. To include more complex semantic variations in the 16-frame video, we sampled 1 frame every 5 frames from the original videos of Charades and resize the sampled frames to a resolution of 128×128. We only keep the objects whose bounding boxes with short edges larger than 16 pixels. In order to avoid overly complex scene graphs that make the representations difficult to infer, we reduce the graph fidelity by cutting out redundant nodes in the scene graph and keep a maximum of 5 object nodes. In addition, each video contains at least 5 video scene graphs so that the video scene graph (VSG) encoder has enough information to infer the graph representations that are not given. In the split dataset, there are 36 object categories and 17 relationship categories. The distribution of object and relationship occurrences are illustrated in Figure 2.

C. Metrics details

Fréchet video distance (FVD). FVD [8] is developed from Fréchett Inception Distance (FID) [2], which is widely-used to evaluate the performance of image generation models. FVD takes into account a distribution over entire videos in order to avoid the disadvantages of frame-level metrics. A pre-trained Inflated 3D Convnet [1] is used to capture video feature distributions. The 2-Wasserstein distance between the ground truth video distribution and the synthetic video distribution is calculated as the metrics.

Structural similarity index measure (SSIM). SSIM [9] is a per-frame perceptual metrics that measures the similarity between two images. The statistical measure combines three different factors: luminance, variance and correlation. We first split the ground truth videos and synthetic videos into single frames. Then we calculate SSIM between the ground truth frames and synthetic frames. The average SSIM of all frames is taken as the final result.

D. Technical implementation details

Video scene graph representation learning framework. In the video scene graph encoder, both the spatial Transformer and temporal Transformer have 3 Transformer layers. We employ 4 attention heads for each attention module, while the dimension d of the input queries, keys, and values is set to 256. The encodings are only added to queries and keys when using the attention modules. For the frame encoder, we adopt the CNN-based model used in [10], which is built upon Inception-v3 model [7]. The input frames are first resized to a resolution of 299×299, while the size of the feature maps extracted by the CNN backbone is $768 \times 17 \times 17$. A 1×1 convolution layer is exploited to reduce the dimension of the feature maps to $d = 256$. Then we use a global average pooling layer to convert the feature maps to the frame vectors. We train the video scene graph encoder and frame encoder using ADAM optimizer [4] with

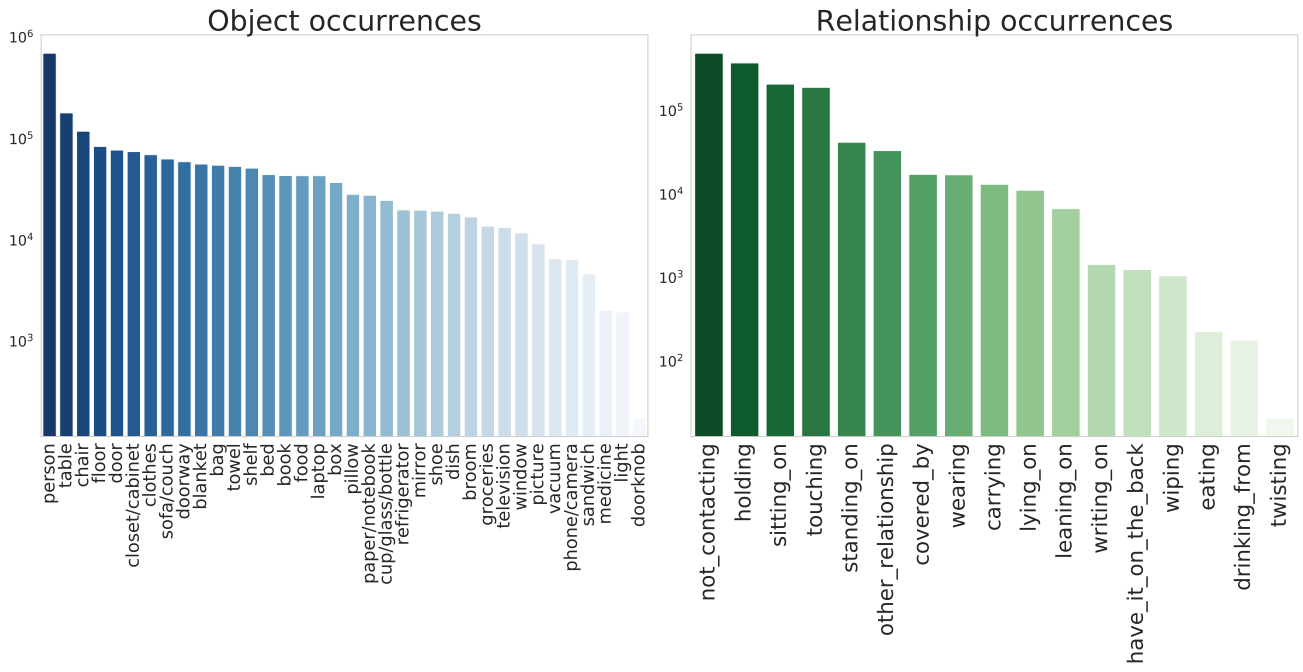

Figure 2. Distributions of object (left) and relationship (right) occurrences for the sub-dataset split from the Action Genome dataset.

a learning rate of 1×10^{-4} and a batch size of 12 images. The training takes about 20 hours on 2 RTX 2080 TI GPUs. The values of different loss functions in the training are shown in Figure 3.

Semantic scene graph-to-video synthesis framework. We adopt the VQ-VAE from [5] and use almost the same hyperparameters. The encoder of the VQ-VAE converts a 128×128 frame into a $512 \times 8 \times 8$ feature map. The 8×8 sub-vectors of the feature map are then quantified to the discrete latent embeddings. The length of the latent codebook is set to 1024 to shorten the training time. The 8×8 discrete latent embeddings are reconstructed to a video frame by the decoder of the VQ-VAE. We train the VQ-VAE using ADAM optimizer [4] with a learning rate of 2×10^{-4} and a batch size of 32 videos on 8 RTX 3090 TI GPUs for about 48 hours.

The auto-regressive Transformer consists of 24 Transformer layers with a head number of 16. Due to the complexity of the auto-regression task, the embedding dimension of the attention module is set to 1024. Therefore, a linear transformation is utilized to project the dimension of video scene graph representations from 256 to 1024, while 1024 latent embeddings with dimension $1024d$ are learned during the training. We train the auto-regressive Transformer using ADAM optimizer [4] with a learning rate of 1×10^{-5} and a batch size of 64 videos on 8 RTX 3090 TI GPUs for about 48 hours. Furthermore, the video scene graph encoder is frozen during the training of the auto-regressive Transformer.

E. Additional qualitative results and limitations

Additional qualitative results. The details such as the human face are not well presented in qualitative examples in the main paper. As discussed, the reason is that the motion in the video is quite large. Another simple example is shown in Figure 4. In the original video, the girl is holding and looking at the book (all the video scene graphs are the simple triplet `person-holding-book`). Although there is some change in the position of the girl's head and book, it is not significant. In this case, SSGVS can render better details and perform well. The original video and some generated frames are omitted because the synthetic frames are very close to the original ones and the motion is small. To visualize the small motion better, we also compute the optical flows for the shown synthetic frames. More videos synthesized by SSGVS are attached to the supplementary material.

In Figure 5, we show the video synthesized by CCVS [5], which only use the first frame as input, and the video synthesized by our SSGVS which use the first frame and also the video scene graphs. With the help of the input video scene graphs, SSGVS can synthesize higher quality frames, especially those far from the starting frame. In this example,

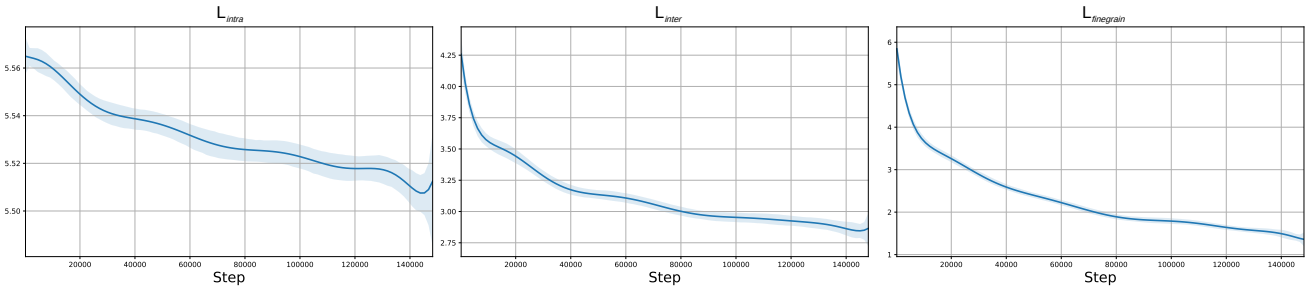

Figure 3. Graphical intra-video contrastive loss, inter-video contrastive loss and fine-grained contrastive loss in the training.

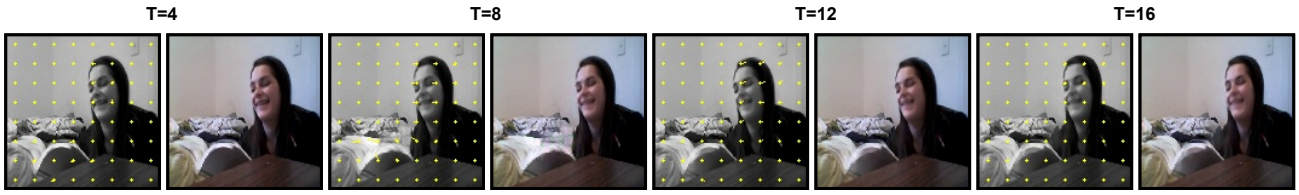

Figure 4. Qualitative result for simple video synthesize, in which the girl keeps holding the book. The motion in the original video is very small. Only 4 synthesized frames and their optical flow are shown. In this case, the details such as the face are well rendered.

there are no significant semantic changes in the video scene graphs. They control SSGVS to generate the frames that maintain the current drinking action, whereas the distortion in the frames generated by CCVS is getting worse.

Limitations. Since the resolution of our generated video is 128×128, this constraint makes some small objects such as the phone and medicine cannot be presented very clearly. In addition, for some videos containing the large motion, the auto-regressive transformer cannot successfully predict the sequence of the latent embeddings. These videos usually involve a change of scene or camera pose. An example is shown in Figure 6. We can increase the size of the synthesis model or adopt a smaller time step to suppress this issue.

F. Ethics statement

As machine learning methods are increasingly used in everyday life, it makes sense to consider the potential social impact of our work. Our work could potentially be used for deep fake as well as other state-of-the-art generative models. Since our model can synthesize videos with specific semantic content, this even makes deep fake more flexible. Developing better models has the potential to be used maliciously to violate human likeness rights or create false information. On the other hand, a good video synthesis model helps the film and video game industries, for example, by replacing live actors in dangerous scenes. It can be also very promising in the metaverse.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[3] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020. 1

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1, 2

[5] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. Ccvs: Context-aware controllable video synthesis. *Advances in Neural Information Processing Systems*, 34:14042–14055, 2021. 2

[6] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526, 2016. 1

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE con-*

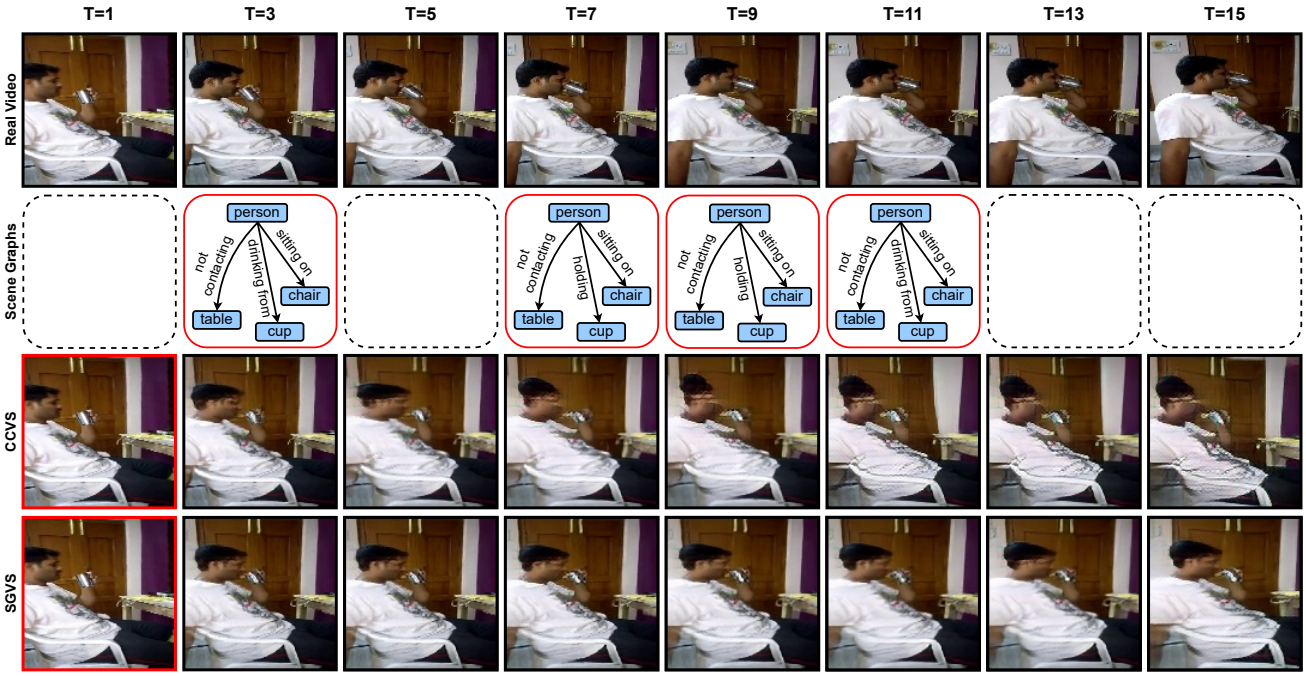

Figure 5. Comparison between the videos synthesized by CCVS and SSGVS. The real frames are given in the first row, while the corresponding video scene graphs are shown in the second row. The video synthesized by SSGVS has higher fidelity with the help of the discrete video scene graphs.

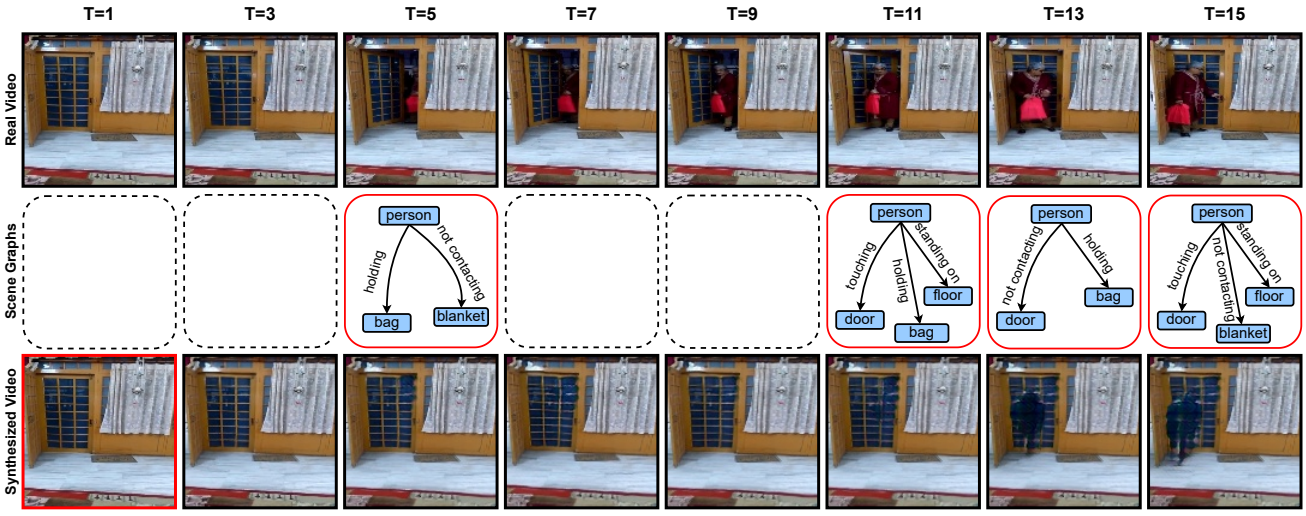

Figure 6. Failure to synthesize a complex video with large motion. The person and the bag cannot be synthesized since they do not appear in the first frame. In this case, only the silhouettes of a standing person are visible in the frames of $T = 13$ and $T = 15$.

ference on computer vision and pattern recognition, pages 2818–2826, 2016. 1

[8] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 1

[9] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[10] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 1