A Recipe for Scaling up Text-to-Video Generation with Text-free Videos
Supplementary Material

Xiang Wang[1]* Shiwei Zhang[2]† Hangjie Yuan[3] Zhiwu Qing[1] Biao Gong[2] Yingya Zhang[2]
Yujun Shen[4] Changxin Gao[1] Nong Sang[1]†

[1]Key Laboratory of Image Processing and Intelligent Control,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2]Alibaba Group [3]Zhejiang University [4]Ant Group

{wxiang,qzw,cgao,nsang}@hust.edu.cn, {zhangjin.zsw,yingya.zyy}@alibaba-inc.com
hj.yuan@zju.edu.cn, {a.biao.gong,shenyujun0302}@gmail.com

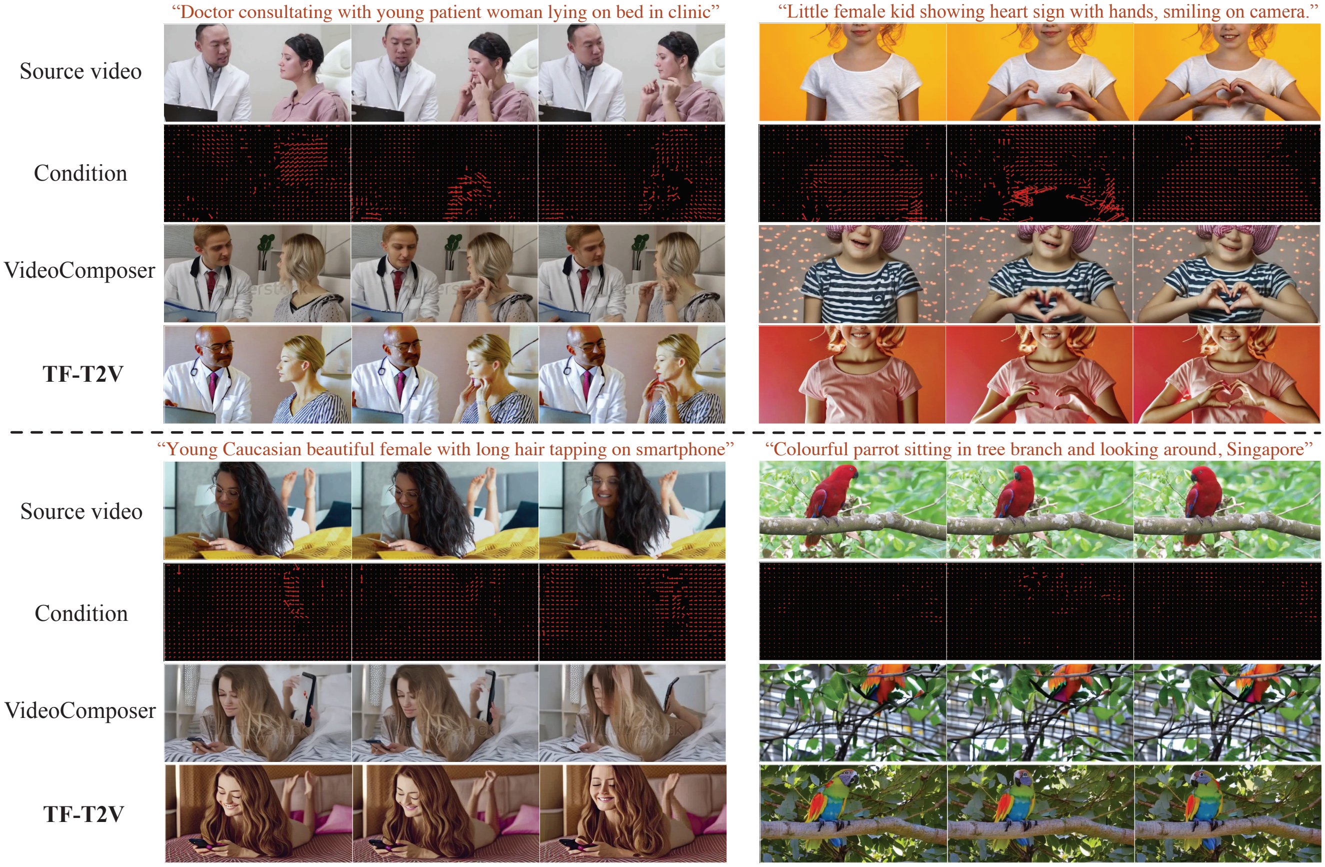
"Doctor consultating with young patient woman lying on bed in clinic"
"Little female kid showing heart sign with hands, smiling on camera."
Source video
Condition
VideoComposer
TF-T2V
"Young Caucasian beautiful female with long hair tapping on smartphone"
"Colourful parrot sitting in tree branch and looking around, Singapore"
Source video
Condition
VideoComposer
TF-T2V

Figure 1. **Qualitative comparison on compositional motion-to-video synthesis**. The videos are generated by taking textual descriptions and motion vectors as conditions. Compared to VideoComposer, TF-T2V produces more realistic and appealing results.

Due to the page limit of the main text, we add more details and experimental results in this appendix. Besides, limitations and future work will also be discussed.

1. More experimental details

In Fig. 1, we show the comparison on compositional motion-to-video synthesis. TF−T2V achieves more appealing results than the baseline VideoComposer [2]. Following

* Intern at Alibaba Group. † Corresponding authors.

Table 1. Ablation study on different training manners.

Setting	FID (↓)	FVD (↓)	CLIPSIM (↑)
Separately	9.22	503	0.2905
Jointly	**8.19**	**441**	**0.2991**

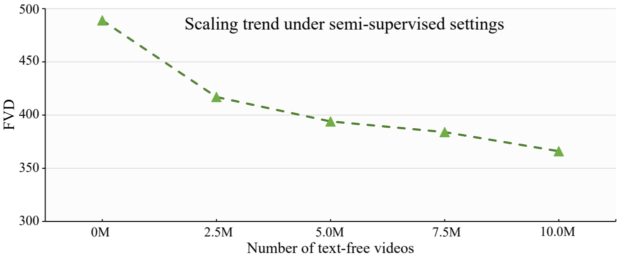

Figure 2. **Scaling trend under semi-supervised settings**. In the experiment, labeled WebVid10M and text-free videos from Internal10M are leveraged.

prior works, we use an off-the-shelf pre-trained variational autoencoder (VAE) model from Stable Diffusion 2.1 [1] to encode the latent features. The VAE encoder has a downsample factor of 8. In the experiment, the network structure of `TF-T2V` is basically consistent with the open source ModelScopeT2V and VideoComposer to facilitate fair comparison. Note that `TF-T2V` is a plug-and-play framework that can also be applied to other text-to-video generation and controllable video synthesis methods. For human evaluation, we randomly generate 100 videos and ask users to rate and evaluate them. The highest score for each evaluation content is 100%, the lowest score is 0%, and the final statistical average is reported.

2. Additional ablation study

Effect of joint training. In our default setting, we jointly train the spatial and temporal blocks in the video diffusion model to fully exploit the complementarity between image and video modalities. An alternative strategy is to separate the spatial and temporal modeling into sequential two stages. We conduct a comparative experiment on these two strategies in Tab. 1. The results demonstrate the rationality of joint optimization in `TF-T2V`.

Scaling trend under semi-supervised settings. In Fig. 2, we vary the number of text-free videos and explore the scaling trend of `TF-T2V` under the semi-supervised settings. From the results, we can observe that FVD (↓) gradually decreases as the number of text-free videos increases, revealing the strong scaling potential of our `TF-T2V`.

3. More experimental results

To further verify that `TF-T2V` can be extended to high-definition video generation, we leverage text-free videos to train a high-resolution text-to-video model, such as 896×512. As shown in Fig. 3, we can notice that in addition to generating 448×256 videos, our method can be easily applied to the field of high-definition video synthesis. For high-definition compositional video synthesis, we additionally synthesize 1280×640 and 1280×768 videos to demonstrate the excellent application potential of our method. The results are displayed in Fig. 4 and Fig. 5.

4. Limitations and future work

In this paper, we only doubled the training set to explore scaling trends due to computational resource constraints, leaving the scalability of larger scales (such as $10\times$ or $100\times$) unexplored. We hope that our approach can shed light on subsequent research to explore the scalability of harnessing text-free videos. The second limitation of this work is the lack of exploration of processing long videos. In the experiment, we follow mainstream techniques and sample 16 frames from each video clip to train our `TF-T2V` for fair comparisons. Investigating long video generation with text-free videos is a promising direction. In addition, we find that if the input textual prompts contain some temporal evolution descriptions, such as "from right to left", "rotation", etc., the text-free `TF-T2V` may fail and struggle to accurately synthesize the desired video. In the experiment, even though we noticed that semi-supervised `TF-T2V` helped alleviate this problem, it is still worth studying and valuable to precisely generate satisfactory videos with high standards that conform to the motion description in the given text.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[2] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *NeurIPS*, 2023. 1

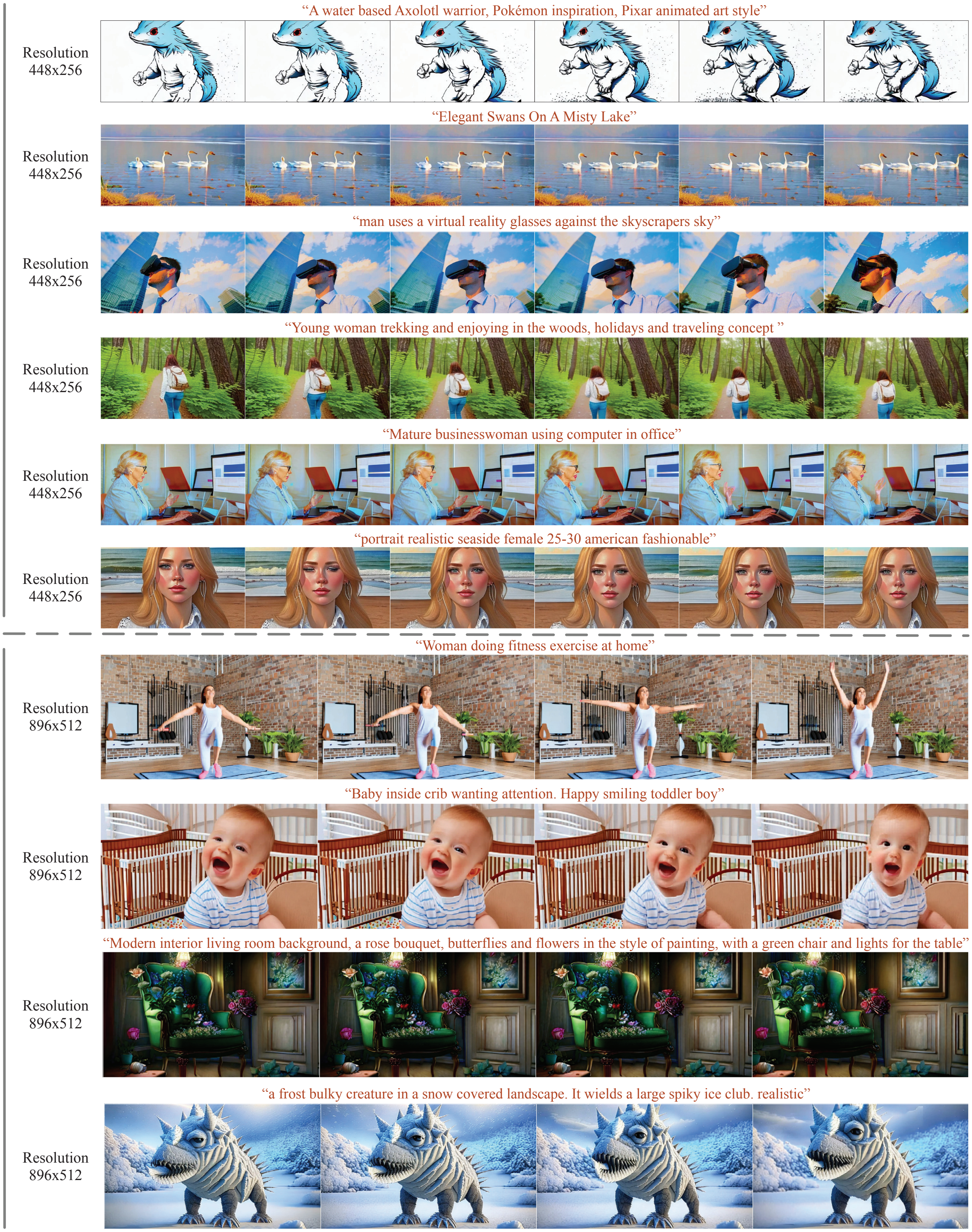
"A water based Axolotl warrior, Pokémon inspiration, Pixar animated art style"
Resolution 448x256
"Elegant Swans On A Misty Lake"
Resolution 448x256
"man uses a virtual reality glasses against the skyscrapers sky"
Resolution 448x256
"Young woman trekking and enjoying in the woods, holidays and traveling concept "
Resolution 448x256
"Mature businesswoman using computer in office"
Resolution 448x256
"portrait realistic seaside female 25-30 american fashionable"
Resolution 448x256
"Woman doing fitness exercise at home"
Resolution 896x512
"Baby inside crib wanting attention. Happy smiling toddler boy"
Resolution 896x512
"Modern interior living room background, a rose bouquet, butterflies and flowers in the style of painting, with a green chair and lights for the table"
Resolution 896x512
"a frost bulky creature in a snow covered landscape. It wields a large spiky ice club. realistic"
Resolution 896x512

Figure 3. **More video results on text-to-video generation** without training on any video-text pairs. Two resolution types of video are generated, 448 × 256 and 896 × 512 respectively.

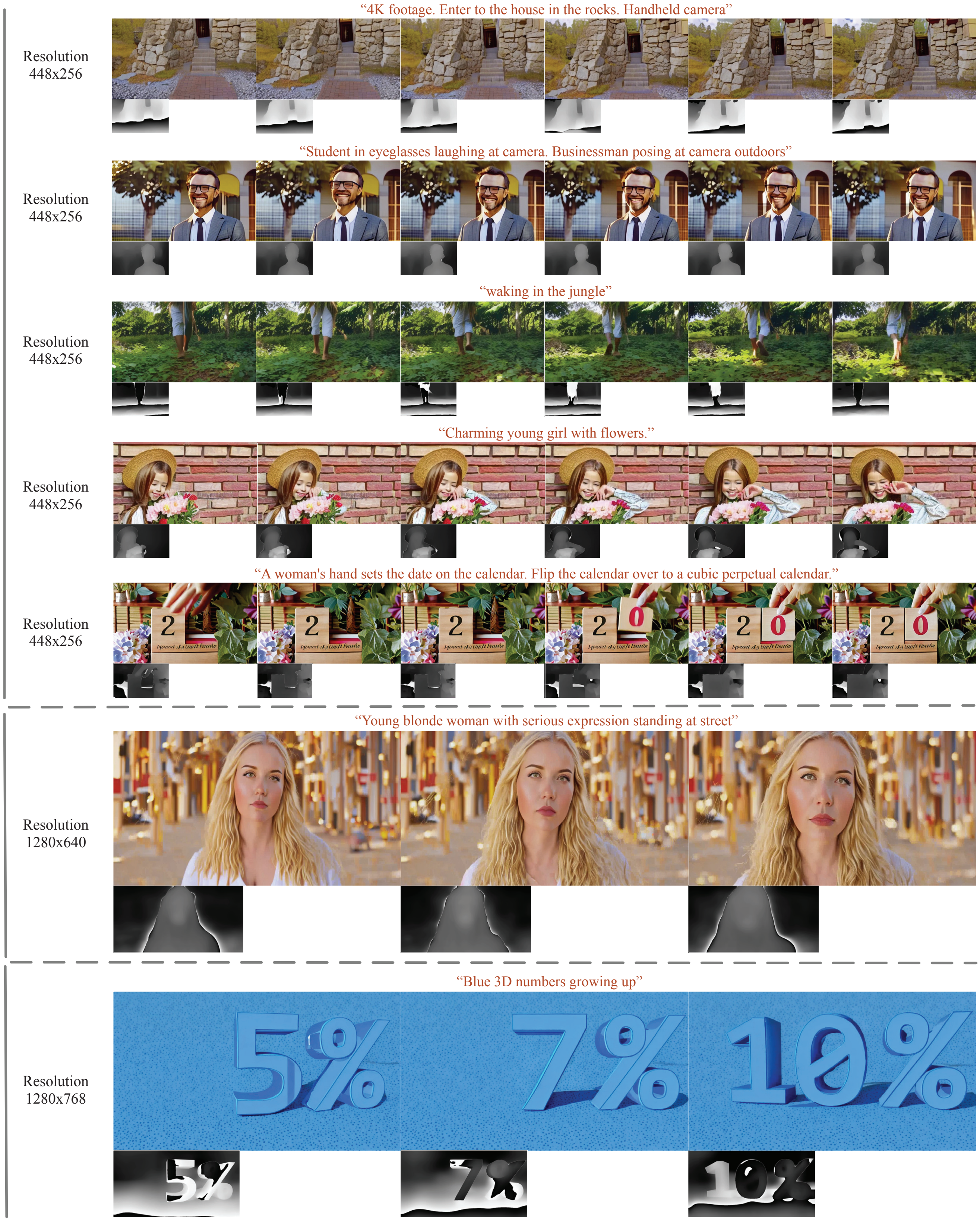
"4K footage. Enter to the house in the rocks. Handheld camera"
Resolution 448x256
"Student in eyeglasses laughing at camera. Businessman posing at camera outdoors"
Resolution 448x256
"waking in the jungle"
Resolution 448x256
"Charming young girl with flowers."
Resolution 448x256
"A woman's hand sets the date on the calendar. Flip the calendar over to a cubic perpetual calendar."
Resolution 448x256
"Young blonde woman with serious expression standing at street"
Resolution 1280x640
"Blue 3D numbers growing up"
Resolution 1280x768

Figure 4. **More video results on compositional depth-to-video synthesis** without training on any video-text pairs. Three resolution types of video are generated, 448 × 256, 1280 × 640 and 1280 × 768 respectively.

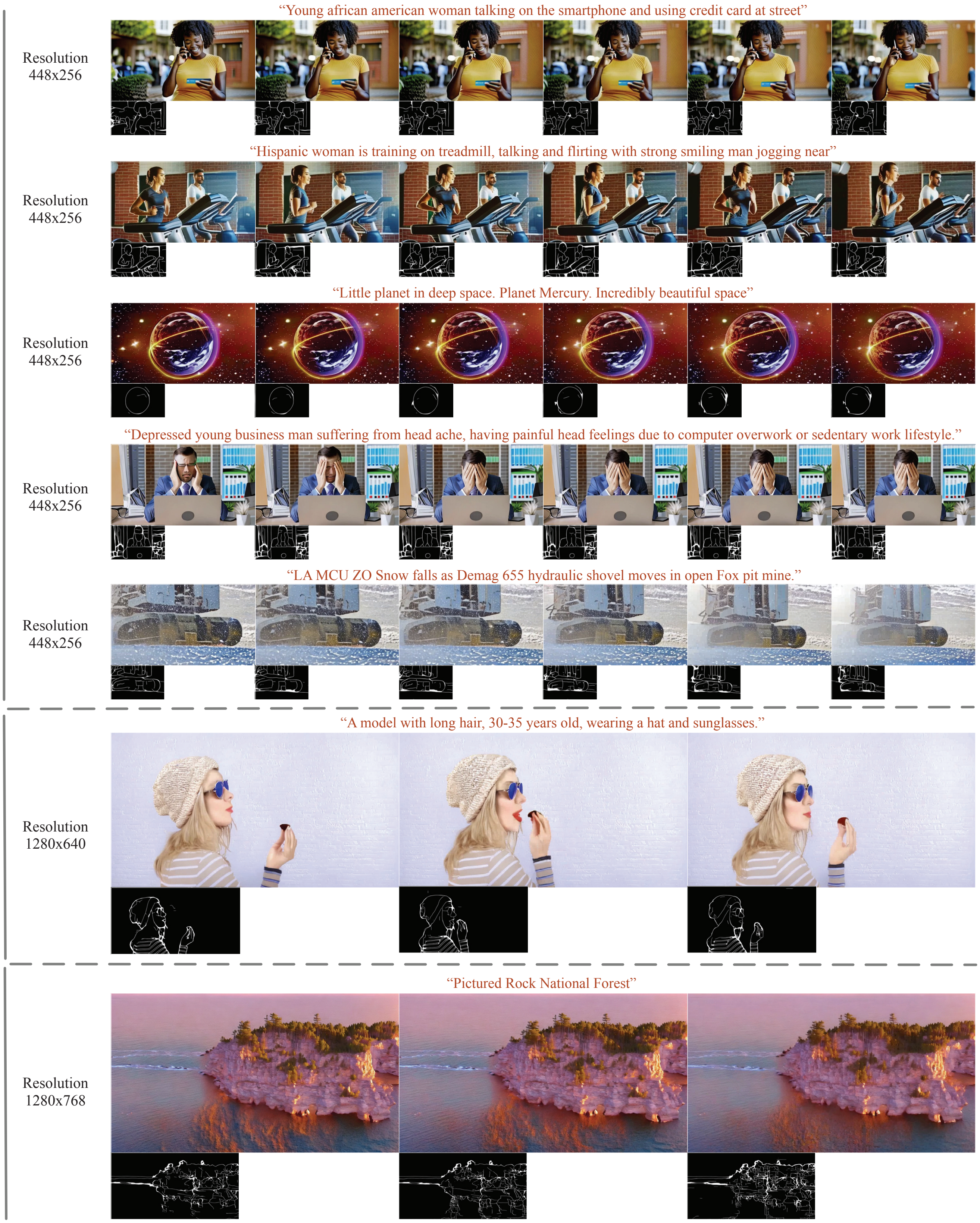
"Young african american woman talking on the smartphone and using credit card at street"
Resolution 448x256
"Hispanic woman is training on treadmill, talking and flirting with strong smiling man jogging near"
Resolution 448x256
"Little planet in deep space. Planet Mercury. Incredibly beautiful space"
Resolution 448x256
"Depressed young business man suffering from head ache, having painful head feelings due to computer overwork or sedentary work lifestyle."
Resolution 448x256
"LA MCU ZO Snow falls as Demag 655 hydraulic shovel moves in open Fox pit mine."
Resolution 448x256
"A model with long hair, 30-35 years old, wearing a hat and sunglasses."
Resolution 1280x640
"Pictured Rock National Forest"
Resolution 1280x768

Figure 5. **More video results on compositional sketch-to-video synthesis** without training on any video-text pairs. Three resolution types of video are generated, 448 × 256, 1280 × 640 and 1280 × 768 respectively.