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

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]

Code: https://github.com/alibaba-mmai-research/MoLo

[1]Huazhong University of Science and Technology

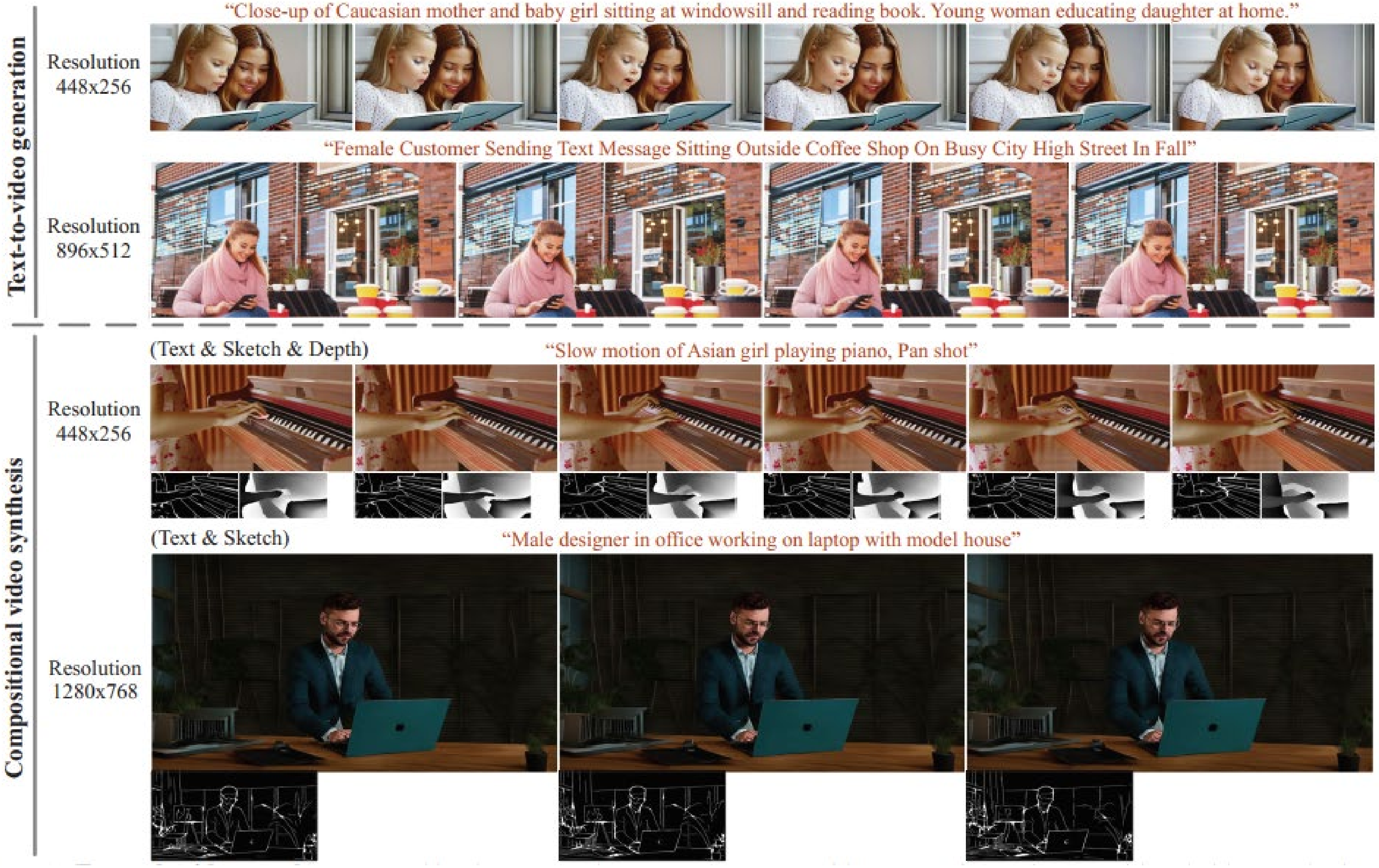
[2]Alibaba Group

[3]Zhejiang University

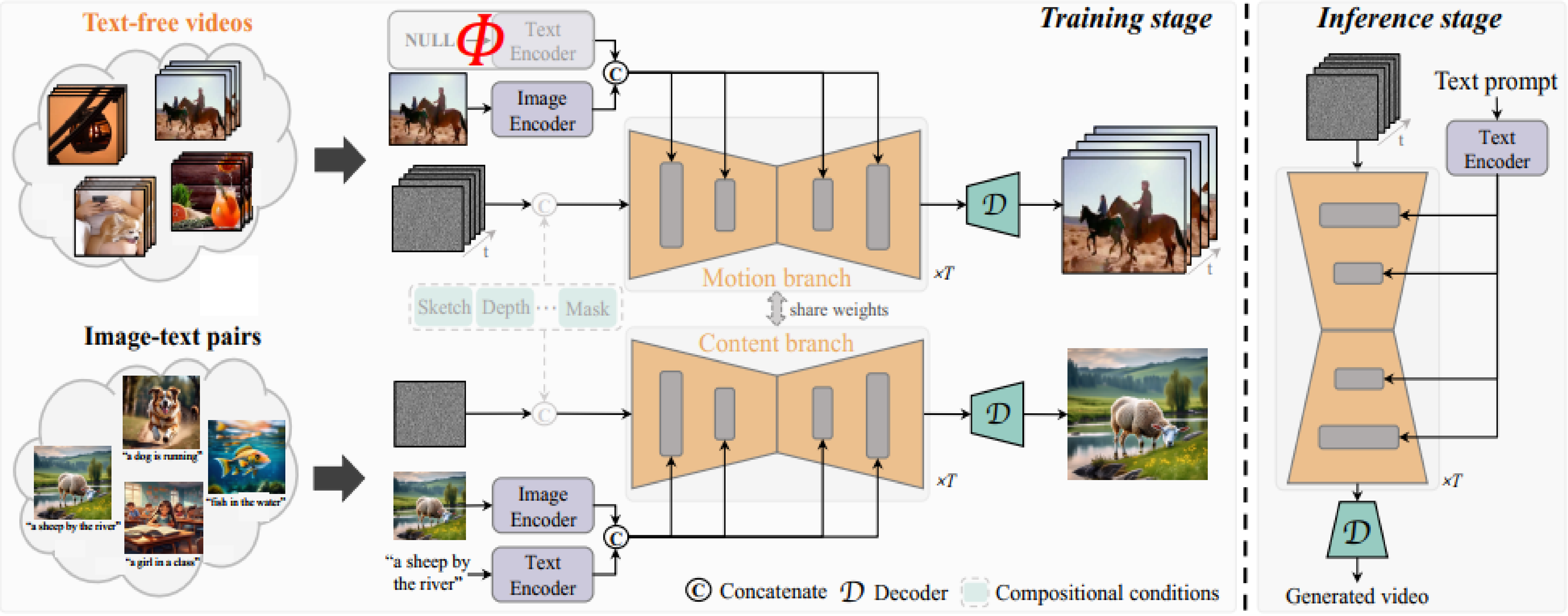
[4]Ant Group

Limitations of text-to-video generation methods

Limitations of the previous approaches:
(1) the limited scale of publicly available video-text data, considering the high cost of video captioning;
(2) the characteristics of scaling potential with text-free videos on video generation are still under-explored.


Text-to-video generation
"Close-up of Caucasian mother and baby girl sitting at windowsill and reading book. Young woman educating daughter at home."
Resolution 448x256
"Female Customer Sending Text Message Sitting Outside Coffee Shop On Busy City High Street In Fall"
Resolution 896x512
Compositional video synthesis
(Text & Sketch & Depth)
"Slow motion of Asian girl playing piano, Pan shot"
Resolution 448x256
(Text & Sketch)
"Male designer in office working on laptop with model house"
Resolution 1280x768

TF-T2V

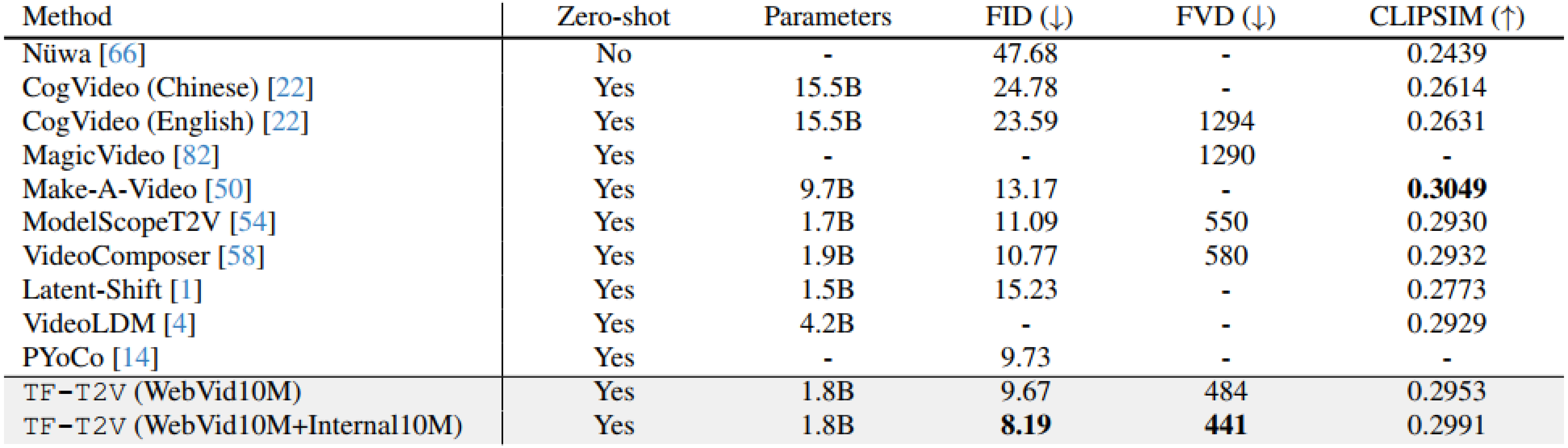


Text-Free Videos for Text-to-Video Generation (TF-T2V)

Generalization performance in different scenarios

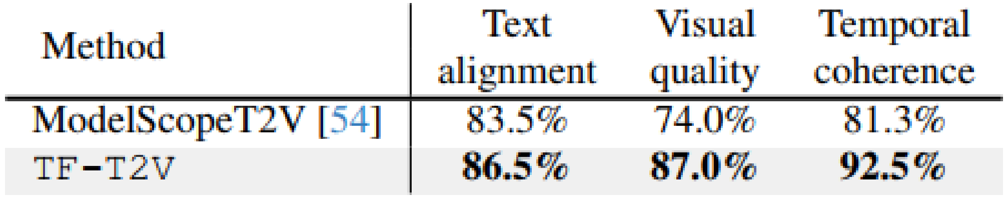
Comparison with state-of-the-art

Table 1. **Quantitative comparison** with state-of-the-art methods for text-to-video task on MSR-VTT in terms of FID, FVD, and CLIPSIM.



Method	Zero-shot	Parameters	FID (↓)	FVD (↓)	CLIPSIM (↑)
Nüwa [66]	No	-	47.68	-	0.2439
CogVideo (Chinese) [22]	Yes	15.5B	24.78	-	0.2614
CogVideo (English) [22]	Yes	15.5B	23.59	1294	0.2631
MagicVideo [82]	Yes	-	-	1290	-
Make-A-Video [50]	Yes	9.7B	13.17	-	**0.3049**
ModelScopeT2V [54]	Yes	1.7B	11.09	550	0.2930
VideoComposer [58]	Yes	1.9B	10.77	580	0.2932
Latent-Shift [1]	Yes	1.5B	15.23	-	0.2773
VideoLDM [4]	Yes	4.2B	-	-	0.2929
PYoCo [14]	Yes	-	9.73	-	-
TF-T2V (WebVid10M)	Yes	1.8B	9.67	484	0.2953
TF-T2V (WebVid10M+Internal10M)	Yes	1.8B	**8.19**	**441**	0.2991

Table 2. **Human preference results** on text-to-video generation.



Method	Text alignment	Visual quality	Temporal coherence
ModelScopeT2V [54]	83.5%	74.0%	81.3%
TF-T2V	**86.5%**	**87.0%**	**92.5%**

Generalization performance in different video scenarios

Qualitative comparison on text-to-video generation

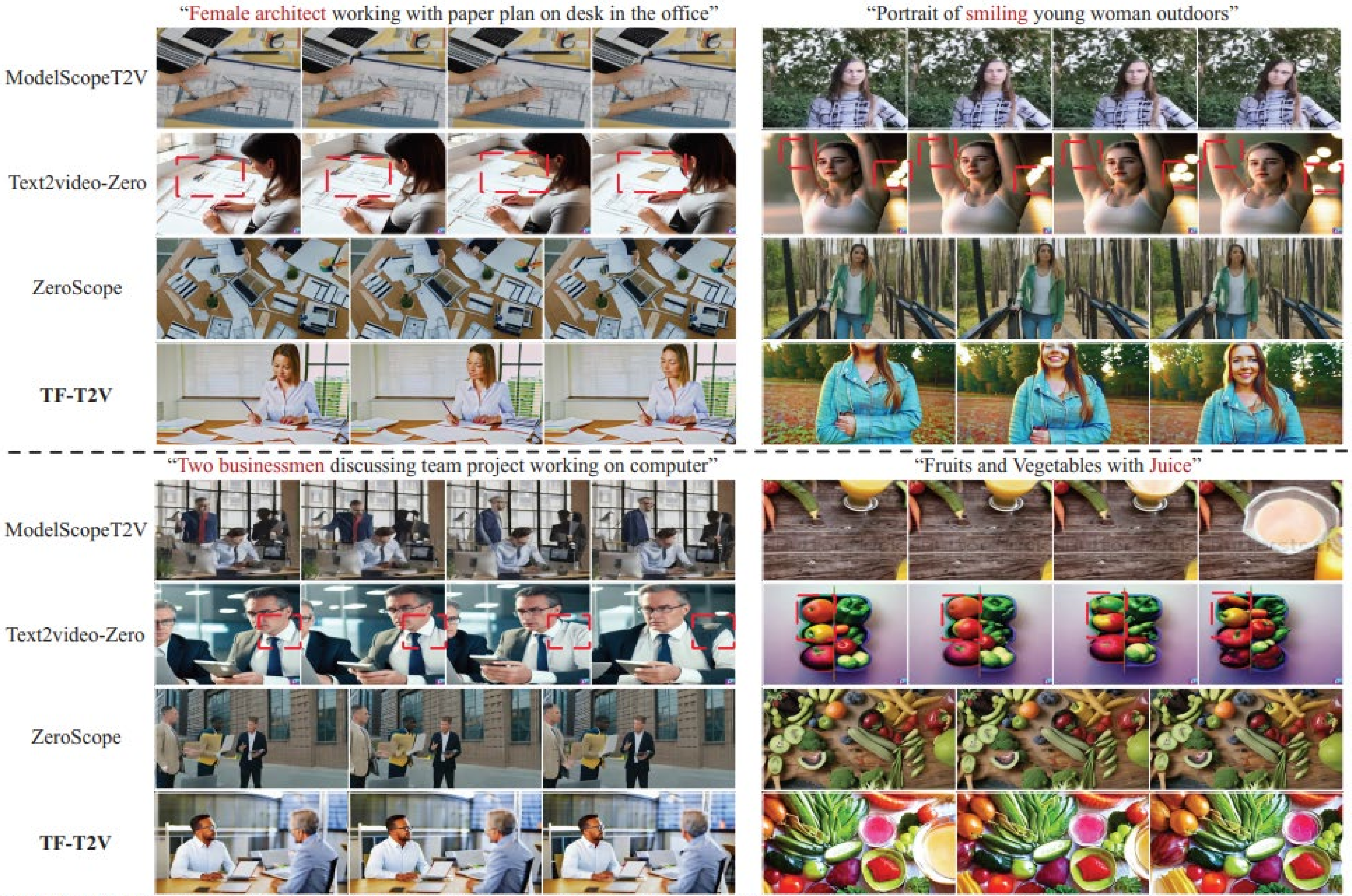
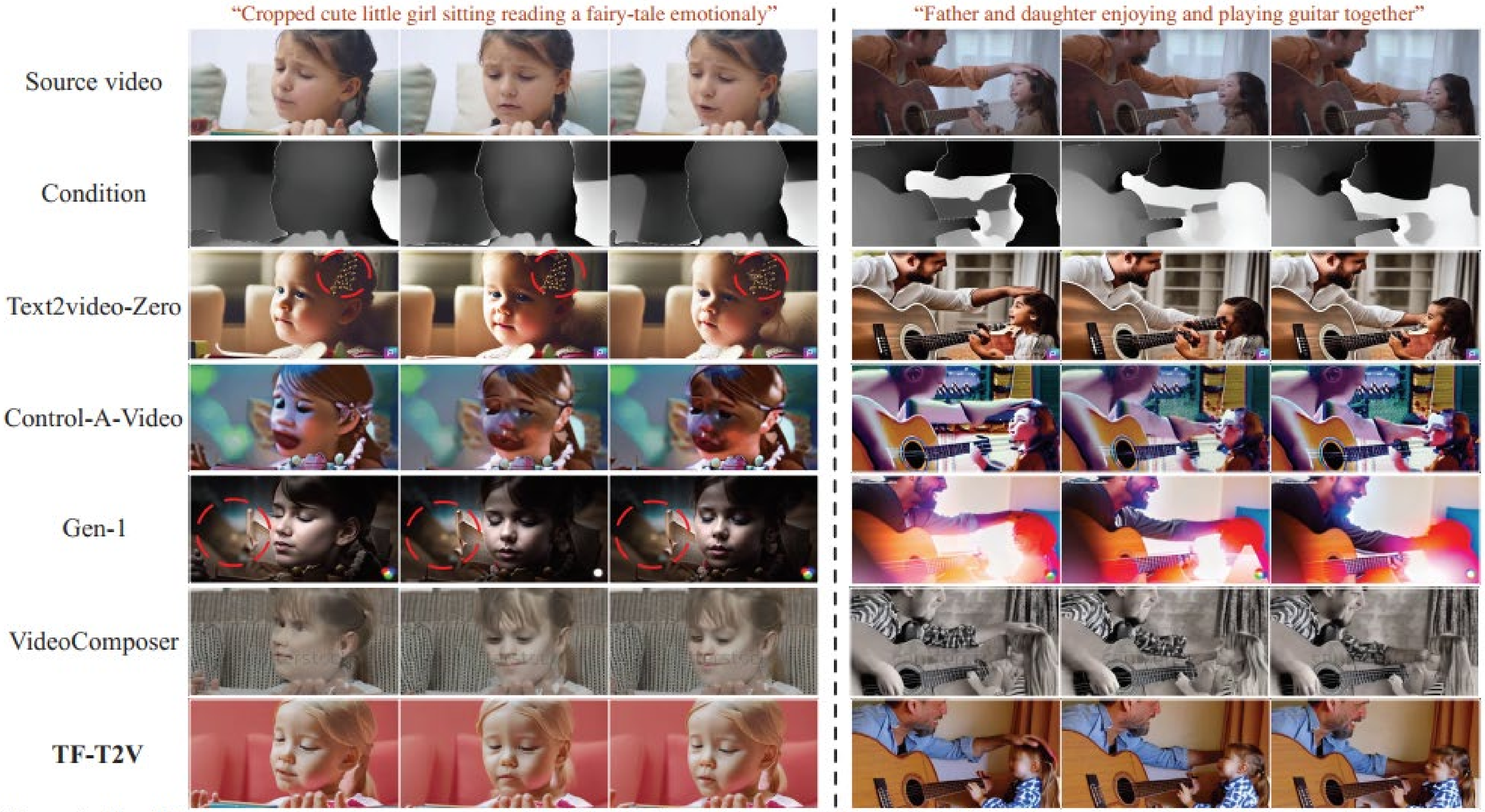

"Female architect working with paper plan on desk in the office"
"Portrait of smiling young woman outdoors"
ModelScopeT2V
Text2video-Zero
ZeroScope
TF-T2V
"Two businessmen discussing team project working on computer"
"Fruits and Vegetables with Juice"
ModelScopeT2V
Text2video-Zero
ZeroScope
TF-T2V

Figure 3. **Qualitative comparison on text-to-video generation**. Three representative open-source text-to-video approaches are compared, including ModelScopeT2V [54], Text2video-Zero [28] and ZeroScope [5]. Please refer to the Appendix for videos and more comparisons.

Generalization performance in different video scenarios

Qualitative comparison on compositional depth-to-video generation



"Cropped cute little girl sitting reading a fairy-tale emotionaly"
"Father and daughter enjoying and playing guitar together"
Source video
Condition
Text2video-Zero
Control-A-Video
Gen-1
VideoComposer
TF-T2V

Figure 4. **Qualitative comparison on compositional depth-to-video generation**. The videos are generated by taking textual prompts and structural guidance as conditions. Compared with existing methods, `TF-T2V` yields more structural compliance and high-fidelity results.

Generalization performance in different video scenarios

Qualitative comparison on compositional sketch-to-video generation

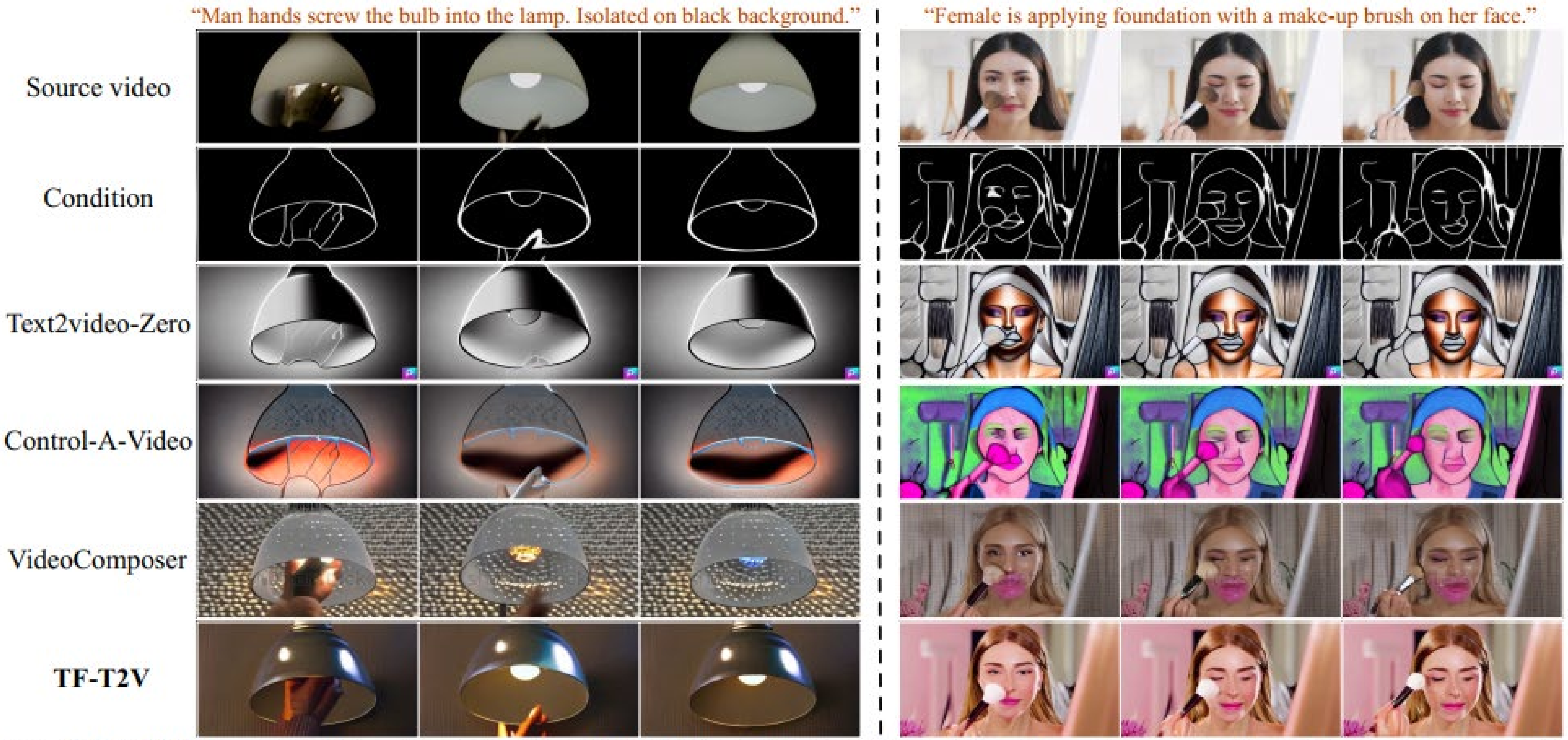

"Man hands screw the bulb into the lamp. Isolated on black background."
"Female is applying foundation with a make-up brush on her face."
Source video
Condition
Text2video-Zero
Control-A-Video
VideoComposer
TF-T2V

Figure 5. **Qualitative comparison on compositional sketch-to-video generation**. The videos are generated by taking textual descriptions and structural guidance as conditions. Compared with other methods, `TF-T2V` produces more realistic and consistent results.

Generalization performance in different video scenarios

Semi-supervised setting

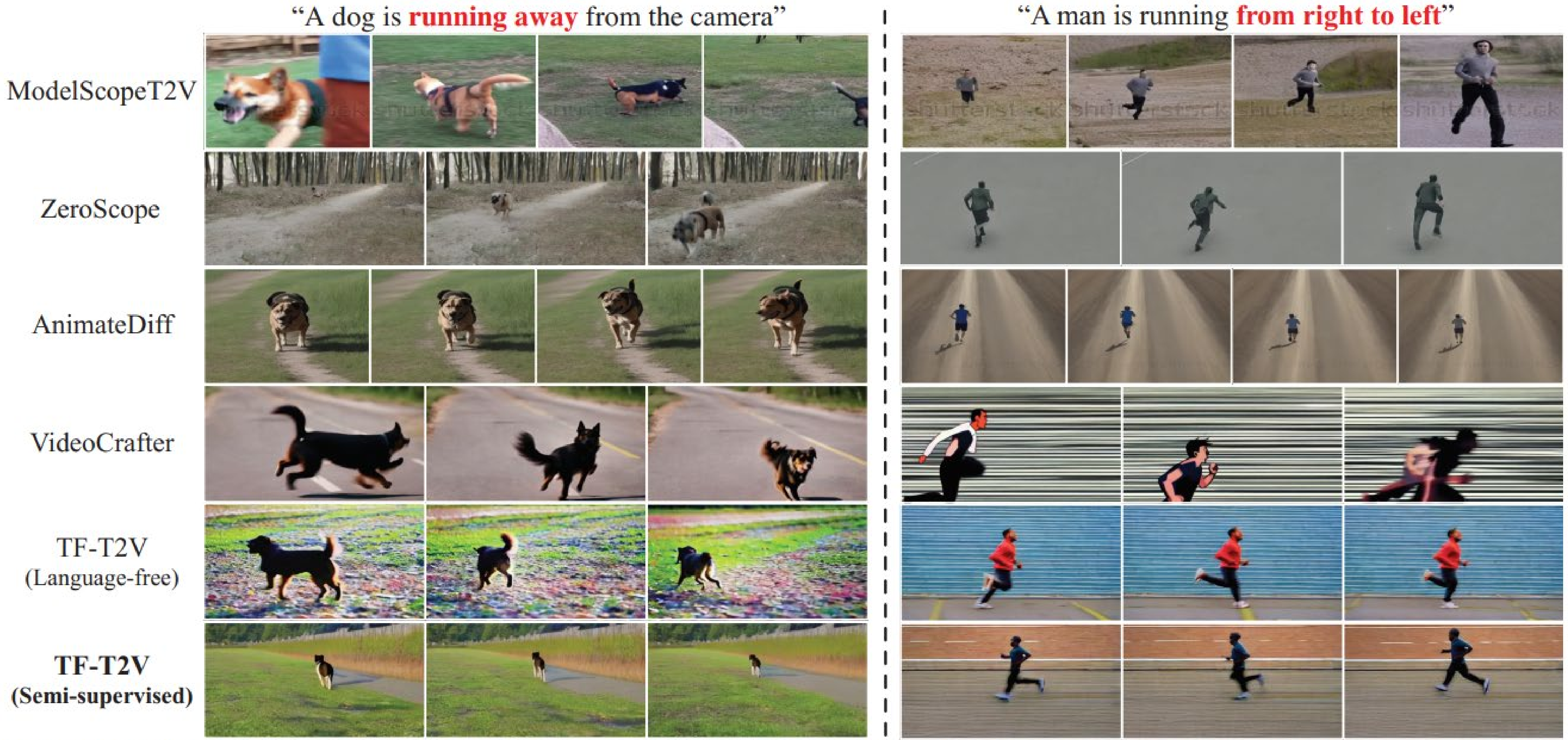


Figure 7. **Qualitative evaluation** on text-to-video generation with temporally-correlated text prompts involving the evolution of movement.

Welcome to visit to our poster!
ID: 01315

Thank you!