EvalCrafter: Benchmarking and Evaluating Large Video Generation Models Supplemental Material

Yaofang Liu[1,2,*] Xiaodong Cun[1,*] Xuebo Liu[3] Xintao Wang[1]
Yong Zhang[1] Haoxin Chen[1] Yang Liu[4†] Tieyong Zeng[4] Raymond Chan[2†] Ying Shan[1]
[1] Tencent AI Lab [2] City University of Hong Kong [3] University of Macau
[4] The Chinese University of Hong Kong
Project Page: http://evalcrafter.github.io

Appendices

Content:

- Analysis of Real-World User Data (Appendix. A).
- Quantitative Results (Appendix. B).
- Qualitative Results (Appendix. C).

A. Detailed Analysis of Real-World User Data

In this section, we present a detailed analysis of the real-world user data collected from text-to-video (T2V) generation discord users, including the FullJourney [2] and PikaLab [4]. We provide insights into the distribution of prompt lengths, important words, and meta classes.

A.1. Prompt Length Distribution

Fig. 1 (a) shows the distribution of prompt lengths in the real-world user data. We find that 90% of the prompts contain words in the range of $[3, 40]$. This observation helps us determine the appropriate length for the prompts in our benchmark.

A.2. Important Words in Prompts

Fig. 1 (b) presents a word cloud of all words in the real-world user data. From this word cloud, we can observe the most frequent words in the prompts and gain insights into the key concepts that users request in T2V generation.

A.3. Meta Classes in Prompts

Fig. 1 (c) shows the distribution of noun types in the real-world user data. We use WordNet [8] to identify the meta classes. Excluding communication, attribute, and cognition words, we find that artifacts (human-made objects), humans, animals, and locations (landscapes) play important roles in the prompts. We also include the most important word `style` from Fig. 1 (b) in the meta classes.

Based on this analysis, we divide the T2V generation into four meta-subject classes: `human`, `animal`, `object`, and `landscape`. This classification helps us create a diverse and representative benchmark for evaluating T2V models.

B. Quantitative Results

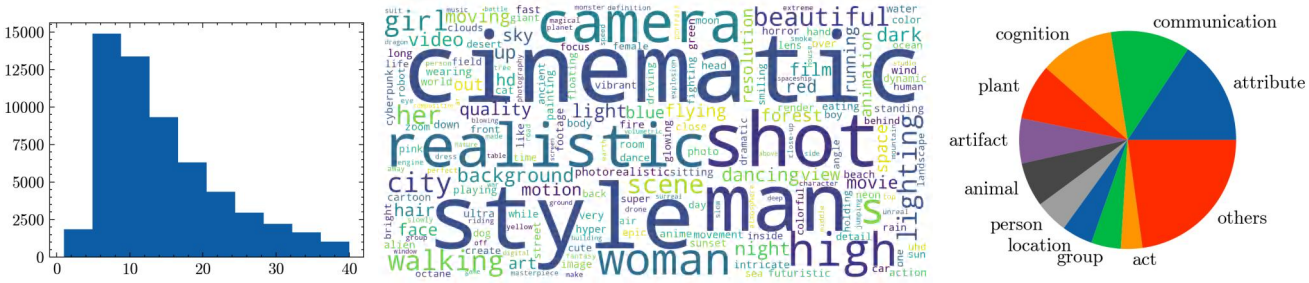
In this part, we present the quantitative results of our evaluation benchmark. We have conducted experiments on various state-of-the-art video generative models and assessed their performance using 17 objective metrics. We provide the raw results of every metric for each model and the correlations between metrics and human labels. The results are illustrated in two tables. The first table (Table 1) shows the raw results of every metric for each model. The second table (Table 2) displays the correlations between metrics and human labels.

B.1. Raw Results of Every Metric for Every Model

Table 1 shows the raw results of all 17 introduced metrics for each of the evaluated models. All metrics are expressed as percentages, except for Warping Error and Flow-Score. The table is organized as follows:

- The first column lists the metrics used for evaluation.
- The following columns display the raw results for each model, including ModelScope [11], Floor33 Pictures [1], and ZeroScope [5], Show-1 [12], Hotshot-XL [3], VideoCrafter1 [6], Gen2 [7], and PikaLab [4].
- Arrows next to the metric names indicate whether higher (↑) or lower (↓) values are better for that particular metric. For Flow-Score, the arrow is replaced with a rightwards arrow (→) as it is a neutral metric.

*Equal Contribution.
†Corresponding Author



(a) Prompt length distribution. (b) All words cloud (c) types of noun

Figure 1. The analysis of the real-world prompts from PikaLab Server [4].

Metrics	Gen2	ModelScope	Pika	Floor33	ZeroScope	VideoCrafter	Show-1	Hotshot
VQA_A ↑	59.44	40.06	59.09	58.7	34.02	66.18	23.19	71.54
VQA_T ↑	76.51	32.93	64.96	52.64	39.94	58.93	44.24	50.52
IS ↑	14.53	17.64	14.81	17.01	14.48	16.43	17.65	17.29
CLIP-Temp ↑	99.94	99.74	99.97	99.6	99.84	99.78	99.77	99.74
Warping Error ↓	0.0008	0.0162	0.0006	0.0413	0.0193	0.0295	0.0067	0.0091
Face Consistency ↑	99.06	98.94	99.62	99.08	99.33	99.48	99.32	99.48
Action-Score ↑	62.53	72.12	71.81	71.66	67.56	68.06	81.56	66.8
Motion AC-Score ↑	44.0	42.0	44.0	74.0	50.0	50.0	50.0	56.0
Flow-Score →	0.7	6.99	0.5	9.26	4.5	5.44	2.07	5.06
CLIP-Score ↑	20.53	20.36	20.46	21.02	20.2	21.33	20.66	20.33
BLIP-BLUE ↑	22.24	22.54	21.14	22.73	21.2	22.17	23.24	23.59
SD-Score ↑	68.58	67.93	68.57	68.7	67.79	68.73	68.42	67.65
Detection-Score ↑	64.05	50.01	58.99	52.44	53.94	67.67	58.63	45.7
Color-Score ↑	37.56	38.72	34.35	41.85	39.25	45.11	48.55	42.39
Count-Score ↑	53.31	44.18	51.46	58.33	41.01	58.11	44.31	49.5
OCR-Score ↓	75.0	71.32	84.31	87.48	82.58	88.04	58.97	63.66
Celebrity ID Score ↓	41.25	44.56	45.21	40.07	46.93	40.18	37.93	38.58

Table 1. Raw results of 17 introduced metrics among the aspects of video quality, text-video alignment, motion quality, and temporal consistency. All metrics are expressed as percentages, except for Warping Error and Flow-Score.

B.2. Correlations Between Metrics and Human Labels

In addition to the raw results, Table 2 presents the correlation analysis between objective metrics and human judgment on T2V generations. We use Spearman's ρ and Kendall's ϕ for correlation calculation. The table is organized into four sections, representing the four aspects of the evaluation: visual quality, motion amplitude, temporal consistency, and text-video alignment. In each section, we compare various methods with our proposed evaluation method, which is highlighted in bold.

As can be seen from the table, our method consistently achieves higher correlation values compared to the average of other methods. This shows the effectiveness of our proposed evaluation method in aligning the objective metrics to users' opinions. For instance, in the visual quality aspect, our method obtains a Spearman's ρ of 55.4 and a Kendall's ϕ of 41.1, which are both higher than the average values of 55.0 and 41.0, respectively. Similar improvements can be observed in other aspects as well.

In addition to the findings mentioned earlier, we can observe that some metrics show negative correlations with human judgment, such as Color-Score and OCR-Score in the TV Alignment aspect. This indicates that these metrics may not be reliable for evaluating the alignment between text and video content in generative models. On the other hand, metrics like Detection-Score and Count-Score exhibit relatively higher correlations with human judgment, suggesting their potential usefulness in evaluating T2V alignment.

Overall, the results in Table 2 provide a comprehensive analysis of various objective metrics and their correlations with human judgment. These results can be valuable for researchers and practitioners in the field of T2V generation to select appropriate metrics for evaluating their models and to better understand the strengths and weaknesses of different evaluation methods.

Aspects	Methods	Spearman	Kendall
Visual Quality	VQA_A	47.8	35.5
	VQA_T	53.6	39.1
	IS	9.9	4.3
	Avg.	54.9	40.9
	Ours	**55.4**	**41.1**
Motion Amplitude	Action-Score	-14.9	-10.4
	Motion AC	-22.1	-16.4
	Flow-Score	-43.3	-30.1
	Avg.	-38.2	-27.7
	Ours	**45.0**	**32.4**
Temporal Consistency	CLIP-Temp	49.7	35.7
	Warping Error	69.0	51.7
	Face Consistency	25.8	17.8
	Avg.	54.4	38.9
	Ours	**56.7**	**41.5**
TV Alignment	CLIP-Score	6.3	4.3
	BLIP-BLEU	26.7	19.0
	SD-Score	-2.8	-2.3
	Detection-Score	11.9	9.4
	Color-Score	-5.5	-3.9
	Count-Score	28.9	22.2
	OCR-Score	-8.3	-6.7
	Celebrity ID Score	-26.0	-19.8
	Avg.	31.9	**22.7**
	Ours	**32.3**	22.5

Table 2. ***Correlation Analysis.*** Whole results of correlations between objective metrics and human judgment on T2V generations. We use Spearman's ρ and Kendall's ϕ for correlation calculation.

C. Qualitative Results

In this part, we present qualitative results of the evaluated T2V models for various aspects of video generation, taking into account the findings listed in the paper. The results are visualized in Fig. 4 to Fig. 6. We discuss the performance of each model in terms of camera motion control, content generation, motion generation, style generation, and task-specific generation.

C.1. Content Generation

In Fig. 2, we present the qualitative results of T2V models and SDXL [9] for four meta types of content generation: human, object, landscape, and animal. Finding #5 shows that resolution does not correlate much with visual appeal, as demonstrated by Gen2 [7] and Hotshot-XL [3], which have small resolutions but are both competitive in visual quality. Besides, we can also find that Gen2 [7] and PikaLab [4] are more distinguishable from SDXL [9] in both video content and style compared with other methods.

C.2. Motion Generation

Fig. 3 displays the qualitative results of T2V models with respect to motion generation. According to Finding #6, larger motion amplitude does not ensure user preference. In our study, most videos that users are fond of are those with slight movements, such as those generated by PikaLab [4] and Gen2 [7].

C.3. Style Generation

The qualitative results of T2V models concerning style generation are shown in Fig. 4. We can see from the figure that most methods have the ability to generate videos with specific styles, which may be inherited from base models. However, various methods like ZeroScope [5] and ModelScope [11] are also struggling to generate high-quality and consistent styled content from prompts.

C.4. Camera Motion Control

Fig. 5 shows the qualitative results of T2V models in terms of prompts with camera motion controls. As indicated by Finding #4, all methodss cannot perform camera motion control using text prompts, which indicates all T2V models lack the understanding of camera motion.

C.5. Task-Specific Generation

Finally, Fig. 6 presents the qualitative results of T2V models and SDXL [9] in terms of different tasks, *i.e.*, face generation, object generation with color, object generation with count, text generation, and activity generation. Finding #8 indicates that many models can sometimes generate completely wrong videos, with severe noises and distortions observed in baseline models like ZeroScope [5], ModelScope [11], and Floor33 Pictures [1]. This could be viewed as a catastrophic forgetting problem, as many current T2V models are finetuned from base models like SD [10].

In conclusion, the qualitative results presented in this appendix provide valuable insights into the strengths and weaknesses of different T2V models in various aspects of video generation. As stated in Finding #10, all current models are not satisfactory enough, and T2V models still have significant room for improvement. Even the best model in our evaluation, Gen2 [7], has limitations like struggling with complex scenes, instruction following, and entity details. These results, along with our proposed evaluation framework and pipeline, enable a more exhaustive and reliable assessment of the performance of large video generation models.

References

[1] Floor33 pictures discord server. https://www.morphstudio.com/. Accessed: 2023-08-30. 1, 3

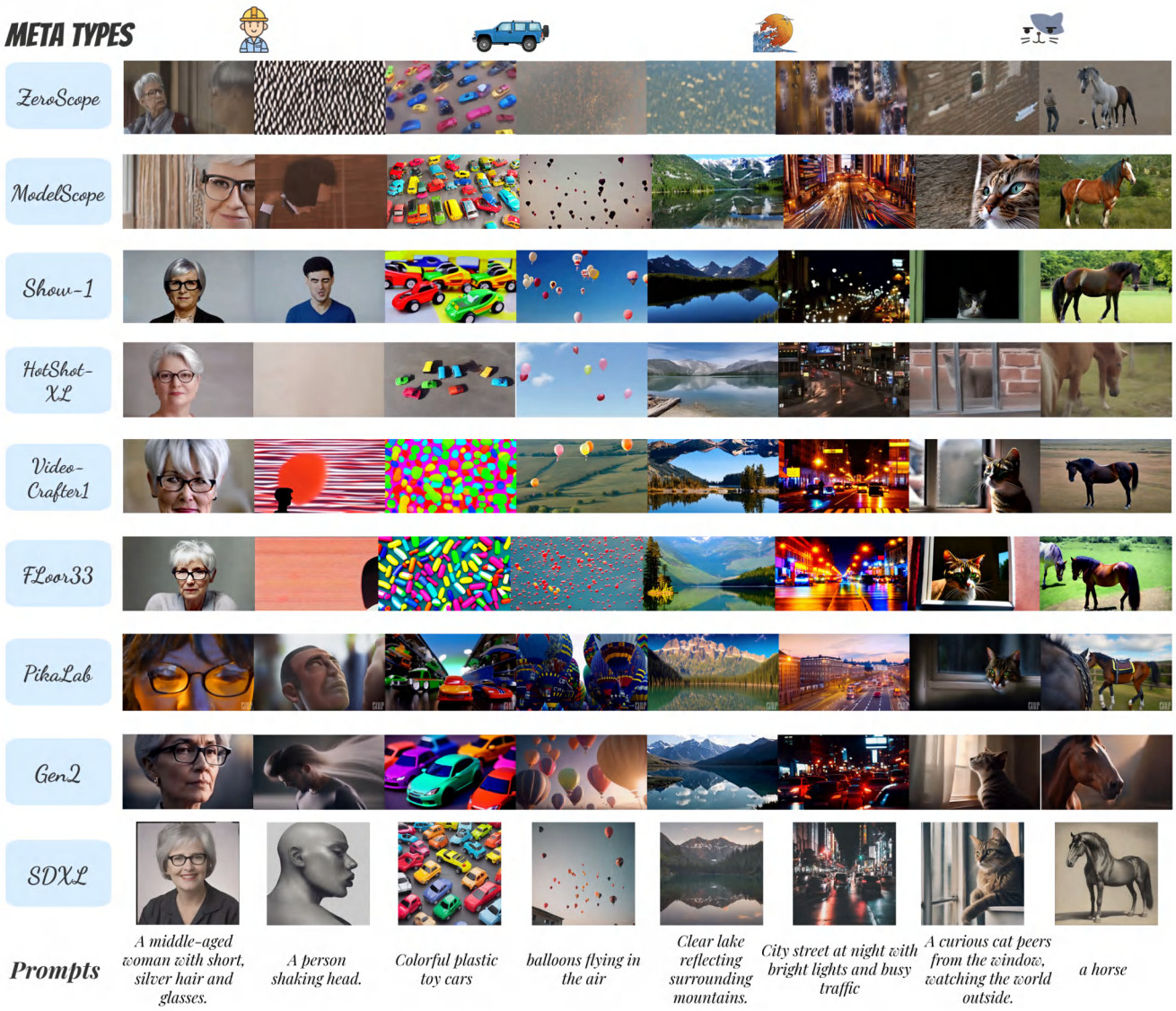


Figure 2. Qualitative results of T2V models in terms of four meta types (i.e., human, object, landscape, and animal)

[2] Fulljourney discord server. https://www.fulljourney.ai/. Accessed: 2023-08-30. 1

[3] Hotshot-xl. https://huggingface.co/hotshotco/Hotshot-XL. Accessed: 2023-10-11. 1, 3

[4] Pika Lab discord server. https://www.pika.art/. Accessed: 2023-08-30. 1, 2, 3

[5] Zeroscope. https://huggingface.co/cerspense/zeroscope_v2_576w. Accessed: 2023-08-30. 1, 3

[6] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 1

[7] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 1, 3

[8] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 1

[9] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[11] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1, 3

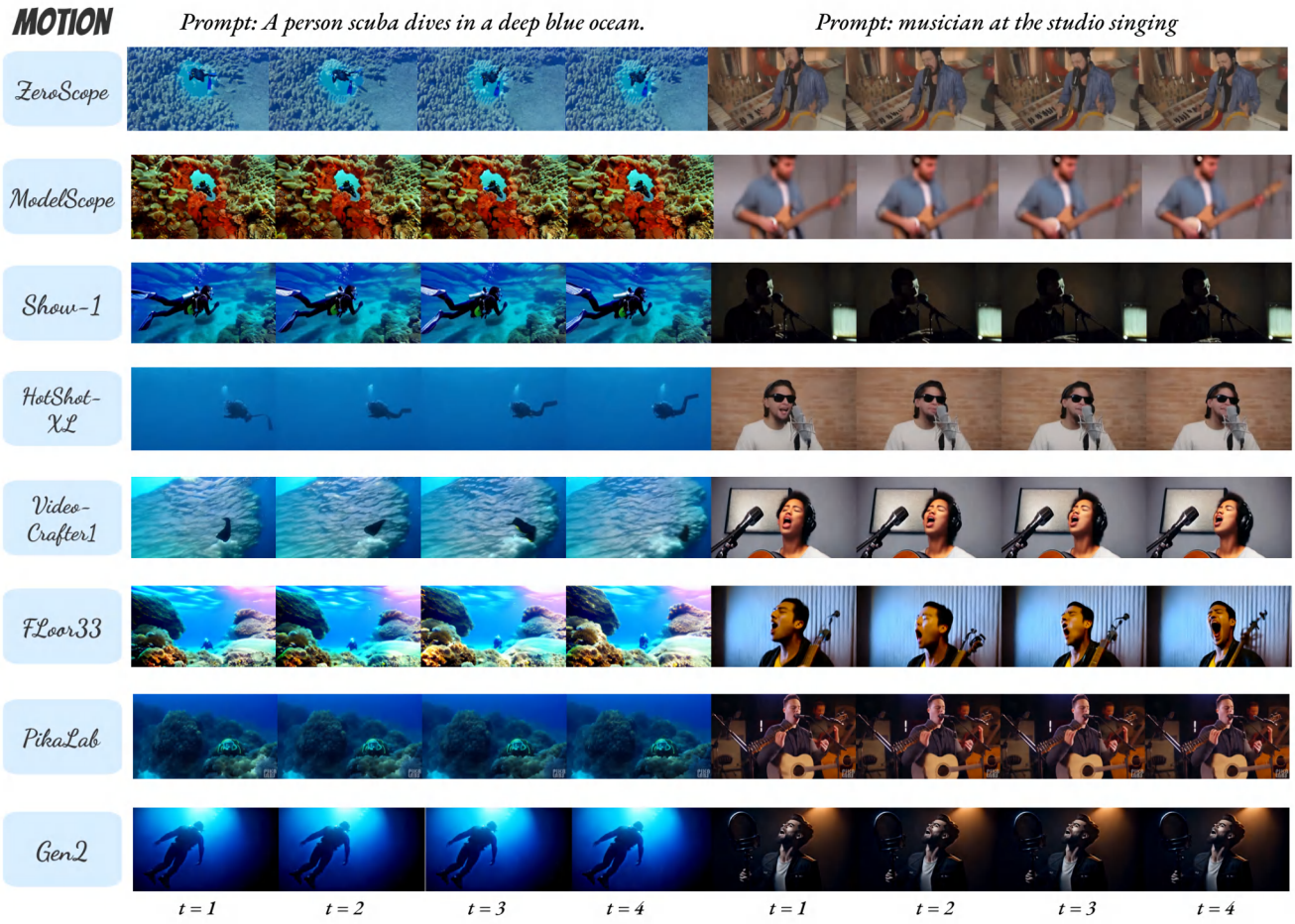


Figure 3. Qualitative results of T2V models w.r.t. motion generation

[12] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 1

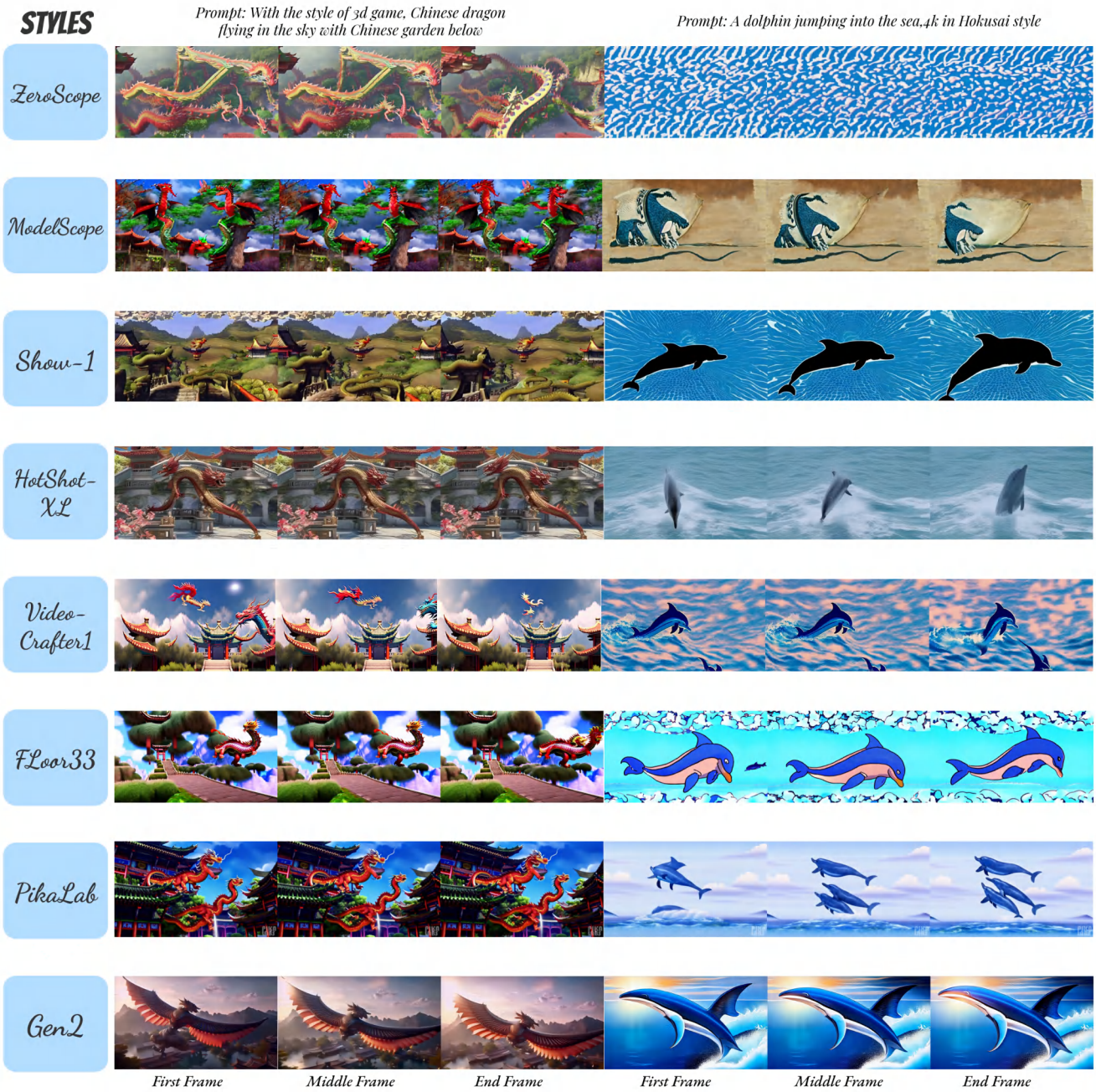


Figure 4. Qualitative results of T2V models w.r.t. style generation

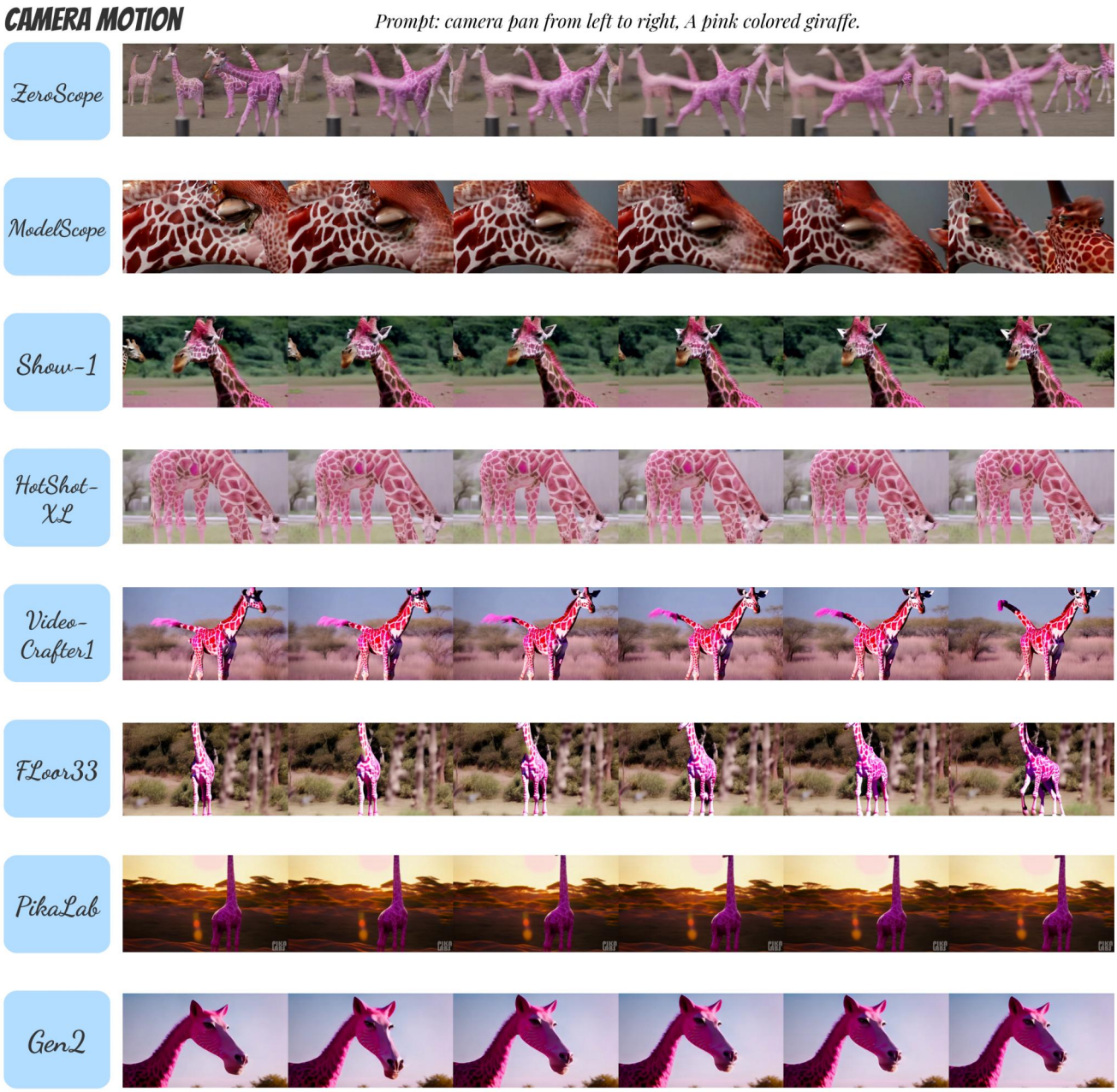


Figure 5. Qualitative results of T2V models in terms of prompts wtih camera motion controls

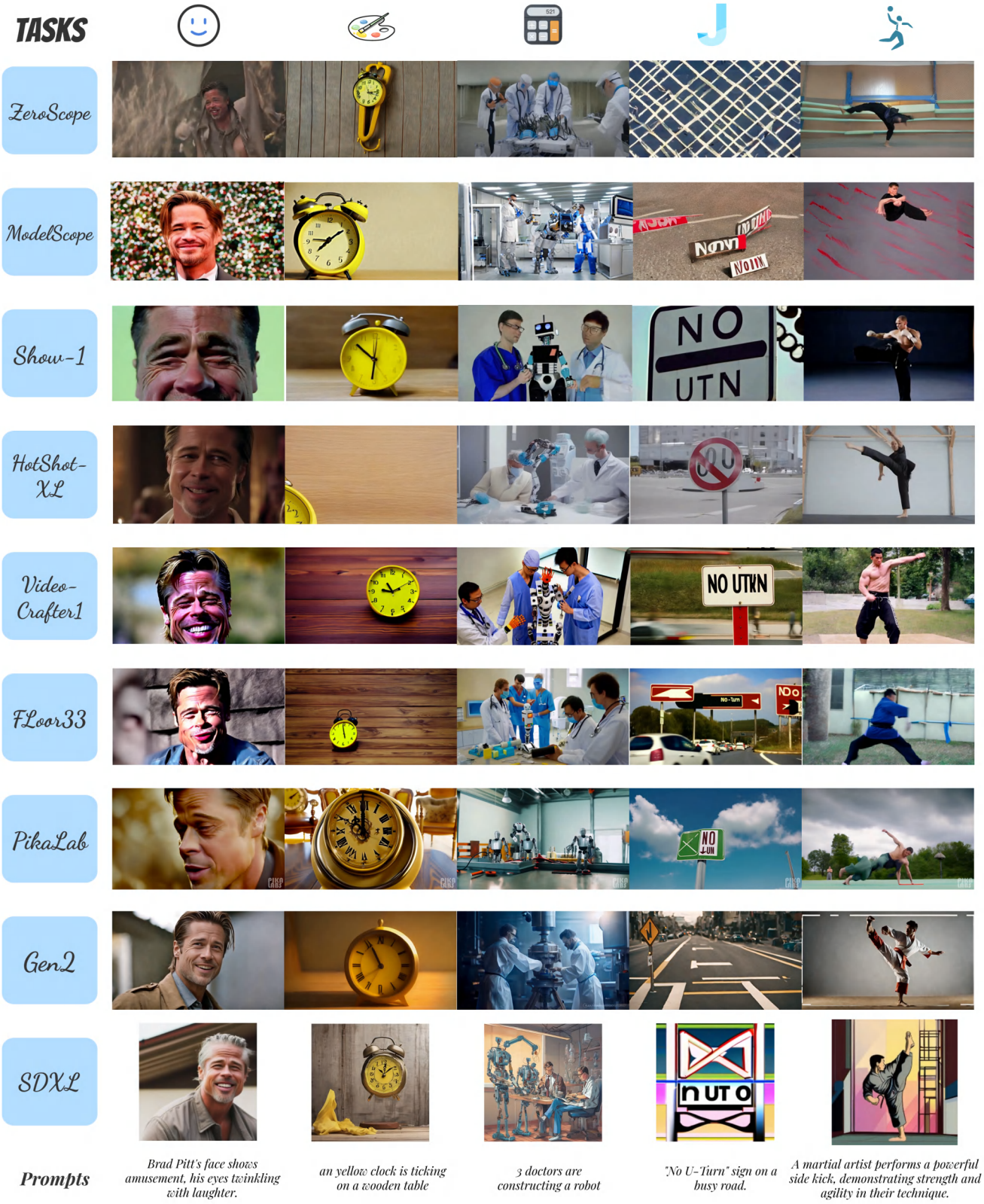


Figure 6. Qualitative results of T2V models in terms of different tasks (i.e., face generation, object generation with color, object generation with count, text generation, and activity generation)