Supplementary Materials for HiGen

Zhiwu Qing[1] Shiwei Zhang[2*] Jiayu Wang[2] Xiang Wang[1]
Yujie Wei[3] Yingya Zhang[2] 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]Fudan University
{qzw, wxiang, cgao, nsang}@hust.edu.cn
{zhangjin.zsw, wangjiayu.wjy, yingya.zyy}@alibaba-inc.com
yjwei22@m.fudan.edu.cn

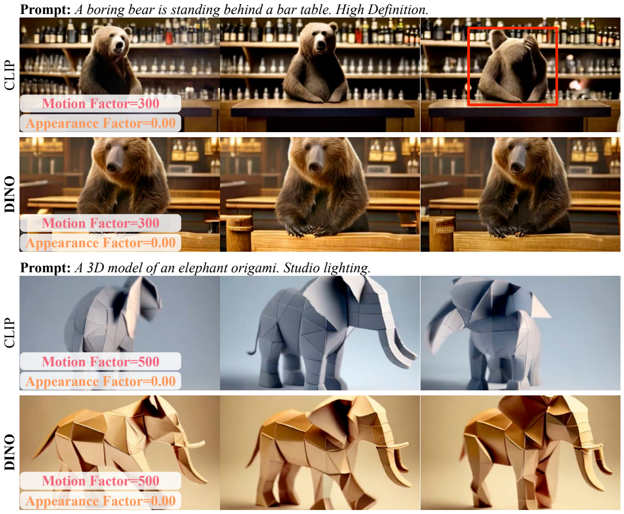


Figure 1. Visualization for different semantic models. In these four generated videos, we set the appearance factor γ^a to zero, indicating that the entire video should have minimal visual changes.

Overview

In this supplementary material, we provide experiments related to the semantic model and features for motion and appearance guidance in Sec. 1. Lastly, in Tab. 2 and Tab. 3, we provide the 69 text prompts used in our ablation experiments.

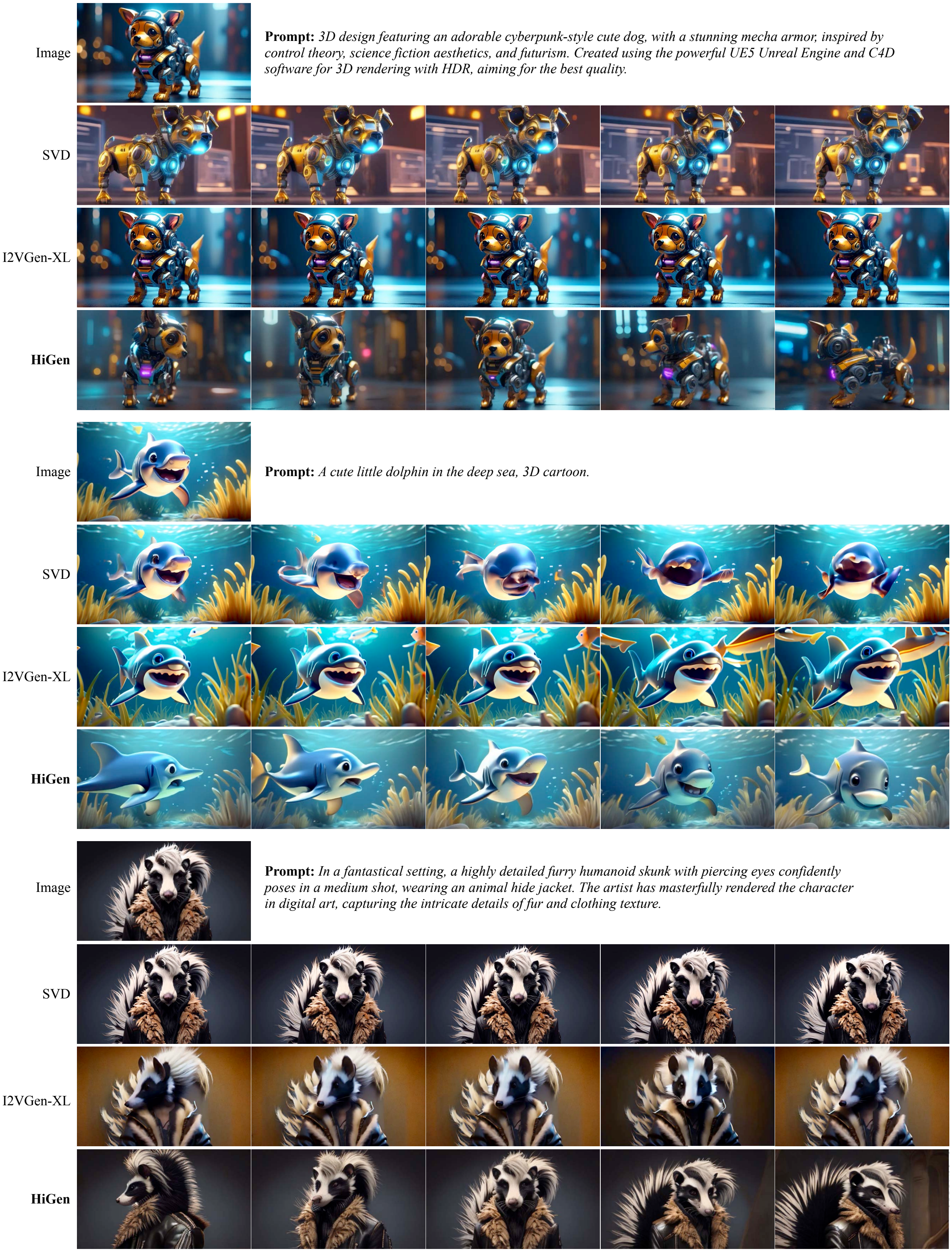
1. More Experiments

Different semantic models. In Fig. 1, we visualize the generated results when using DINO [2] and CLIP [3] as semantic models in the appearance analysis. It can be observed that videos generated using DINO as the semantic model exhibit better temporal stability, whereas CLIP may result in unexpected and irrational abrupt changes. Fig.10 in the main paper also supports this observation by demonstrating a lower correlation between appearance and motion factors obtained from DINO features, thereby enabling better independent control.

How to generate motion and appearance guidance? In Fig.3 of the main paper, we default to using a vector composed of $F-1$ frame difference elements to generate motion guidance, while a similarity matrix is used to generate appearance guidance. The reason behind this choice is that frame difference cannot capture the motion information between frames with significant visual differences, whereas a semantic model can effectively model the appearance correlation between any pair of video frames. In Tab. 1, we quantitatively analyze the combinations of motion and appearance guidance using vector-based and matrix-based methods. We conducted evaluations with three different motion factors and semantic factors for each combination, and then measured changes in terms of Temporal Consistency and CLIPSIM. It can be observed that different combinations exhibit similar spatial quality for the generated videos (*i.e.*, minimal changes in CLIPSIM), but using frame difference vectors for motion guidance and similarity matrices for appearance guidance leads to more significant temporal variations (*i.e.*, ± 0.0276).

Comparison with image-to-video approaches. In Fig. 2, we compare HiGen with state-of-the-art image-to-video generation methods. The images are generated using advanced text-to-image methods such as Midjourney. We directly incorporate these images as spatial priors in the temporal reasoning step. It can be observed that, compared to Stable Video Diffusion [1] and I2VGen-XL [4], the videos generated by HiGen exhibit significantly more vivid temporal dynamics and creativity.

[*] Corresponding authors.
Project page: https://higen-t2v.github.io/.



Prompt: *3D design featuring an adorable cyberpunk-style cute dog, with a stunning mecha armor, inspired by control theory, science fiction aesthetics, and futurism. Created using the powerful UE5 Unreal Engine and C4D software for 3D rendering with HDR, aiming for the best quality.*

Prompt: *A cute little dolphin in the deep sea, 3D cartoon.*

Prompt: *In a fantastical setting, a highly detailed furry humanoid skunk with piercing eyes confidently poses in a medium shot, wearing an animal hide jacket. The artist has masterfully rendered the character in digital art, capturing the intricate details of fur and clothing texture.*

Figure 2. Comparison with advanced image-to-video methods, *i.e.*, Stable Video Diffusion [1] and I2VGen-XL [4].

Motion Guidance	Appearance Guidance	Temporal Consistency	CLIPSIM
Vector	Vector	0.9314 ± 0.0154	0.3171 ± 0.0015
Matrix	Matrix	0.9380 ± 0.0198	0.3165 ± 0.0010
Matrix	Vector	0.9392 ± 0.0167	0.3159 ± 0.0017
Vector	Matrix	0.9367 ± **0.0276**	0.3166 ± 0.0013

Table 1. Different methods for generating motion and appearance guidance.

2. Prompts

Finally, we will provide the 69 text prompts that were used in Tables 1, 2, and Figure 8 of our main paper.

References

[1] Blattmann Andreas, Dockhorn Tim, Kulal Sumith, Mendelevitch Daniel, Kilian Maciej, Lorenz Dominik, Levi Yam, English Zion, Voleti Vikram, Letts Adam, Jampani Varun, and Rombach Robin. Stable video diffusion: Scaling latent video diffusion models to large datasets. 2023. 1, 2

[2] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1

[4] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 1, 2

1. *Close up of grapes on a rotating table. High Definition*
2. *Raw fresh beef meat fillet on a wooden plate with dill*
3. *Close up coffee being poured into a glass. Slow motion*
4. *Close-up milky liquid being poured. slow motion*
5. *A waving flag close up realistic animation seamless loop*
6. *Face of happy macho mature man looking at camera*
7. *Face of happy macho mature man smiling*
8. *A girl is looking at the camera smiling. High Definition*
9. *Woman in sunset*
10. *Young girl eye macro, shot in raw, 4k*
11. *Blue sky clouds timelapse 4k time lapse big white clouds cumulus growing cloud formation sunny weather background*
12. *Campfire at night in a snowy forest with starry sky in the background*
13. *Ocean waves hitting headland rocks pacifica california slow motion*
14. *There is a table by a window with sunlight streaming through illuminating a pile of books*
15. *Beautiful sexy lady in elegant white robe. close up fashion portrait of model indoors. beauty blonde woman*
16. *Fire burning in a forest*
17. *Wildfire in mountain of thailand (pan shot)*
18. *Fireworks*
19. *Melting pistachio ice cream dripping down the cone*
20. *A 3D model of an elephant origami. Studio lighting*
21. *Strawberry close-up on a black background swinging, slow motion. water flows down the berry*
22. *A spaceship is landing.*
23. *A giant spaceship is landing on mars in the sunset. High Definition*
24. *Drone flythrough of a tropical jungle covered in snow. High Definition*
25. *Fog at the end of the path in the summer-autumn forest. nobody present. scary scene. peaceful. quiet*
26. *Cars running on the highway at night*
27. *A man is riding a horse in sunset*
28. *close up of a clown fish swimming. 4K*
29. *a boring bear is standing behind a bar table. High Definition*
30. *Beautiful pink rose background. blooming rose flower rotation, close-up*
31. *A cat eating food out of a bowl, in style of Van Gogh*
32. *Wide shot of woman worker using welding machine on her work in site construction.*
33. *celebration of christmas*
34. *A fire is burning on a candle.*
35. *Pov driving high rural mountain country road snow winter blue sky nature environment sierra nevada usa*
36. *Lid taken off gift box with puppy inside on table top with holiday gifts.*
37. *Silhouette of retired caucasian american couple enjoying the sunrise having kayaking trip on the lake outdoors red dragon*
38. *Aerial autumn forest with a river in the mountains.*
39. *London, uk - november 16,2014:traffic. buses and cars in baker street move slowly in london england. congested city traffic*
40. *A corgi is swimming fastly*
41. *There is a table by a window with sunlight streaming through illuminating a pile of books*
42. *A glass bead falling into water with a huge splash. Sunset in the background*
43. *Aerial autumn forest with a river in the mountains*
44. *astronaut riding a horse*
45. *A clear wine glass with turquoise-colored waves inside it*
46. *A bear dancing and jumping to upbeat music, moving his whole body*
47. *A bigfoot walking in the snowstorm*
48. *An iron man surfing in the sea*
49. *Filling a glass with warm coffee*
50. *3d fluffy Lion grinned, closeup cute and adorable, long fuzzy fur, Pixar render*

Table 2. Prompts Part-I.

51. *A big palace is flying away, anime style, best quality*
52. *A teddy bear is drinking a big wine*
53. *A giant spaceship is landing on mars in the sunset. High Definition*
54. *A happy elephant wearing a big birthday hat walking under the sea, 4k*
55. *Albert Einstein washing dishes*
56. *Blue sky clouds timelapse 4k time lapse big white clouds cumulus growing cloud formation sunny weather background*
57. *drone flythrough interior of sagrada familia cathedral*
58. *Close up of grapes on a rotating table. High Definition*
59. *A stunning aerial drone footage time lapse of El Capitan in YosemiteNational Park at sunset*
60. *Aerial autumn forest with a river in the mountains*
61. *An astronaut dances in the desert*
62. *Blue sky clouds timelapse 4k time lapse big white clouds cumulus growing cloud formation sunny weather background*
63. *Beautiful pink rose background. blooming rose flower rotation, close-up*
64. *Fog at the end of the path in the summer-autumn forest. nobody present. scary scene. peaceful. quiet*
65. *A beautiful sunrise on mars, Curiosity rover. High definition,timelapse, dramatic colors*
66. *Van Gogh is smoking*
67. *A shiny golden waterfall flowing through glacier at night*
68. *A teddy bear painting a portrait*
69. *Fog at the end of the path in the summer-autumn forest. nobody present. scary scene. peaceful. quiet*

Table 3. Prompts Part-II.