Supplementary: Synthesizing Compositional Videos from Text Description

1. Additional Results and Details

As our results are based on Text-2-Video models, we provide all the details and additional video results in our project page: https://prajwalsingh.github.io/Video-ASTAR/. It can be observed that the videos generated by Video-ASTAR are consistent and coherent. The Figure 2 shows a qualitative comparison with other methods over the interaction prompt in the T2V-CompBench benchmark. Our proposed approach is able to understand the concepts from the given prompt and synthesize missing entities in the other methods. Figure 1 shows that our proposed method generates dynamic motion given the first frame mask selection strategy.

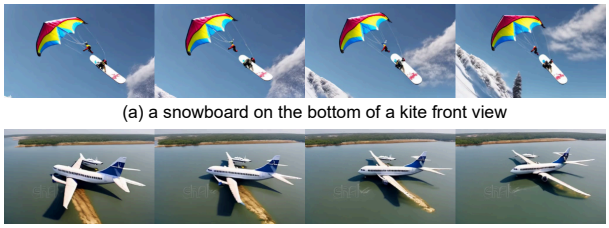

Figure 1. The figure illustrates dynamic motion on different prompts, suggesting that the motion is not restricted due to the first frame mask selection strategy. These results were already present in the project page, and have been resurfaced here for visibility.

2. Video-ASTAR vs ASTAR

ASTAR was originally proposed for compositional text-to-image generation [1], where optimization is applied on still images using attention segregation and retention losses. In contrast, Video-ASTAR extends this idea to text-to-video by addressing video-specific challenges: a) *Segregation and Retention across frames:* We modified the attention segregation and retention loss to operate consistently over video sequences, ensuring that tokens referring to different entities remain distinct throughout the entire clip, not just within a single image. b) *Attention masking:* We introduce a first-frame attention masking strategy that provides a stable reference while still allowing motion across frames. c) *Layer selection:* We determine effective layers for attention manipulation in the VideoCrafter2 U-Net (layers {9,17,18,19}), balancing spatial precision with semantic grounding. This is a non-trivial extension since the temporal dynamics of video require careful choice. d) *Additional modules:* Beyond these adaptations, we also introduce centroid loss for motion consistency and token swapping + latent interpolation to handle multi-action prompts. These design choices make Video-ASTAR specifically suited for video generation, where temporal stability and multi-action consistency are essential.

3. Intuition Behind Loss Functions

The three proposed loss terms play complementary roles in stabilizing compositional video generation.

- Attention Segregation ($\mathcal{L}_{seg}$): encourages distinct tokens (e.g., "cat" and "dog") to specialize in different spatial regions by minimizing the overlap between their attention maps. This prevents multiple tokens from competing for the same pixels.
- Attention Retention ($\mathcal{L}_{ret}$): enforces consistency of token attention across denoising steps, reducing drift and ensuring that tokens remain associated with the same regions throughout optimization.
- Centroid Loss ($\mathcal{L}_{cent}$): promotes coherence between paired tokens by aligning their spatial centroids across frames. This stabilizes interactions such as "a person holding a ball," keeping the ball near the hand while both move across time.

4. Cross Attention Layers

During test-time optimization, we extract token-level cross-attention maps from specific transformer blocks of the VideoCrafter2 U-Net. After empirical validation, we found layers {9, 17, 18, 19} strike the best balance between spatial detail and semantic grounding. Early layers tend to produce noisy, low-level attention that is not semantically reliable, while very late layers overly collapse attention to sparse regions. Using this mid-to-late layer set provides stable token localization while retaining temporal consistency.

5. Token Selection

Our proposed Video-ASTAR method works on cross-attention maps pooled from the spatial layer in the Video-

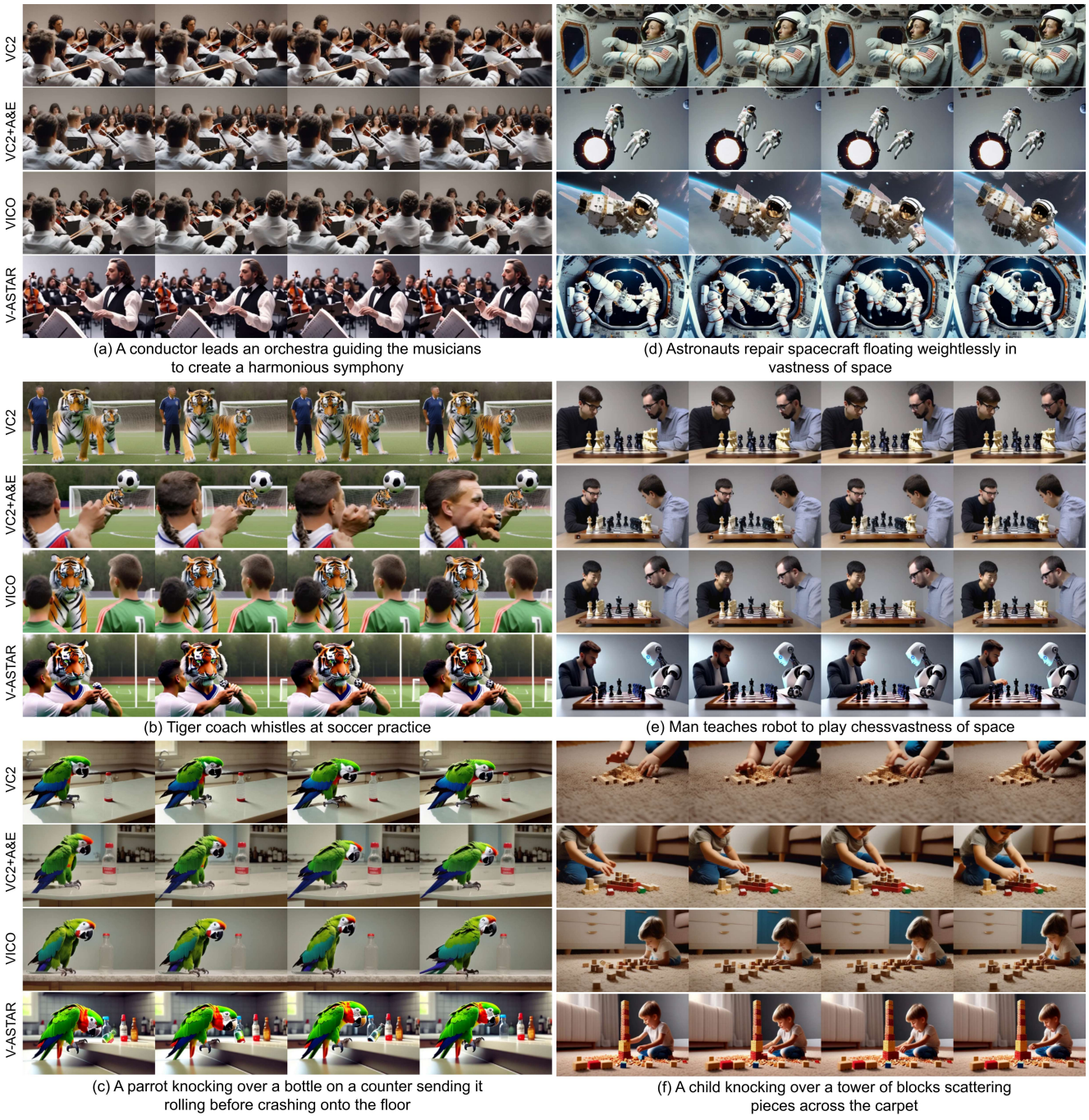

Figure 2. Qualitatitve comparison on T2V-CompBench [5] over the interaction prompts.

Carfter2 [3] T2V framework. Attention maps are sensitive to text tokens; therefore, selecting the correct token for the optimization process is an important factor for compositional video generation. The token selection method extracts 2–6 keywords from any text prompt to capture main objects, actions, relationships, or settings, using spaCy [4] for natural language processing. It starts by converting the prompt to lowercase and removing common suffixes. The prompt is then processed with spaCy to split it into words, identify parts of speech (nouns, verbs, adjectives), and detect named entities like ”Tokyo.” Unimportant words, such as articles (”a,” ”the”), prepositions (”on,” ”in”), and conjunctions (”and”), are removed to focus on meaningful content. Multi-word phrases, like ”teddy bear” or ”medieval castle,” are simplified to a single word, prioritizing the last noun (e.g., ”bear”) or the last word if no noun exists. The method checks for specific patterns: if the prompt contains ”left,” ”right,” ”top,” or ”bottom,” it selects one word be-

fore the preposition, the preposition itself, and one word after (e.g., ”bicycle on the left of a car” yields [”bicycle”, ”left”, ”car”]); if ”and” is present, it selects one word for each object (e.g., ”keyboard and cell phone” yields [”keyboard”, ”phone”]); otherwise, it selects up to 6 words, prioritizing subjects, actions, and settings (e.g., ”panda on a surfboard in the ocean at sunset” yields [”panda”, ”surfboard”, ”ocean”, ”sunset”]). The output is a list of 2–6 words within a list, ensuring at least 2 words if available or falling back to one word (e.g., ”Fireworks” yields [”fireworks”]). The method handles diverse prompts, including those without suffixes or with unusual structures, ensuring concise and robust output.

Before passing the given prompt to the above state, we pre-process it to remove the comma and add ”4k high resolution” as a suffix to all the prompts. We observe that sometimes it helps VC2 [2] to generate coherent videos.

Algorithm 1 Token Selection Method

Require: Text prompt
Ensure: List of 2–6 tokens in a list

Lowercase prompt, remove suffixes (e.g., “4k high resolution”)
Process prompt with NLP to get tokens, POS, entities
Remove stop words (e.g., “a”, “the”, “on”, “and”)
Identify phrases (e.g., “teddy bear”)
for each phrase **do**
    Pick last noun or last word
end for
Keep nouns, verbs, adjectives, entities
if prompt has “left”, “right”, “top”, or “bottom” **then**
    Output [word before, preposition, word after]
else if prompt has “and”, ≥ 2 words **then**
    Output [first word, last word]
else
    Output up to 6 words
end if
if output < 2 words **then**
    Add words to reach 2
end if
if output > 6 words **then**
    Keep first 6
end if
if no words **then**
    Use first noun/verb/adjective
end if
return [tokens]

References

[1] Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2283–2293, 2023. 1

[2] 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. 3

[3] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 2

[4] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 2

[5] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *NeurIPS*, 36: 78723–78747, 2023. 2