SteerX: Creating Any Camera-Free 3D and 4D Scenes with Geometric Steering

Supplementary Material

A. Proofs

Proposition 1. *Given the reverse generative process in (1), let q_t be the transition kernel satisfying*

$$\frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q_t(\mathbf{x}_{t-1}|\mathbf{x}_t)} = 1 + \epsilon_t(\mathbf{x}_{T:t}), \tag{11}$$

with $|\epsilon_t(\mathbf{x}_{T:t})| \leq \varepsilon$ uniformly. Also assume that the error from the reward computed at the approximate state $\hat{\mathbf{x}}_0$ is bounded, i.e. $|r_\phi(\hat{\mathbf{x}}_0) - r_\phi(\mathbf{x}_0)| \leq \eta$. Then, given the defined max potentials in (9),(10), Alg. 1 samples from

$$\tilde{p}_\theta(\mathbf{x}_0) \propto p_\theta(\mathbf{x}_0) \exp(\lambda r_\phi(\mathbf{x}_0))(1 + \mathcal{O}(T\varepsilon + \lambda\eta)) \tag{12}$$

Proof. From the conditions, the unnormalized weight assigned to a complete path is

$$W_{\mathbf{x}_{T:0}} = \prod_{t=1}^{T} [(1 + \epsilon_t(\mathbf{x}_{T:t}))G_t(\mathbf{x}_{T:t})]\, G_0(\mathbf{x}_{T:0}) \tag{13}$$

$$= \exp(\lambda r_\phi(\hat{\mathbf{x}}_0)) \prod_{t=1}^{T} (1 + \epsilon_t(\mathbf{x}_{T:t})), \tag{14}$$

where for the second equality, we used

$$G_0(\mathbf{x}_{T:0}) \prod_{t=1}^{T} G_t(\mathbf{x}_{T:t}) = \exp(\lambda r_\phi(\hat{\mathbf{x}}_0)). \tag{15}$$

Let $r_\phi(\hat{\mathbf{x}}_0) = r_\phi(\mathbf{x}_0) + \delta(\mathbf{x}_0)$. We have

$$\exp(\lambda r_\phi(\hat{\mathbf{x}}_0)) = \exp(\lambda r_\phi(\mathbf{x}_0)) \exp(\lambda\delta(\mathbf{x}_0)). \tag{16}$$

Given $|\delta(\mathbf{x}_0)| \leq \eta$, we use the Taylor expansion

$$\exp(\lambda\delta(\mathbf{x}_0)) = 1 + \mathcal{O}(\lambda\eta). \tag{17}$$

Further, we have that

$$\prod_{t=1}^{T} (1 + \epsilon_t(\mathbf{x}_{T:t})) = 1 + \mathcal{O}(T\varepsilon). \tag{18}$$

Combining (14),(17), and (18), the full weight reads

$$W(\mathbf{x}_{T:0}) = \exp(\lambda r_\phi(\mathbf{x}_0))(1 + \mathcal{O}(T\varepsilon + \lambda\eta)). \tag{19}$$

Integrating out the latent variables $\mathbf{x}_{T:1}$, the proof is complete. □

Algorithm 2 SteerX (rectified flow)

Required: rectified flow model $\mathbf{v}_\theta$, reward function r_ϕ, number of particles k, and initial noise $\{\mathbf{x}^j_{t_N}\}_{j=1}^k \sim \mathcal{N}(0, I)$.

Sampling:

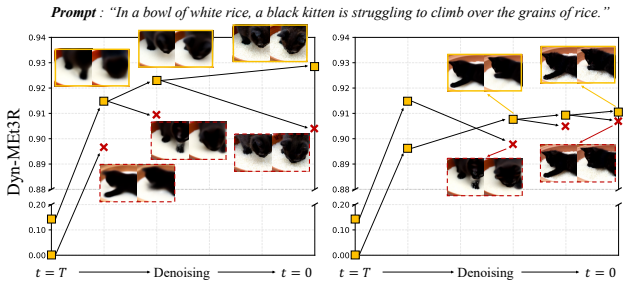
for $i \in \{N-1, \ldots, 0\}$ **do**
  for $j \in \{1 \ldots k\}$ **do**
    $\hat{\mathbf{x}}^j_{t_0} \leftarrow \mathbf{x}^j_{t_{i+1}} - t_{i+1}\mathbf{v}_\theta(\mathbf{x}^j_{t_{i+1}})$
    $\mathbf{s}^j_{t_i} \leftarrow r_\phi(\hat{\mathbf{x}}^j_{t_0})$ ▷ *Intermediate rewards*
    $G^j_{t_i} \leftarrow \exp(\lambda \max_{l=t_i}^{t_N}(\mathbf{s}^j_l))$ ▷ *Potential*
  end for
  $\{\hat{\mathbf{x}}^j_{t_0}\}_{j=1}^k \sim \text{Multinomial}(\{\hat{\mathbf{x}}^j_{t_0}, G^j_{t_i}\}_{j=1}^k)$
  $\boldsymbol{z} \sim \mathcal{N}(0, I)$
  $\mathbf{x}^j_{t_i} \leftarrow (1 - t_i)\{\hat{\mathbf{x}}^j_{t_0}\}_{j=1}^k + t_i\boldsymbol{z}$
end for
$l \leftarrow \arg\max_{i \in \{1,\ldots,k\}} r_\phi(\mathbf{x}^i_{t_0})$
return $\mathbf{x}^l_{t_0}$

B. Geometric steering on rectified flow models

Rectified flow-based video generative models [24, 35, 62] follow a straight Ordinary Differential Equation path, making it challenging to apply geometric steering since resampling particles does not introduce diverse sampling trajectories. Therefore, to introduce a stochastic process into the generation process, we provide additional modifications to adapt geometric steering for rectified flow models, as shown in Algorithm 2. The process of computing intermediate rewards and potentials remains the same as before. However, instead of resampling new particles from the existing particles, we resample the expected $\hat{x}_{t_0}$ from the multinomial distribution. Then, project the resampled particles onto a valid manifold at each noise level. This approach effectively enables geometric steering in rectified flow models and ensures that the model explores diverse trajectories.

C. Additional Results

We present additional experiments and results to further validate the scalability and effectiveness of SteerX. In Section C.2, we explore how increasing the number of particles or extending video length impacts geometric steering, providing insights into the scaling properties of SteerX. We show qualitative comparisons for Text-to-4D generation in Section C.3, and additional qualitative results for both Text-to-4D and Image-to-3D scene generation in Section C.4.



(a) Early Resampling (b) Late Resampling

Figure 9. **Resampling analysis** for $k = 2, M = 2$ in Text-to-4D.

Method	k	Aesthetic↑	Temporal↑	Dynamic↑	Dyn-MEt3R↑
Mochi [62]	1	0.491	0.243	-	0.884
+ **SteerX**	4	0.500	0.248	-	0.929
+ **SteerX**	8	**0.526**	**0.251**	-	**0.945**
HunyuanVideo [35]	1	0.549	0.241	-	0.911
+ **SteerX**	4	0.555	0.243	-	0.964
+ **SteerX**	8	**0.570**	**0.244**	-	**0.979**
CogVideoX [80]	1	0.592	-	0.158	0.880
+ **SteerX**	4	0.596	-	0.170	0.909
+ **SteerX**	8	**0.600**	-	**0.172**	**0.930**

Table 6. **Ablation study on the number of particles.**

Method	k	N	Temporal↑	Dyn-MEt3R↑
HunyuanVideo [35]	1	25	0.241	0.911
HunyuanVideo [35]	1	49	0.245	0.940
+ BoN	4	25	0.239	0.931
+ BoN	4	49	0.246	0.948
+ **SteerX**	4	25	0.243	0.964
+ **SteerX**	4	49	**0.248**	**0.978**

Table 7. **Ablation study on the number of frames.**

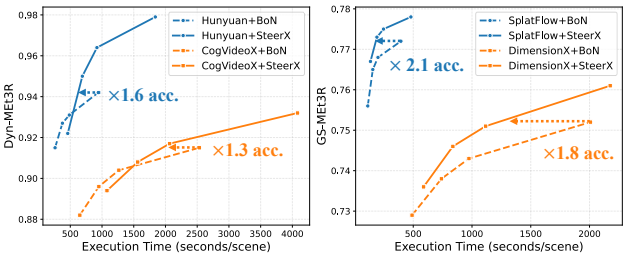


Figure 10. **Scalability analysis** with $k = 2, 3, 4, 8$. We use 100 randomly selected samples in VBench-I2V for Image-to-3D/4D.

C.1. Analysis on design choices

Linear resampling places resampling steps at uniform intervals across the entire timestep T. Also, as shown in Fig. 9, the generative model tends to form a coarse geometric structure around $0.8T$ and focuses on fine details after $0.6T$. Based on this observation, early and late resampling are uniformly scheduled between $0.8T - 0.6T$ and $0.4T - 0.2T$, respectively. Early resampling allows the model to build upon the coarse structure, refine local geometry, and gradually incorporate fine details by exploring diverse generation trajectories. In contrast, reward values tend to plateau in the later steps, indicating limited exploration at late resampling.

C.2. Scalability of SteerX

We further explore the scaling property of SteerX by increasing the number of particles k and video length N. Figure 10 presents the execution time versus reward values for all generation tasks as the number of particles increases. Although SteerX incurs additional computational overhead by forwarding the scene reconstruction model multiple times, it demonstrates better inference-time scalability than BoN. Also, as the number of particles increases, SteerX achieves greater performance gains by exploring more diverse sampling trajectories, rather than relying on post-hoc selection. Table 6 presents quantitative results on the performance of 4D scene generation as the number of particles increases. We observe that Dyn-MEt3R remains highly correlated with other evaluation metrics, further demonstrating the robustness of SteerX's scalability. Also, Fig. 12 and Table 7 show the impact of extending video length on Text-to-4D scene generation. We observe that as video length increases, the generated videos become more dynamic and tend to be more object-centric. Compared to the best-of-N approach, SteerX generates more visually plausible and dynamic objects, effectively capturing camera motion.

C.3. Additional comparisons in Text-to-4D

We further present qualitative comparisons to demonstrate the effectiveness of SteerX in following the given camera descriptions, as shown in Figure 16. SteerX successfully aligns with both the specified camera trajectories and object motions, resulting in highly natural 4D scenes.

C.4. Additional qualitative results

As shown in Figures 13 to 15, we provide additional qualitative results for Text-to-4D and Image-to-3D scene generation, demonstrating SteerX's ability to generate diverse 3D and 4D scenes only from images or text prompts. We also provide video results in Fig. 11.

D. Limitations and Discussions

While SteerX effectively enhances both visual quality and geometric alignment in 3D and 4D scene generation, it has certain limitations that present opportunities for future improvements. First, SteerX currently relies on existing feed-forward scene reconstruction models, meaning it cannot directly reconstruct 4D Gaussian Splats (4DGS). Second, video generative models for 4D scene generation struggle to produce video frames with large inter-frame camera motion, limiting the overall scene scale. Future advancements in video generation models that better handle broad camera motion ranges will further enhance SteerX's effectiveness in large-scale 4D scene generation.


A woman in a dress lost her balance and fell on the steps, with the camera tilting down.

Filmed from a first-person perspective, the camera passes through the graffiti alley in Melbourne, Australia, where the graffiti walls are covered with artwork from many artists.

Figure 11. **4D demo.** Please click each example in Acrobat Reader.

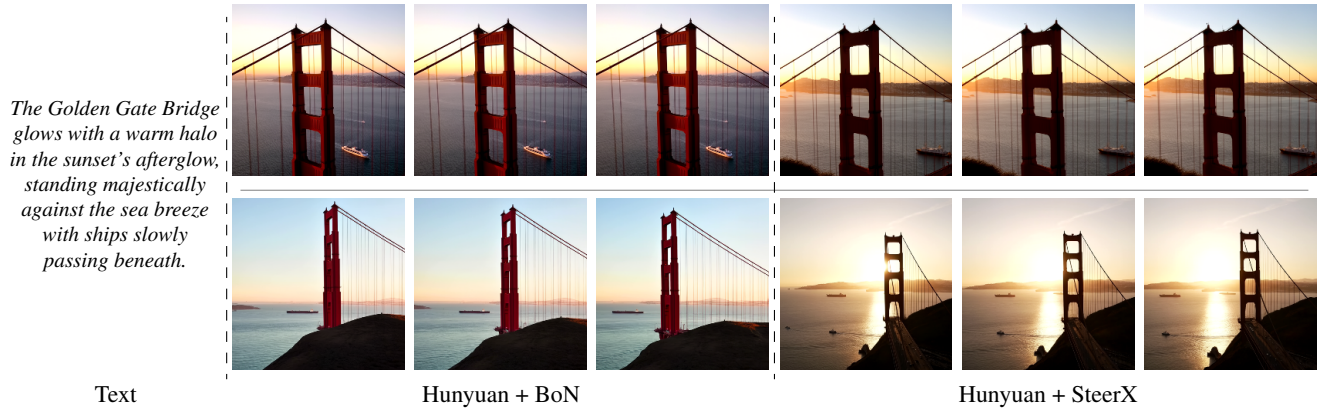
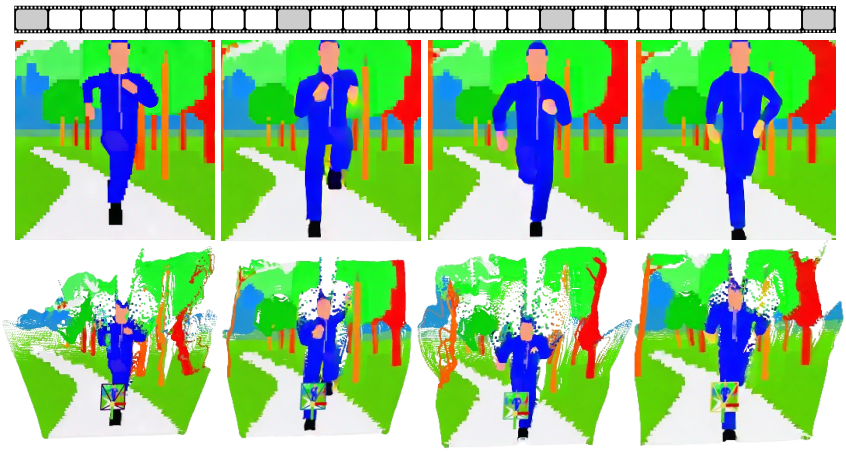


Figure 12. **Qualtitave ablation on video length.** We use four particles and visualize frames with $N = 25$ (top) and $N = 49$ (bottom).

Figure 13. **Additional qualitative results in Image-to-3D.**

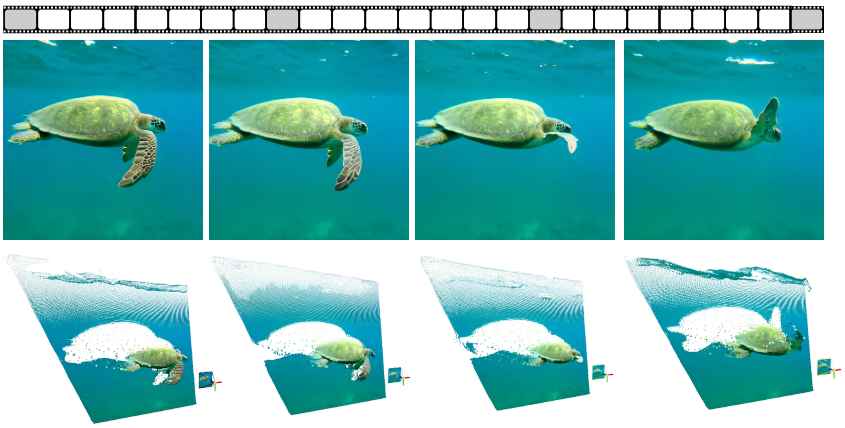


"Create a pixel art video of a man running on a park trail in a blue tracksuit and black running shoes."

"A red sports car is drifting quickly around the corner."

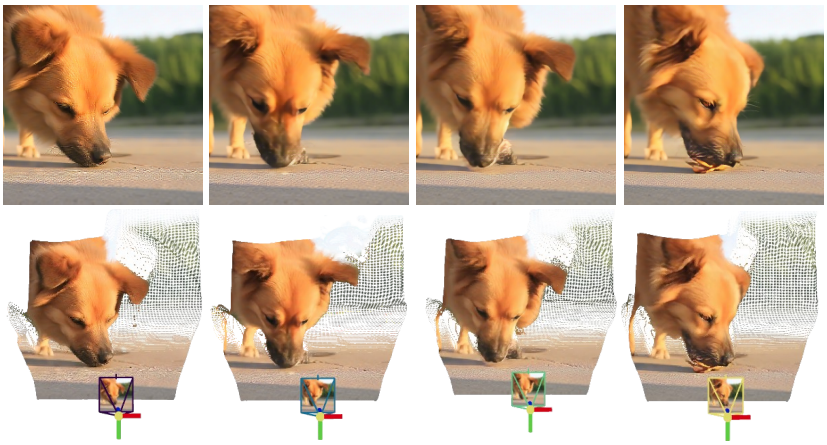
"In the bustling square at night, a woman around 60 years old is happily dancing the square dance. ..."

Figure 14. **Additional qualitative results in Text-to-4D.**



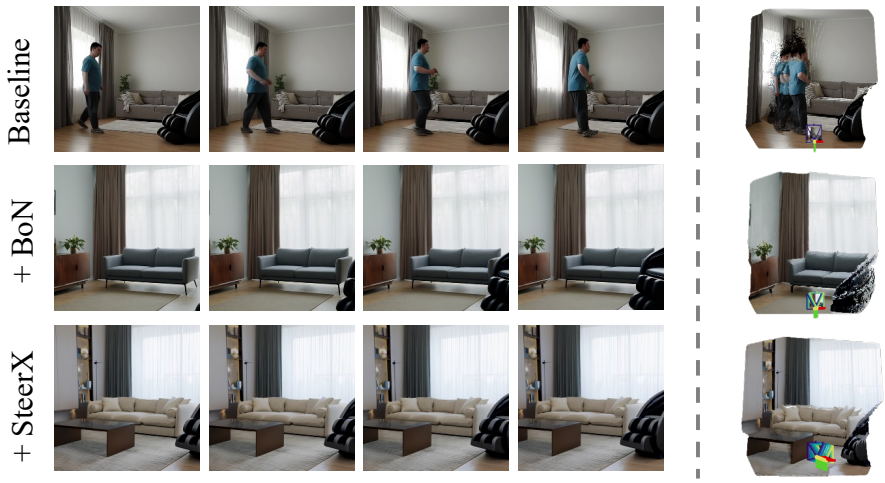
"In the ocean, a large sea turtle covered with green algae on its shell swims in the sea. ..."

"In the park, a little girl in a pink dress is on the swing, in a full shot."



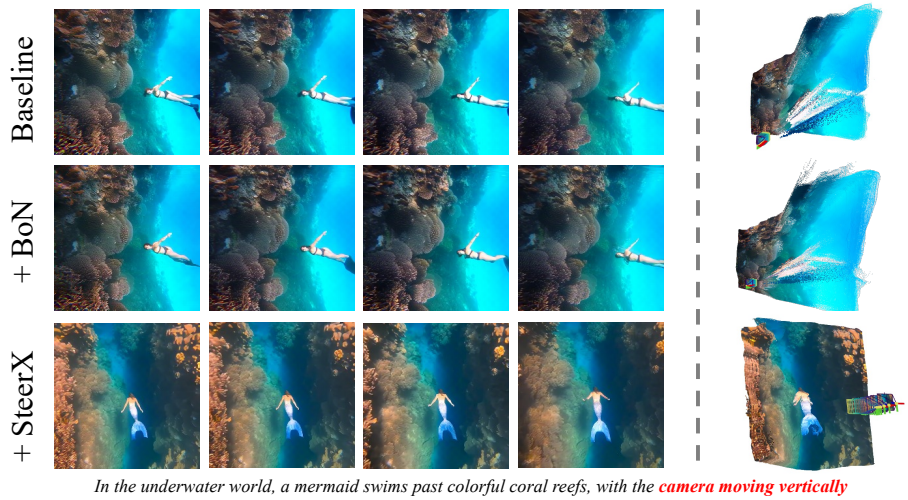
"Under the warm sunshine, a little dog is eating slowly."

Figure 15. **Additional qualitative results in Text-to-4D.**



In the center of the living room, there is a sofa, and to the right of the sofa is a massage chair. ***The camera horizontally moving from left to right.***

A khaki-colored fisherman's hat made of canvas, with a wide, round brim, is hanging on a coat rack behind the door. ***The camera zooms in*** *to highlight the small daisy pattern embellished on the hat.*



In the underwater world, a mermaid swims past colorful coral reefs, with the ***camera moving vertically from top to bottom during filming.***

Figure 16. **Qualitative comparisons on Text-to-4D.**