A Unified Approach for Text- and Image-guided 4D Scene Generation

Yufeng Zheng[1,2,3], Xueting Li[1], Koki Nagano[1], Sifei Liu[1], Otmar Hilliges[2], Shalini De Mello[1]
[1]NVIDIA, [2]ETH Zurich, [3]Max Planck Institute for Intelligent Systems

1. Societal impact

We note that our work could be potentially used to generate fake 4D content that is violent or harmful. Building upon pre-trained large-scale diffusion models, it inherits the biases and limitation of these models. Therefore, the 4D videos generated with our method should be carefully examined and labeled as synthetic content.

2. More Results

Video results Please refer to our web page (https://research.nvidia.com/labs/nxp/dream-in-4d/) for video results of the text-to-4D, image-to-4D and personalized 4D tasks. We also provide qualitative comparisons of our method and ablation baselines.

More rendering angles In Fig. 1, we show renderings from the top view and the bottom view. Our 4D assets produce plausible renderings from different elevation angles.

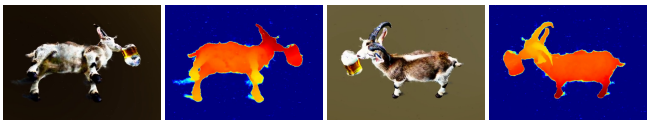


Figure 1. **Renderings from different elevation angles**.

3. Network Architecture

Canonical NeRF. We use a hash-encoded multi-resolution feature grid with 16 resolution levels, where the base resolution is $16 \times 16 \times 16$ and maximum resolution is $4096 \times 4096 \times 4096$. The feature grid is followed by two shallow MLPs to produce the density and color values. Both of the MLP networks have 1 hidden layer and 64 neurons per layer.

Deformation field. The deformation field uses a hash-encoded multi-resolution feature grid with 12 levels, where the base resolution is $4 \times 4 \times 4 \times 4$ and maximum resolution is $232 \times 232 \times 232 \times 232$. The feature grid is followed by an MLP with 4 hidden layers and 64 neurons per layer to predict the displacement values $\mathbf{d}$ for scene deformation. We then calculate the canonical point location by $\mathbf{x_c} = \mathbf{x_d} + \mathbf{d}$.

Prompts	λ_{2D}	λ_{3D}
A superhero dog wearing a red cape flying through the sky	1.2	1
A cat singing	1.2	1
A dog riding a skateboard	1	1
Clown fish swimming through coral reef	1	1
A fox playing a video game	1.2	1
A goat drinking beer	1	1
A monkey eating a candy bar	1	1
An emoji of a baby panda reading a book	1.2	1
A baby panda eating ice cream	1	1
A squirrel riding a motorcycle	1.2	1

Table 1. 2D and 3D guidance loss weights for the static stage.

Background. We model the background with an MLP, which takes the viewing direction as input and outputs a color value. This assumes that the background is located infinitely far away from the camera. The MLP has 3 hidden layers and 64 neurons per layer.

4. Training Schedule

4.1. Static Stage

For the static stage, we render multi-view images of resolution 64×64 with a batch size of 8 for the first 5000 iterations, and resolution 256×256 with a batch size of 4 for the last 5000 iterations. For MVDream [6] guidance, the images are upsampled to 256×256. For StableDiffusion [5], we upsample to 512×512. We use guidance scale of 50 for MVDream and 100 for StableDiffusion.

We use an AdamW [3] optimizer with a learning rate of 0.001 for all the MLP parameters and 0.01 for the parameters of hash-encoded multi-resolution feature grid. The β parameters are set to 0.9 and 0.99, respectively. We train the networks for 10000 iterations on a NVIDIA V100 GPU, which takes 4.5 hours.

We found that balancing the 3D and 2D guidance weights to be important for achieving view-consistent, text-aligned and realistic results. In Tab. 1, we list the loss weights used for all prompts in the paper.

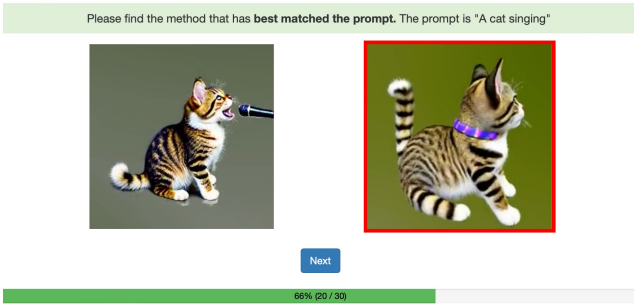
4.2. Dynamic Stage

To learn deformation in the dynamic stage, we use guidance from the Zeroscope video diffusion model [2], where we render 24-frame videos of resolution 144×80. We found our method to also work well with guidance from the Modelscope video diffusion model [1], in which case we rendered videos of resolution 64×64 for the first 7000 iterations and then upsample to 256×256. We use a guidance scale of 100 for both Zeroscope and Modelscope guidance, and gradually decrease the time step used for the SDS loss [4] from $[0.99, 0.99]$ to $[0.2, 0.5]$.

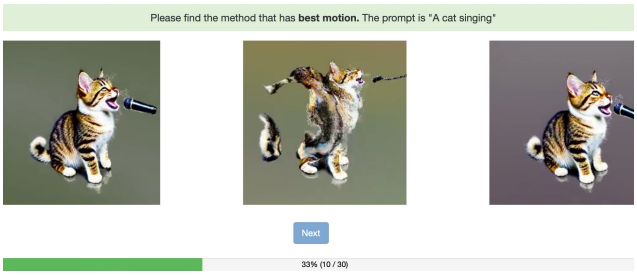
In the dynamic stage, we optimize the deformation parameters with an AdamW [3] optimizer with learning a rate of 0.001 and $\beta = [0.9, 0.99]$. We train the deformation network for 10000 iterations on a NVIDIA A100 or RTX A6000 GPU. We start by using the first 4 levels of the multi-resolution features, and gradually include the higher resolution features, adding 1 level every 500 iterations. The dynamic stage takes 9 hours for Zeroscope, and 6 hours for Modelscope due to the lower inference resolution of the diffusion model.

5. User Study

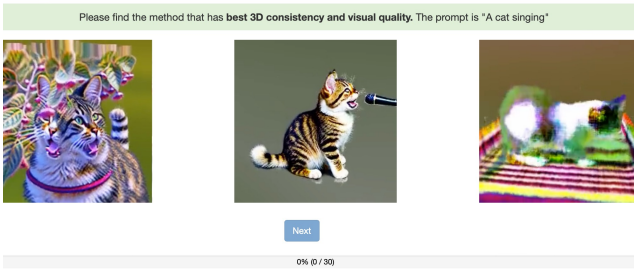
In this section, we introduce more details of the user study. As discussed in Sec. 4.1 of the main paper, we compare against state-of-the-art method (MAV3D) and ablative baselines through a user preference study. Specifically, we present the results from our method and the baseline(s) to a user and give instructions to "Please find the method that best matched the prompt / has best motion / has best 3D consistency and visual quality." The example interface is shown in Fig. 2.



(a) Interface example corresponding to Table 1 (a) in the main paper.



(b) Interface example corresponding to Table 1 (b) in the main paper.



(c) Interface example corresponding to Table 1 (c) in the main paper.

Figure 2. **User study interface.**

References

[1] Modelscope. https://huggingface.co/damo-vilab/text-to-video-ms-1.7b. 2

[2] Zeroscope. https://huggingface.co/cerspense/zeroscope_v2_576w. 2

[3] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1, 2

[4] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

[6] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 1