Supplementary Material
Generative Image Dynamics

Zhengqi Li Richard Tucker Noah Snavely Aleksander Holynski
Google Research

1. Spectral volumes as image-space modal bases

In this section, we provide details on the connection between our predicted spectral volumes and image-space modal bases, which can be used to simulate interactive dynamics through modal analysis. We refer readers to the original work in computer graphics and structure engineering for more thorough theory and analysis [3, 5, 16, 18]. In particular, we treat the pixels of the input image as a set of points (with cardinality $|P|$) that are linked to each other via mass-spring-damper system. Starting from an object (modeled as a harmonic oscillator) at rest, that object's motion response to an external force $\mathbf{f}(t)$ follows the equation of motion:

$$M\ddot{\mathbf{u}}(t) + C\dot{\mathbf{u}}(t) + K\mathbf{u}(t) = \mathbf{f}(t) \tag{1}$$

where $\mathbf{u}$ is the motion displacement of a given scene represented as a vector of size $|P|$, and M, C, and K are the mass, damping and stiffness matrices respectively, dictating the intrinsic dynamics of the object [16].

One can transform this equation of motion into modal space, in which M, C, and K reduce to diagonal matrices and we can decouple the equation of motions into a set of $|P|$ independent single-degree-of-freedom systems:

$$\ddot{\mathbf{q}}_i(t) + c_i\dot{\mathbf{q}}_i(t) + k_i\mathbf{q}_i(t) = \frac{f_i(t)}{m_i} \tag{2}$$

where m_i, c_i, and k_i are the diagonalized elements of the mass damping and stiffness matrices in modal space; and $\mathbf{q}_i$ and f_i correspond to the motion displacement $\mathbf{u}$ and force $\mathbf{f}$ in modal space. As in prior work, we do not directly estimate these matrices, but instead only use the intermediate modal space to derive equations of response. We can further simplify Equation 2 under the common assumptions of Rayleigh damping: $c_i = \alpha m_i + \beta k_i$, where we empirically set $\alpha = 0.4$ and $\beta = 0.08$ in our case. This gives rise to

$$\ddot{\mathbf{q}}_i(t) + \gamma_i\dot{\mathbf{q}}_i(t) + \omega_i^2\mathbf{q}_i(t) = \frac{f_i(t)}{m_i} \tag{3}$$

$$\omega_i^2 = \frac{k_i}{m_i}, \quad \gamma_i = \alpha + \beta\omega_i^2 \tag{4}$$

where ω is the damped natural frequency.

Moreover, Davis *et al*. [4] shows that the temporal Fourier transform of per-pixel motion trajectory is approximately proportional to the image-space projection of mode shapes at resonant frequencies. Therefore, we can treat the spectral volume as a basis and interpret the modal displacement $\mathbf{q}_i$ as the displacement of the object, in order to model and simulate image-space object dynamics. In particular, the final displacement of a pixel $\mathbf{p}$ in image-space, F_t, can be written as a sum of modal displacements weighted by the corresponding coefficients in the spectral volume over selected frequency bands f_j:

$$F_t(\mathbf{p}) = \sum_{f_j} S_{f_j}(\mathbf{p})\mathbf{q}_{f_j}(t). \tag{5}$$

In our setting, we use all predicted frequency bands from our latent diffusion modal as a basis in order to to avoid manual mode selection performed in prior work. To simulate complex modal displacement $\mathbf{q}_{f_j,t}$, we perform explicit Euler integration over Equation 4 to update the modal displacement, velocity, and acceleration over time:

$$\ddot{\mathbf{q}}_i(t+\delta t) = \frac{f_i(t)}{m_i} - \gamma_i\dot{\mathbf{q}}_i(t) - \omega_i^2\mathbf{q}_i(t) \tag{6}$$

$$\dot{\mathbf{q}}_i(t+\delta t) = \dot{\mathbf{q}}_i(t) + \delta_t\ddot{\mathbf{q}}_i(t+\delta t) \tag{7}$$

$$\mathbf{q}_i(t+\delta t) = \mathbf{q}_i(t) + \delta_t\ddot{\mathbf{q}}_i(t+\delta t) \tag{8}$$

where we set $m_i = 1$ throughout our experiment.

Further, we project force vectors into modal space in order to initial the complex modal displacement. In particular, as in original work of Davis *et al*., we compute the magnitude of the initial state of the modal displacement as

$$||\mathbf{q}_{f_j}(0)|| = ||\frac{\mathbf{f}(0)}{||\mathbf{f}(0)||_2} \cdot S_{f_j}||_2 \tag{9}$$

and we compute the phase ϕ of the initial state of the modal displacement for forces from a "drag and release" interaction as:

$$\phi_{\text{drag}}(\mathbf{q}_{f_j}(0)) = -\phi(\frac{\mathbf{f}(0)}{||\mathbf{f}(0)||_2} \cdot S_{f_j}). \tag{10}$$

(a) Input image (b) Motion derived W

Figure 1. We visualize input RGB images and corresponding weights derived from the magnitude of motion texture.

2. Image-based rendering

In the main manuscript we describe how we use the predicted motion magnitude to determine the contributing weight of each source pixel mapped to its destination location, using motion magnitude as a proxy for depth following the work of Davis, *et al*. [4]. In particular, we compute a per-pixel weight, $W(\mathbf{p}) = \frac{1}{T}\sum_t ||F_t(\mathbf{p})||_2$ as the average magnitude of the predicted motion texture in order to determine the contributions of colliding source pixels at destination time (shown in Fig. 1):

$$I'_t(\mathbf{p} + F_t\mathbf{p}) = \frac{\sum I_0(\mathbf{p}) \cdot W(\mathbf{p})}{\sum W(\mathbf{p})}. \tag{11}$$

We use motion-derived weights instead of learnable ones because we observe that in the single-view case, learnable weights are not effective for addressing disocclusion ambiguities, as shown in the second column of Figure 2.

Moreover, we choose to perform per-frame independent rendering instead of creating layered representations [17, 19], since we found that the latter configuration can lead to significantly more visible artifacts and distortions near object boundaries. In addition, since most of the motions we produce are small, the method is only required to inpaint relatively small unseen regions during rendering. Therefore, we find that per-frame refinement is sufficient in our case.

3. Frequency-adaptive normalization

We show additional visualization of spectral volume coefficients at different frequency in Fig. 3, where we observe

(a) Average-splat (b) Learned W (c) W from motion

Figure 2. From left to right, we show a rendered future frame with (a) average splatting in RGB pixel space, (b) softmax splatting with learnable weights [9], and (c) motion-aware feature splatting.

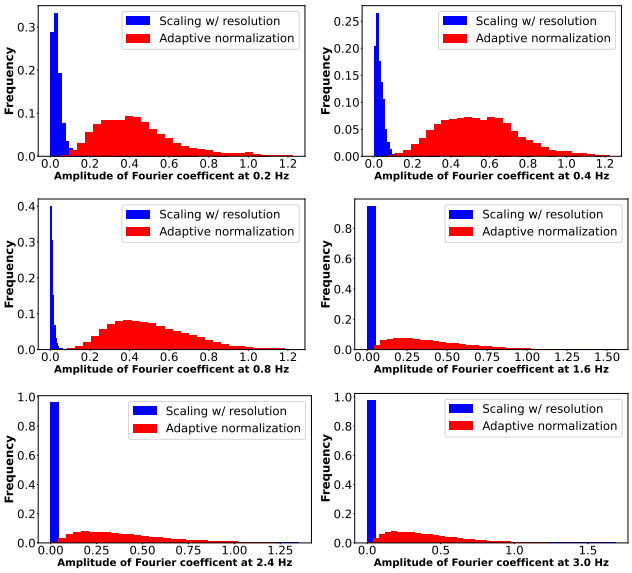

Figure 3. Histogram of the amplitudes of Fourier terms across frequencies after (1) scaling amplitude by image width and height (blue), or (2) frequency adaptive normalization (red).

that our frequency-adaptive normalization redistributes coefficients more evenly across different frequencies.

4. Additional implementation details

4.1. Network architecture

We use a VAE of continuous latent dimension 4 for each frequency slice of the spectral volume, and use base channel 128 with channel multiplier 1, 2, and 4 for the VAE. We perform whitening to normalize encoded VAE latent features. For the 2D diffusion model, we use base channels of 128, channel multipliers 1, 4, and 8, and attention resolutions 32, 16, and 8 for each block. We downsample the input RGB image using a standard ResNet encoder to produce 16-channel features and concatenate them with a 4-channel noisy latent (or Gaussian noise map during inference) before

Figure 4. **Visual comparisons of generated future frames and corresponding motion fields.** We show generated future frame (odd rows) and estimated motion fields between the input and corresponding generated images. By inspecting differences with a reference image from the ground truth video, we observe that our approach produces more realistic textures and motions compared with baselines.

feeding them to the denoising network.

During training, we use 1,000 diffusion steps, and a square linear noise schedule to perform latent denoising. We adopt Adam [11] to train the LDM model for 750K steps with batch size 96 and initial learning rate 5^{-5}. To avoid overfitting, we apply random data augmentation by performing color jittering, random horizontal flips, random image scaling, random rotation within five degrees, and random crops. During inference, each step of DDIM sampling takes 0.5 seconds for motion prediction.

We adopt a ResNet-34 [8] as a feature extractor in our image-based rendering module. Specifically, we encode I_0 through a feature extractor network to produce a multi-scale feature map $\mathcal{M} = \{M_j | j = 0, ..., J\}$. For each individual feature map M_j at scale j, we resize and scale the predicted 2D motion field F_t according to the resolution of M_j. With the motion field F_t and weights W, we apply softmax splatting to warp the feature map at each scale to produce a warped feature $M'_{j,t} = \mathcal{W}_{\text{softmax}}(M_j, F_t, W)$, where $\mathcal{W}_{\text{softmax}}$ is the softmax splatting operation. The image synthesis network is based on a co-modulation StyleGAN architecture [12, 21], where we inject both style features mapped from Gaussian noise and warped features $M'_{j,t}$ into decoder blocks at athe corresponding scale to produce a final rendered image $\hat{I}_t$. We adopt Adam [11] to train the rendering model with batch size 32 and initial learning rate 10^{-4} for 300K iterations.

4.2. Data

As mentioned in the main manuscript, we collect and process a set of 3,015 videos depicting scenes with oscillation dynamics from online footage websites as well as from our own captures. In terms of videos from online sources, we collect this kind of footage by using query texts like “Static shot, trees/flowers/candles/clothes/lanterns, wind/breeze”, and we also use more specific names of everyday trees and flowers to mine more data, (examples include maples, oak, beech, elm, pine, spruce, redwood, rose, daisy, carnation, tulips, chrysanthemum, dahlia, sunflowers, orchid, lily, iris, cherry blossom, bushes, ivy, and dandelions). Furthermore,

we remove the clips with strong camera motions by checking if 95% pixels have average magnitude of motion trajectory larger than one pixel; we also remove the clips without scene motion by checking if average magnitude of estimated motion trajectories is less than one pixel.

To generate ground truth spectral volumes, we find the choice of optical flow method to be crucial. In particular, we observe that deep-learning based flow estimators tend to produce over-smoothed flow fields, which results in unrealistic animation through image based rendering. Instead, we apply a coarse-to-fine image pyramid-based flow algorithm [2, 13] between selected starting image and every future frame within a video sequence to derive motion trajectory. We treat every 10th frame from each video as a starting image, compute optical flows from the starting image and the following 149 frames, and derive spectral volumes by applying temporal FFT to the estimated per-pixel motion trajectory and selecting the first K frequency slices in Fourier domains.

4.3. Large video diffusion model baselines

AnimateDiff [7] Since the original implementation of AnimateDiff only supports artificial images generated with Stable Diffusion as a starting frame, we perform DDIM inversion [14] to generate a video from a given real input RGB image. We use the model from Realistic Vision V2.0 as a backbone since it is most related to the natural images and motion we focus on. We manually add additional text inputs according to the context of input picture. For example, we use the prompt "flowers, wind, 8k uhd, dslr, soft lighting, high quality, film grain, Fujifilm XT3" for an input image depicting flowers.

ModelScope [20]. The ModelScope model is based on the work of VideoComposer [20], which supports multimodaility inputs such as text, input, sketch, motion vectors or depth maps. Therefore, we feed our input image to the model to generate corresponding video clip.

GEN-2 [1]. GEN-2 is a recent commercial video generation solution that supports text- and image-to-video tasks. For each input, we not only feed the image but also manually provide text describing the image, as we found that additional text descriptions systematically improve video generation. For instance, for a picture of trees, we provide the text prompt "A static shot of trees swaying in the wind, masterpiece."

User study. We perform a user study and compare our generated animations with the three baselines mentioned above. On a randomly selected 30 videos from the test set, we ask users "which video is more realistic?". A total of 35 users finished the study. We found a 71.1% preference rate for our method over AnimateDiff, a 83.7% preference rate over ModelScope, and a 82.6% preference rate over GEN-2. We also provide visual video comparisons in the supplementary website.

5. Additional qualitative comparisons

We provide additional comparisons of the quality of individual frames and motions synthesized by our approach and by other baselines [6, 10, 15] by visualizing the predicted video frame $\hat{I}_t$ and its corresponding motion displacement field at time $t = 128$. Figure 4 shows that the frames generated by our approach exhibit fewer artifacts and distortions compared to other methods, and our corresponding 2D motion fields most resemble the reference displacement fields estimated from the corresponding real videos. In contrast, the background content generated by other methods tends to drift, as shown in the flow visualizations in the even-numbered rows. Moreover, the video frames generated by other methods exhibit significant color distortion or ghosting artifacts, suggesting that the baselines are less stable when generating videos with long time duration.

References

[1] Gen-2. https://research.runwayml.com/gen2.

[2] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 25–36. Springer, 2004.

[3] Abe Davis, Katherine L Bouman, Justin G Chen, Michael Rubinstein, Fredo Durand, and William T Freeman. Visual vibrometry: Estimating material properties from small motion in video. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5335–5343, 2015.

[4] Abe Davis, Justin G Chen, and Frédo Durand. Image-space modal bases for plausible manipulation of objects in video. *ACM Transactions on Graphics (TOG)*, 34(6):1–7, 2015.

[5] Myers Abraham Davis. *Visual vibration analysis*. PhD thesis, Massachusetts Institute of Technology, 2016.

[6] Yuki Endo, Yoshihiro Kanamori, and Shigeru Kuriyama. Animating landscape: Self-supervised learning of decoupled motion and appearance for single-image video synthesis. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2019)*, 38(6):175:1–175:19, 2019.

[7] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Aleksander Holynski, Brian L Curless, Steven M Seitz, and Richard Szeliski. Animating pictures with Eulerian motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5810–5819, 2021.

[10] Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and

Shuchang Zhou. A dynamic multi-scale voxel flow network for video prediction. *ArXiv*, abs/2303.09875, 2023.

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In *European Conference on Computer Vision*, pages 515–534. Springer, 2022.

[13] Ce Liu. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009.

[14] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[15] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18444–18455, 2023.

[16] Ahmed A Shabana. *Theory of vibration*, volume 2. Springer, 1991.

[17] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8028–8038, 2020.

[18] Jos Stam. Stochastic dynamics: Simulating the effects of turbulence on flexible structures. *Computer Graphics Forum*, 16(3), 1997.

[19] Qianqian Wang, Zhengqi Li, David Salesin, Noah Snavely, Brian Curless, and Janne Kontkanen. 3d moments from near-duplicate photos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3906–3915, 2022.

[20] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023.

[21] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2021.