

Supplemental Material: Tracking by Predicting 3-D Gaussians Over Time

Anonymous CVPR submission

Paper ID *****

1. Supplementary Material

This document provides a detailed description of the supplementary videos accompanying our paper, "Tracking by Predicting 3-D Gaussians Over Time." These videos offer qualitative insights into the performance of our proposed *Video-GMAE* model, complementing the quantitative evaluations presented in the main manuscript. The videos, provided as MP4 files in the supplementary zip and referenced by filename in the sections below, showcase zero-shot tracking capabilities, tracking after fine-tuning, and visualizations of the pre-training reconstruction process. In addition to these videos, we provide full-page static figures that summarize representative frame sequences and qualitative comparisons between *Video-GMAE* and prior work.

1.1. Zero-shot Tracking Examples

Our *Video-GMAE* model is pre-trained through a self-supervised objective that enforces temporal correspondence by predicting the evolution of 3D Gaussian primitives over time. This pre-training allows the model to perform point tracking in a zero-shot manner, without any direct supervision from labeled tracking data. The following videos illustrate this emergent capability. This provides a qualitative complement to the zero-shot quantitative results reported in Table 2 of the main paper.

- 1. Zero-shot tracking on TAP-Vid Davis dataset:** Selected points are tracked across frames, showcasing the model's ability to maintain correspondence for objects undergoing motion and deformation in diverse visual contexts. These visualizations correspond to the supplementary video files `zeroshot-davis.mp4` and `zeroshot_davis2.mp4`.
- 2. Zero-shot tracking on TAP-Vid Kinetics dataset:** Similar to the DAVIS examples, this video presents zero-shot tracking results on sequences from the Kinetics dataset. The scenes in Kinetics often feature a wide array of human actions and complex interactions, providing a different set of challenges. The video highlights the model's capacity to track points even in the presence of varied object classes and complex motion. These visualizations correspond to the supplementary

video files `zeroshot-kinetics.mp4` and `zeroshot-kinetics2.mp4`.

- 3. Failure cases in zero-shot tracking:** To provide a balanced view of our model's zero-shot capabilities, this video illustrates common failure modes. As discussed in the main paper's Limitations section, *Video-GMAE*'s zero-shot tracking performance can degrade under certain conditions. Specifically, scenes with complex, high-frequency backgrounds, combined with significant camera motion, pose a challenge. This is attributed to two main factors: (i) the pre-training assumption of a static camera, which is violated in such scenarios, and (ii) the fixed budget of 256 Gaussian primitives used during pre-training. In visually dense backgrounds, a substantial portion of these Gaussians may be allocated to modeling the static parts of the scene, leaving insufficient capacity to accurately represent the motion of foreground objects or to disambiguate object motion from camera-induced motion. This video exemplifies these situations where tracking accuracy diminishes. These visualizations correspond to the supplementary video file `zeroshot-failure.mp4`.

Figures 1, 2, and 3 provide static visualizations corresponding to these zero-shot tracking examples. The figures highlight typical point trajectories on TAP-Vid DAVIS and TAP-Vid Kinetics, as well as characteristic failure modes in challenging settings, complementing the dynamic behavior shown in the videos.

1.2. Qualitative Comparison with GMRW-C

The main paper reports quantitative zero-shot tracking metrics and presents a small set of qualitative comparisons between *Video-GMAE*-zeroshot and the strongest self-supervised baseline, GMRW-C [1]. Here, we expand on these analyses by providing additional examples on TAP-Vid DAVIS and TAP-Vid Kinetics in Figures 4 and 5. The supplementary videos `comparison_grid1.mp4` and `comparison_grid2.mp4` provide dynamic visualizations of these qualitative comparisons.

Across many sequences, we observe that *Video-GMAE*-zeroshot tends to produce temporally coherent tracks: once a point is visible, its identity is usually preserved without short-



Figure 1. **Zero-shot tracking on TAP-Vid DAVIS.** Representative zero-shot point tracking results of *Video-GMAE* on TAP-Vid DAVIS. Colored points denote tracked locations over time. The examples illustrate that our model can maintain long-range correspondences for deforming objects and complex motions without task-specific supervision.

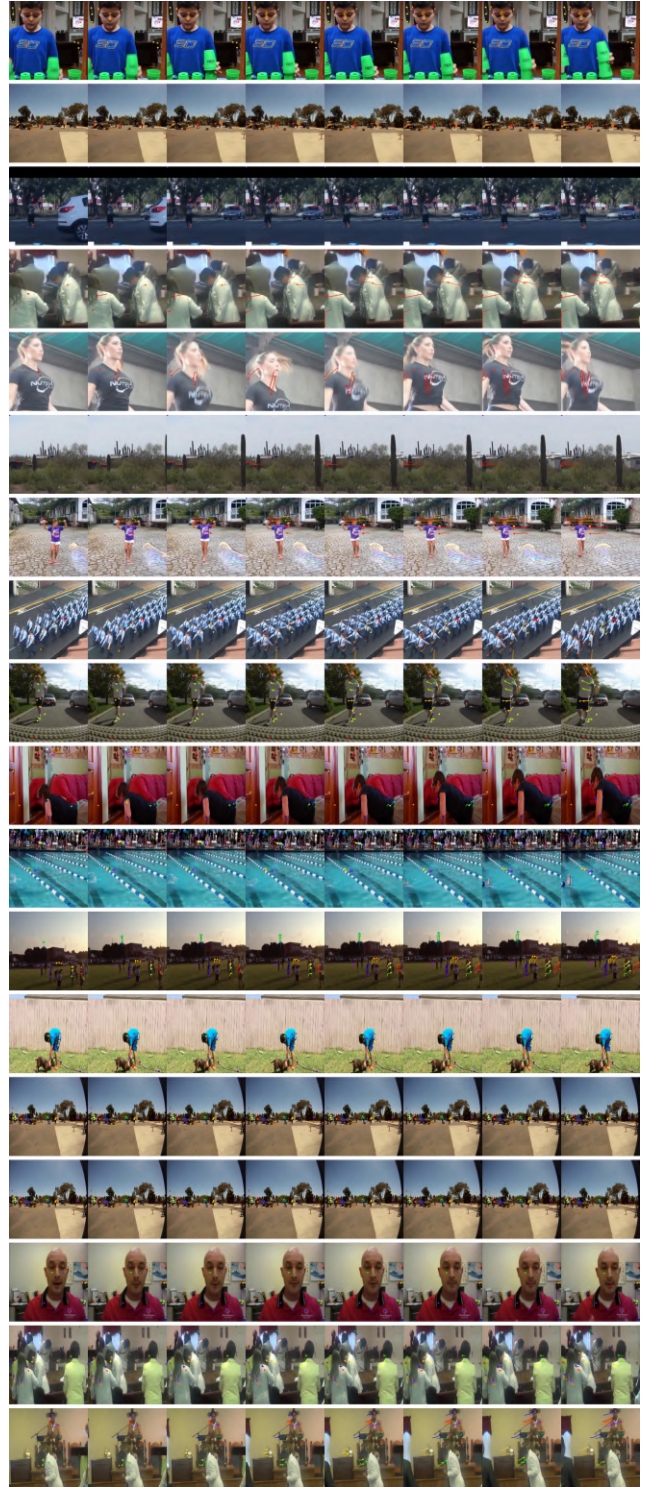


Figure 2. **Zero-shot tracking on TAP-Vid Kinetics.** Static visualization of zero-shot tracking with *Video-GMAE* on TAP-Vid Kinetics sequences. Colored points denote tracked locations over time. The scenes involve diverse human actions and object interactions, and the model is able to track points across rapid motions and appearance changes.

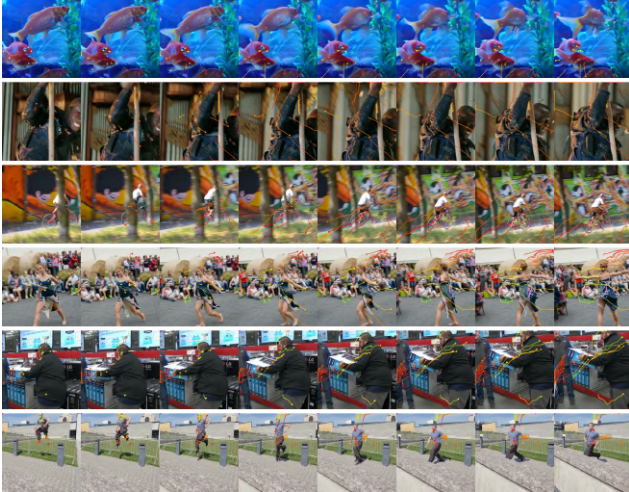


Figure 3. **Zero-shot failure cases.** Examples of challenging zero-shot tracking scenarios where *Video-GMAE* degrades, typically due to complex, high-frequency backgrounds and substantial camera motion. The figure reflects failure modes discussed in the main paper, where the limited 256-Gaussian budget and the static-camera pre-training assumption can reduce tracking accuracy on moving foreground objects.



Figure 4. **Qualitative comparison with GMRW-C on TAP-Vid DAVIS.** Representative zero-shot tracking sequences comparing *Video-GMAE*-zeroshot (the sequences with the green border) to GMRW-C [1] on TAP-Vid DAVIS. Each strip shows point trajectories over time for both methods on challenging motions and occlusions. On TAP-Vid Davis, *Video-GMAE* is conservative at occluding tracks, but has a slightly worse tracking precision, which leads to a similar Average Jaccard score compared to GMRW-C.

1.3. Fine-tuned Tracking Examples

While *Video-GMAE* exhibits strong zero-shot tracking, its representations can be further adapted for enhanced perfor-

lived disappearances or rapid switches, leading to smoother trajectories over time. In contrast, GMRW-C sometimes exhibits frame-to-frame instabilities such as early occlusion predictions, brief drop-outs, or re-activations of points that reduce the effective temporal continuity of the track. We also see that our method often maintains point visibility until objects fully exit the frame, making it more conservative around occlusions, whereas GMRW-C may mark points as occluded several frames before they leave view and occasionally re-activate them spuriously.

At the same time, these comparison strips highlight scenarios where GMRW-C can be advantageous. In particular, when small, visually detailed regions undergo large displacements, GMRW-C sometimes preserves point locations more faithfully than *Video-GMAE*-zeroshot. This is particularly evident for the TAP-Vid Davis videos, where *Video-GMAE*-zeroshot predicts occlusions better but sacrifices point location precision, which is qualitatively consistent with the Average Jaccard score being similar between the two methods. This behavior is consistent with a limitation of our representation: modeling each scene with only 256 Gaussian primitives constrains the spatial resolution at which fine-scale details can be reconstructed and tracked. Overall, these additional comparisons reinforce the quantitative results and suggest that our Gaussian-based zero-shot tracker is well-suited for stable, occlusion-aware tracking, while leaving room for future work to improve fine-detail fidelity under large motions.

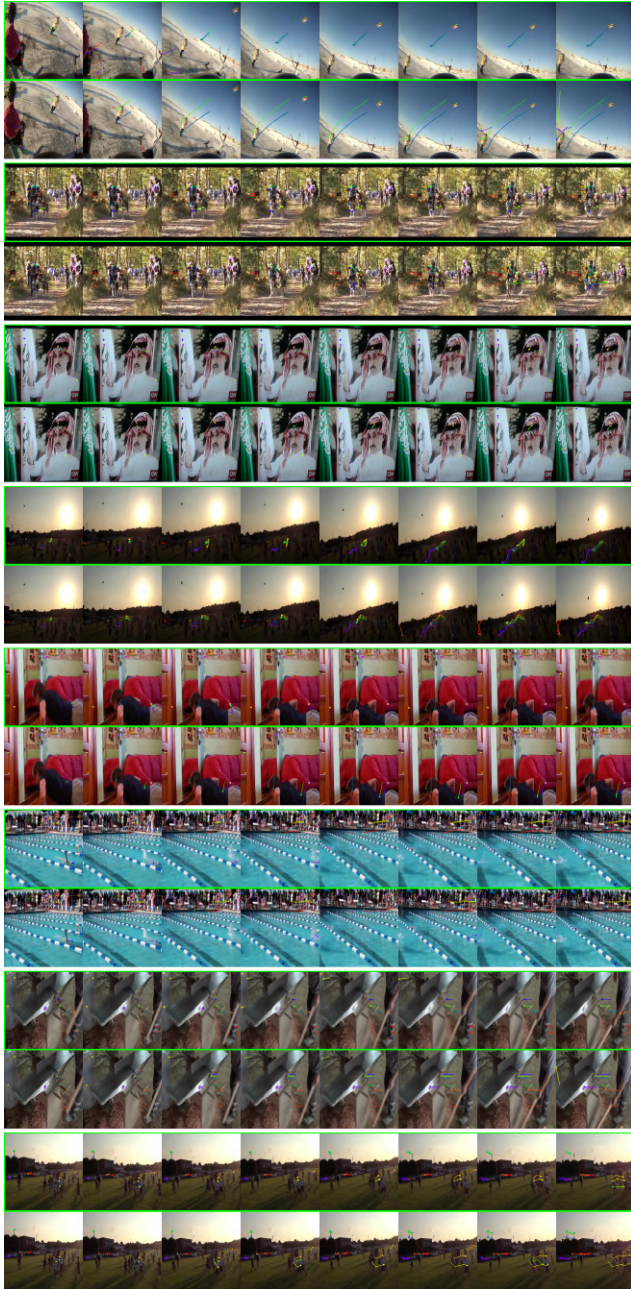


Figure 5. **Qualitative comparison with GMRW-C on TAP-Vid Kinetics.** Additional zero-shot tracking comparisons on TAP-Vid Kinetics sequences. While *Video-GMAE*-zeroshot (the sequences with the green border) often yields smoother, more persistent tracks, GMRW-C can better preserve points on small, fine-scale structures undergoing large displacements. These examples illustrate the trade-offs between the two approaches and complement the quantitative metrics reported in the main paper.

mance via supervised fine-tuning on task-specific datasets. The following videos demonstrate the capabilities of our model after such fine-tuning, corresponding to the fine-tuned

results presented in Table 2 of our main paper.

1. **Fine-tuned tracking on TAP-Vid Davis dataset:** This video showcases the improved point tracking precision and robustness of *Video-GMAE* on the TAP-Vid DAVIS dataset after supervised fine-tuning. The tracking is visibly more stable and accurate compared to the zero-shot results, especially in challenging sequences involving occlusions, rapid motion, and multi-object interactions. These sequences are provided in the supplementary video files `finetune-davis.mp4` and `finetune-davis2.mp4`.
2. **Fine-tuned tracking on TAP-Vid Kinetics dataset:** Similar to the prior video, this video shows the improved point tracking on TAP-Vid Kinetics dataset after supervised fine-tuning. These sequences are provided in the supplementary video files `finetune-kinetics.mp4` and `finetune-kinetics2.mp4`.

Figures 6 and 7 provide static examples of fine-tuned tracking on both TAP-Vid DAVIS and TAP-Vid Kinetics, illustrating the reduction in drift and jitter as well as improved robustness to occlusions compared to the zero-shot case.

1.4. Video-GMAE Pre-training Gaussians Rendered

The core of *Video-GMAE* lies in its self-supervised pre-training strategy, where the model learns to reconstruct video sequences by predicting the parameters and temporal dynamics of 3D Gaussian primitives within an MAE framework. The videos in this section show what the reconstructed videos look like.

1. **Rendered Videos:** These three videos (renders1.mp4, renders2.mp4, and renders3.mp4) visualize the output of the *Video-GMAE* decoder during the pre-training phase. For a given input video, the model predicts a set of 3D Gaussians for the initial frame and their subsequent transformations for the following frames. These Gaussians are then rendered to reconstruct the video sequence. The visualizations demonstrate the model’s ability to represent scene content using a collection of dynamic Gaussian primitives.

Figures 8 and 9 show static renderings from two sets of pre-training reconstructions. These stills highlight how a relatively small number of Gaussians can capture coarse geometry, appearance, and motion, while also revealing the limitations of the representation in highly textured or fine-detail regions.

References

- [1] Ayush Shrivastava and Andrew Owens. Self-supervised any-point tracking by contrastive random walks. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 3

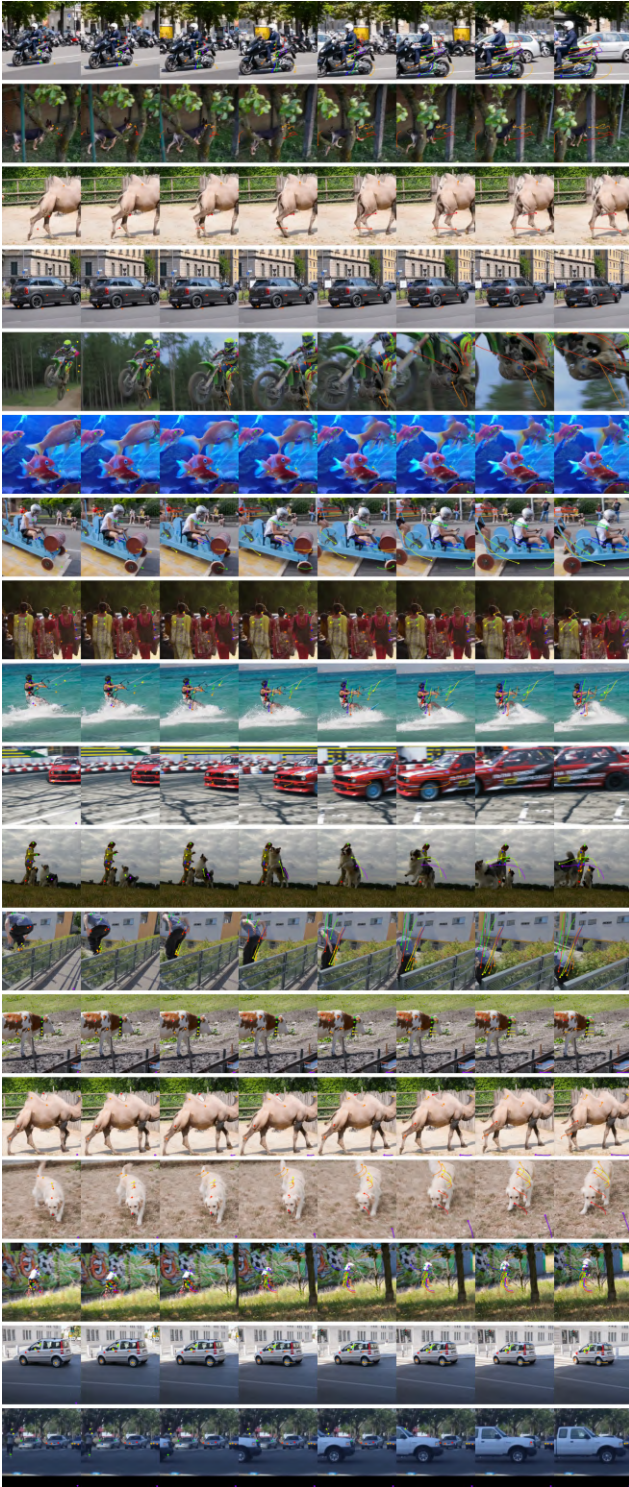


Figure 6. **Fine-tuned tracking on TAP-Vid DAVIS.** Still-frame visualization of point tracking with *Video-GMAE* after supervised fine-tuning on TAP-Vid DAVIS. Compared to the zero-shot sequences in Figures 1 and 3, the fine-tuned model exhibits more precise localization, reduced temporal jitter, and improved handling of occlusions and object interactions.

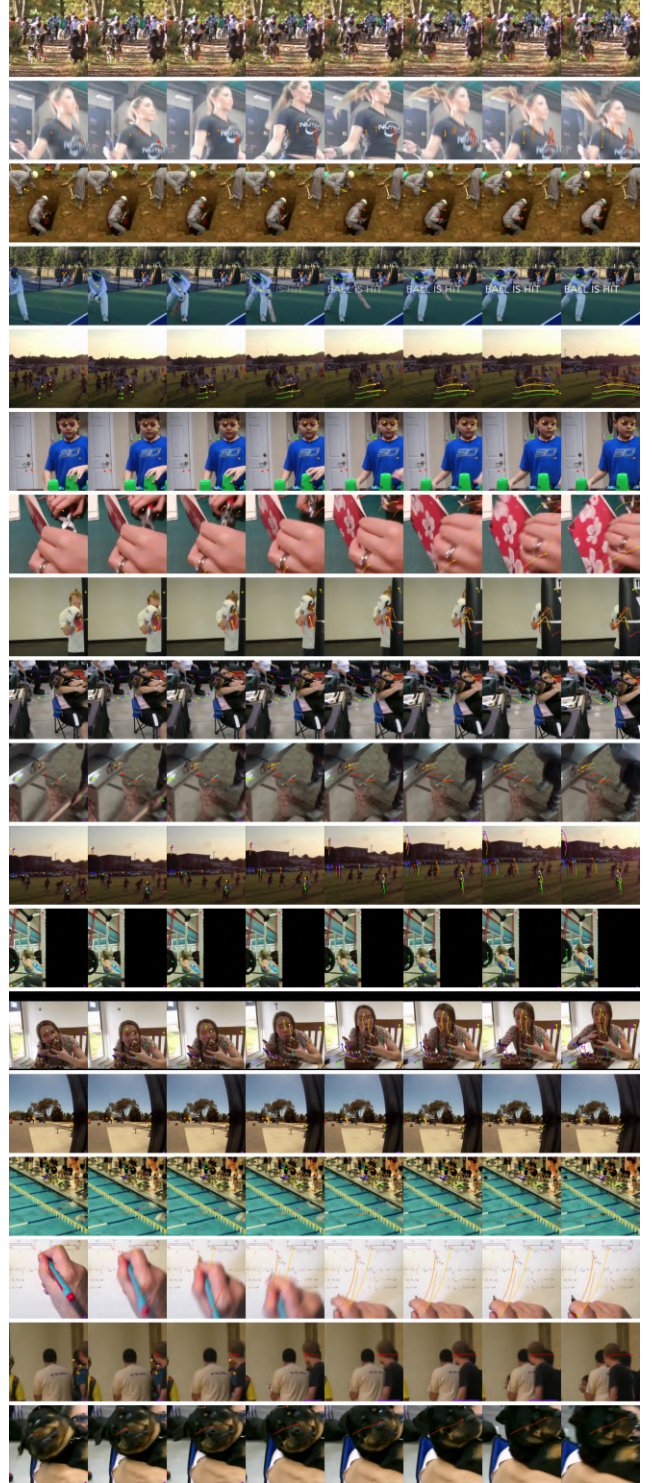


Figure 7. **Fine-tuned tracking on TAP-Vid Kinetics.** Static example of fine-tuned tracking on TAP-Vid Kinetics. Fine-tuning leads to crisper trajectories and more stable point identities across complex human actions, complementing the qualitative improvements observed in the corresponding videos.

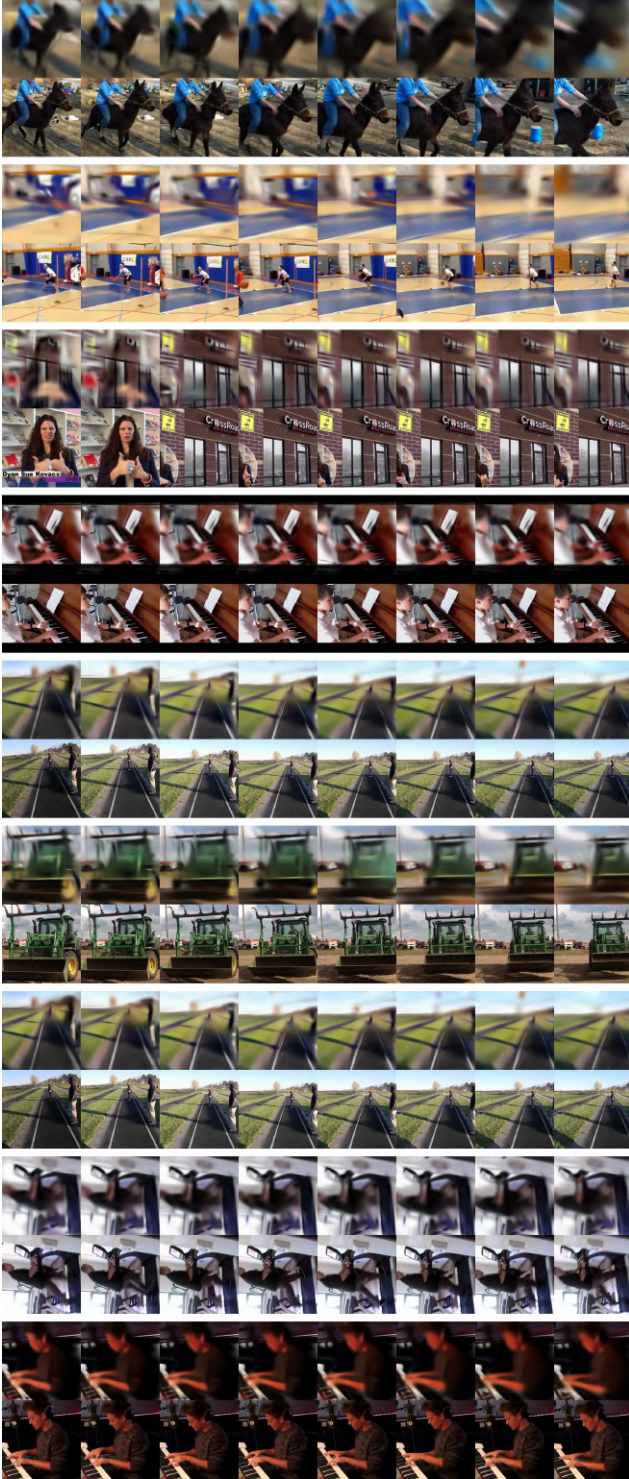


Figure 8. **Video-GMAE** pre-training reconstructions (set 1). Example frames from the rendered outputs of the *Video-GMAE* decoder during pre-training. For each sequence, the model predicts an initial set of 3D Gaussians and their temporal evolution, which are then rendered to reconstruct the video. The visualizations show that a compact set of dynamic Gaussian primitives can capture the dominant structure and motion in the scene.

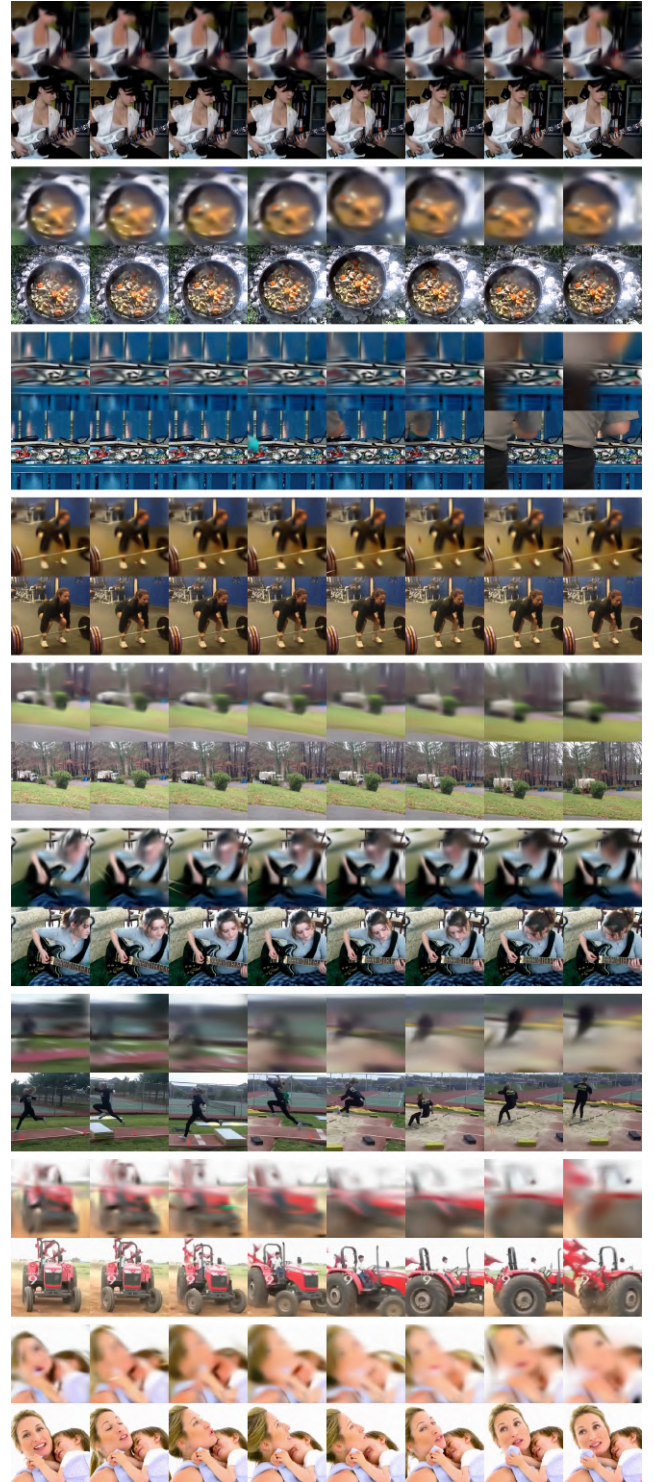


Figure 9. **Video-GMAE** pre-training reconstructions (set 2). Additional examples of pre-training reconstructions on different sequences. While the Gaussians capture coarse geometry and dynamics, fine-scale textures and small structures are limited by the 256-Gaussian budget, consistent with the discussion of representation capacity in the main paper.