

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

Tanish Baranwal

Himanshu Gaurav Singh

Jathushan Rajasegaran

Jitendra Malik

University of California, Berkeley

tanish@berkeley.edu

A. 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 zeroshot tracking capabilities, tracking after fine-tuning, and visualizations of the pretraining 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.

A.1. Zeroshot Tracking Algorithm

For completeness, we include a compact pseudocode summary of the zeroshot tracking procedure used in Section 4 of the main paper in Algorithm 1.

Algorithm 1 Pseudocode for our zeroshot point tracking algorithm.

Require: Gaussians, camera K , $[R|t]$, points $p^{(0)}$

Ensure: Tracks $p^{(t)}$, visibility $v^{(t)}$

```
1: for  $t = 0$  to  $T-2$  do
2:   Project centers at  $t, t+1$ ; render  $F^{(t)}$  and  $\alpha^{(t)}$ 
3: end for
4: for  $t = 0$  to  $T-2$  do
5:    $a \leftarrow p^{(t)} + F^{(t)}(p^{(t)})$ ;  $s \leftarrow$  weighted top- $k$  avg
6:    $\omega^{(t)} \leftarrow$  sum of top- $k$  weights
7:   if  $\omega^{(t)} \geq \tau_{\text{vis}}$  then
8:      $p^{(t+1)} \leftarrow (1-\beta)a + \beta s$ ;  $v^{(t+1)} \leftarrow 1$ 
9:   else
10:     $p^{(t+1)} \leftarrow s$ ;  $v^{(t+1)} \leftarrow 0$ 
11:   end if
12: end for
```

A.2. Higher Resolution Figures from the Paper

Higher-resolution versions of Figures 5, 6, and 7 from the main paper are provided below.

A.3. Zeroshot Tracking Examples

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

1. **Zeroshot 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. **Zeroshot tracking on TAP-Vid Kinetics dataset:** Similar to the Davis examples, this video presents zeroshot 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 zeroshot tracking:** To provide a balanced view of our model’s zeroshot capabilities, this video illustrates common failure modes. As discussed in the main paper’s Limitations section, *Video-GMAE*’s zeroshot 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



Figure 9. Higher-resolution version of Figure 5 from the main paper.



Figure 10. Higher-resolution version of Figure 6 from the main paper.

main factors: (i) the pretraining 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 ac-

curately 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 12, 13, 14, 15, and 16 provide static visualiza-

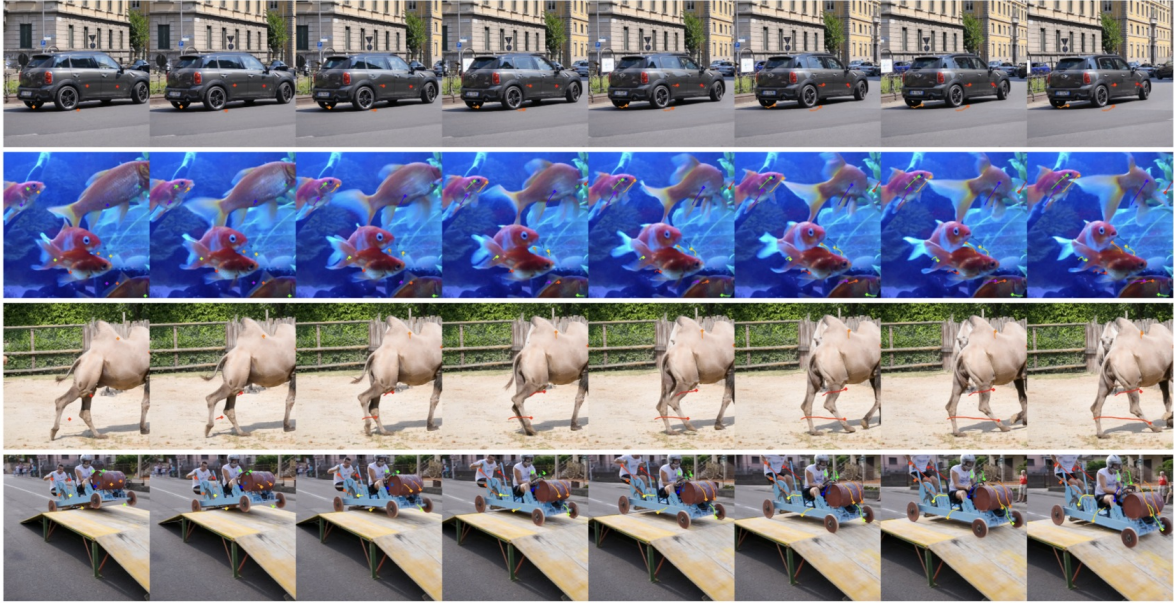


Figure 11. Higher-resolution version of Figure 7 from the main paper.

tions corresponding to these zeroshot 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.

A.4. Qualitative Comparison with GMRW-C

The main paper reports quantitative zeroshot 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 17, 18, 19, and 20. 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-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 zeroshot tracker is well-suited for stable, occlusion-aware tracking, while leaving room for future work to improve fine-detail fidelity under large motions.

A.5. Fine-tuned Tracking Examples

While *Video-GMAE* exhibits strong zeroshot tracking, its representations can be further adapted for enhanced performance 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 zeroshot results, especially in challenging sequences involving occlusions, rapid motion, and multi-object in-

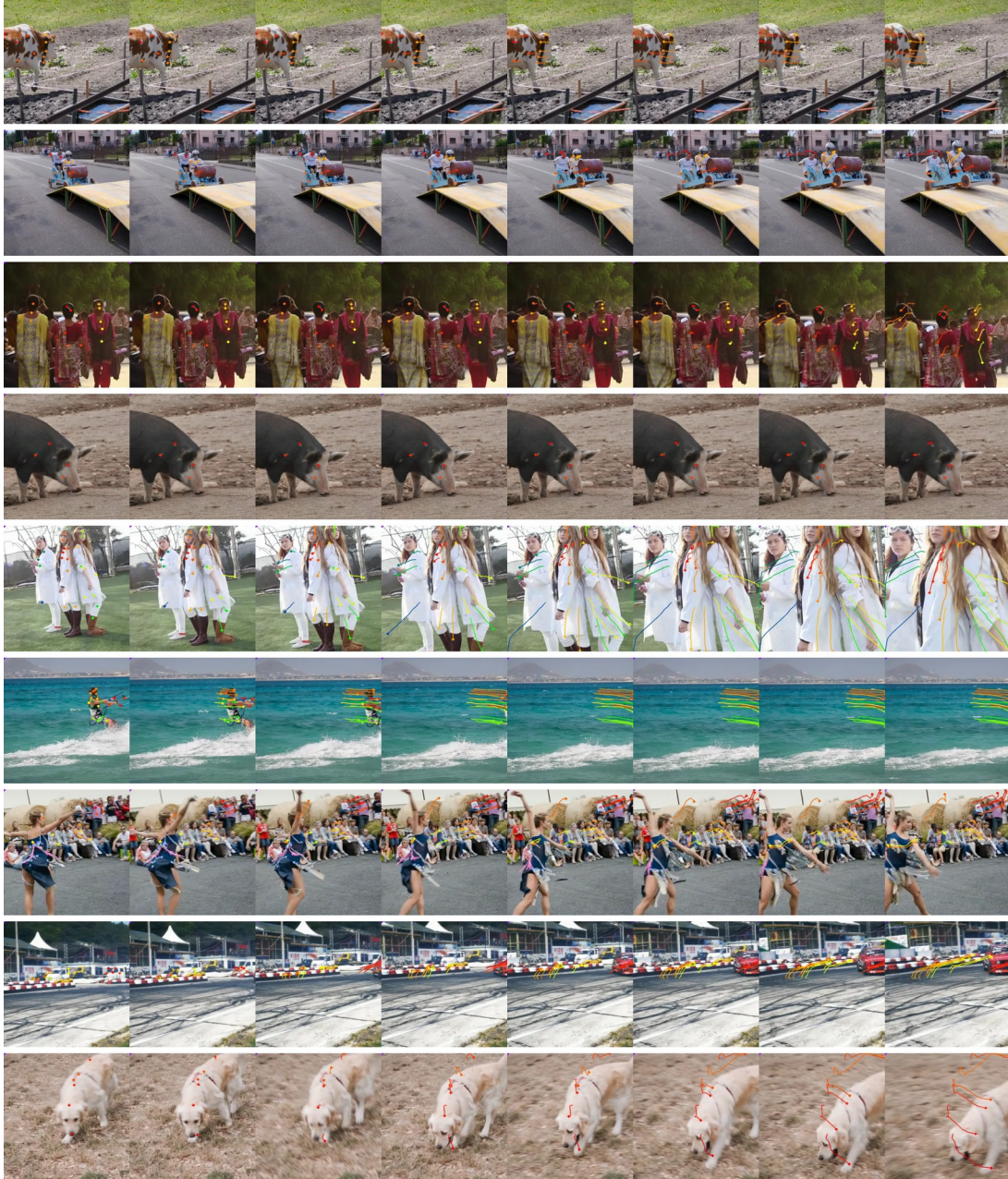


Figure 12. **Zeroshot tracking on TAP-Vid Davis (part 1).** Representative zeroshot point tracking results of *Video-GMAE* on TAP-Vid Davis. Colored points denote tracked locations over time. The examples illustrate that our model can maintain long-range correspondences for deforming objects and complex motions without task-specific supervision.

teractions. 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 21, 22, 23, and 24 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 zeroshot case.

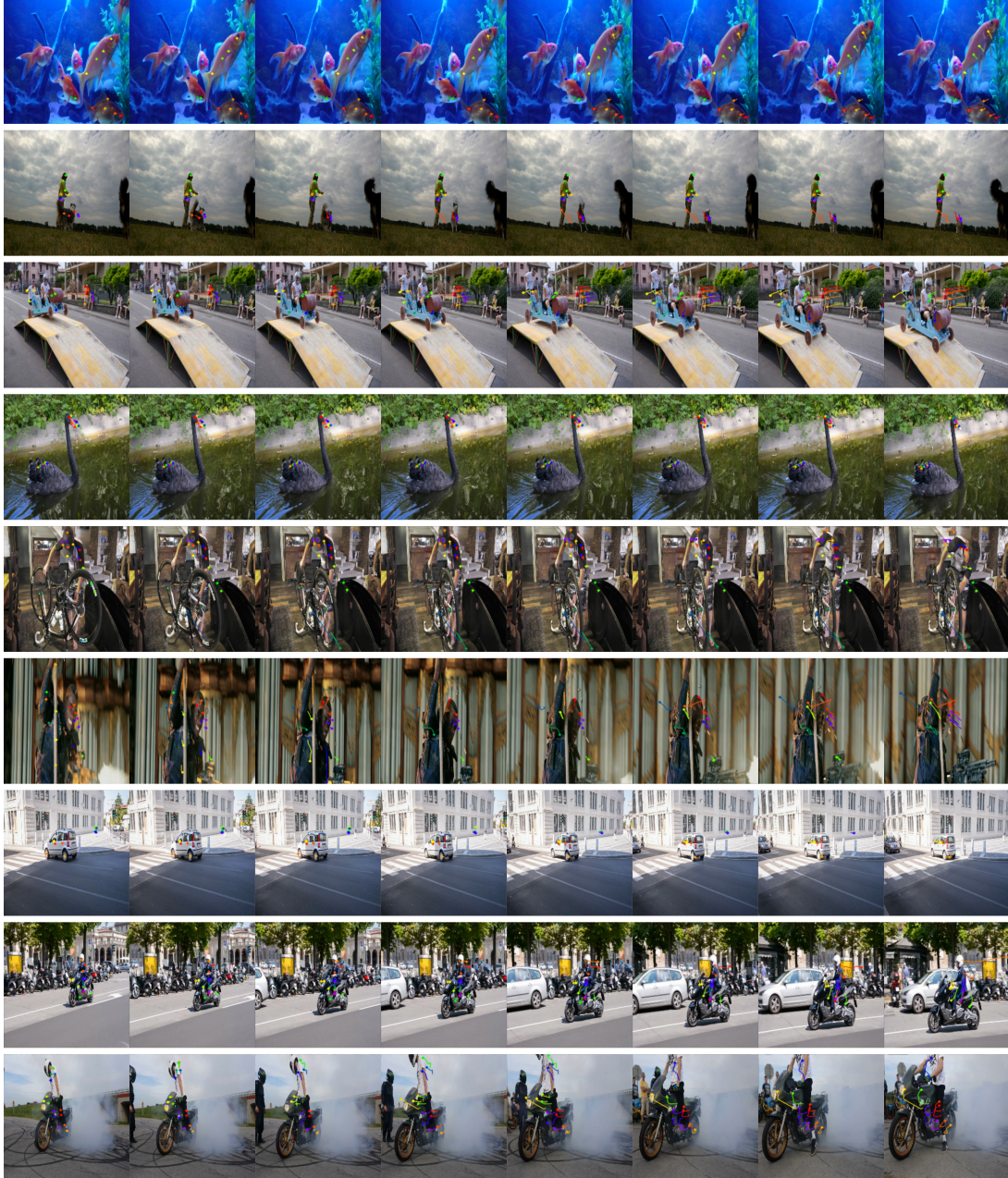


Figure 13. **Zeroshot tracking on TAP-Vid Davis (part 2).** Continuation of the Davis zeroshot tracking examples in Figure 12.

A.6. Identity Switching Analysis

To better quantify temporal consistency, we evaluate identity switching on Kubric, which provides ground-truth object instance labels for each frame. The goal of this analysis is to measure not only whether the tracker remains spatially accurate, but also whether it continues to follow the same underlying object over time rather than drifting to another visually similar instance. This is particularly important in scenes with occlusions or close object interactions, where a tracker may remain locally plausible while silently switching

identity.

For each predicted point track, we assign a ground-truth object instance ID at every frame using the Kubric instance annotations at the predicted point location. We then count an identity switch whenever this assigned instance ID changes between consecutive frames. We report switching rates separately over visible and occluded intervals, yielding visible switching rate (Vis-SR) and occluded switching rate (Occ-SR).

Our results support the qualitative claim in the main paper that *Video-GMAE* better preserves identity, especially under



Figure 14. **Zeroshot tracking on TAP-Vid Kinetics (part 1).** Static visualization of zeroshot 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.

occlusion. On Kubric, *Video-GMAE* achieves a **Vis-SR** of **0.025** and an **Occ-SR** of **0.037**, compared to GMRW-C, which obtains a Vis-SR of 0.055 and an Occ-SR of 0.222. The largest gap appears in the occluded setting, where *Video-GMAE* is substantially less likely to drift to a different object after temporary disappearance. These results suggest that the correspondence-aware Gaussian representation encourages more stable object-level identity over time, complementing the standard tracking metrics reported in the main paper.

A.7. *Video-GMAE* Pretraining Gaussians Rendered

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

1. **Rendered Videos:** These three videos



Figure 15. **Zeroshot tracking on TAP-Vid Kinetics (part 2).** Continuation of the Kinetics zeroshot tracking examples in Figure 14.

(renders1.mp4, renders2.mp4, and renders3.mp4) visualize the output of the *Video-GMAE* decoder during the pretraining phase. For a given input video, the model predicts a set of 3-D 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 25, 26, and 27 show static renderings from the first set of pretraining reconstructions. Figures 28, 29, and 30

show the second set. 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 *Computer Vision – ECCV 2024*, pages 267–284, 2024. 3, 9

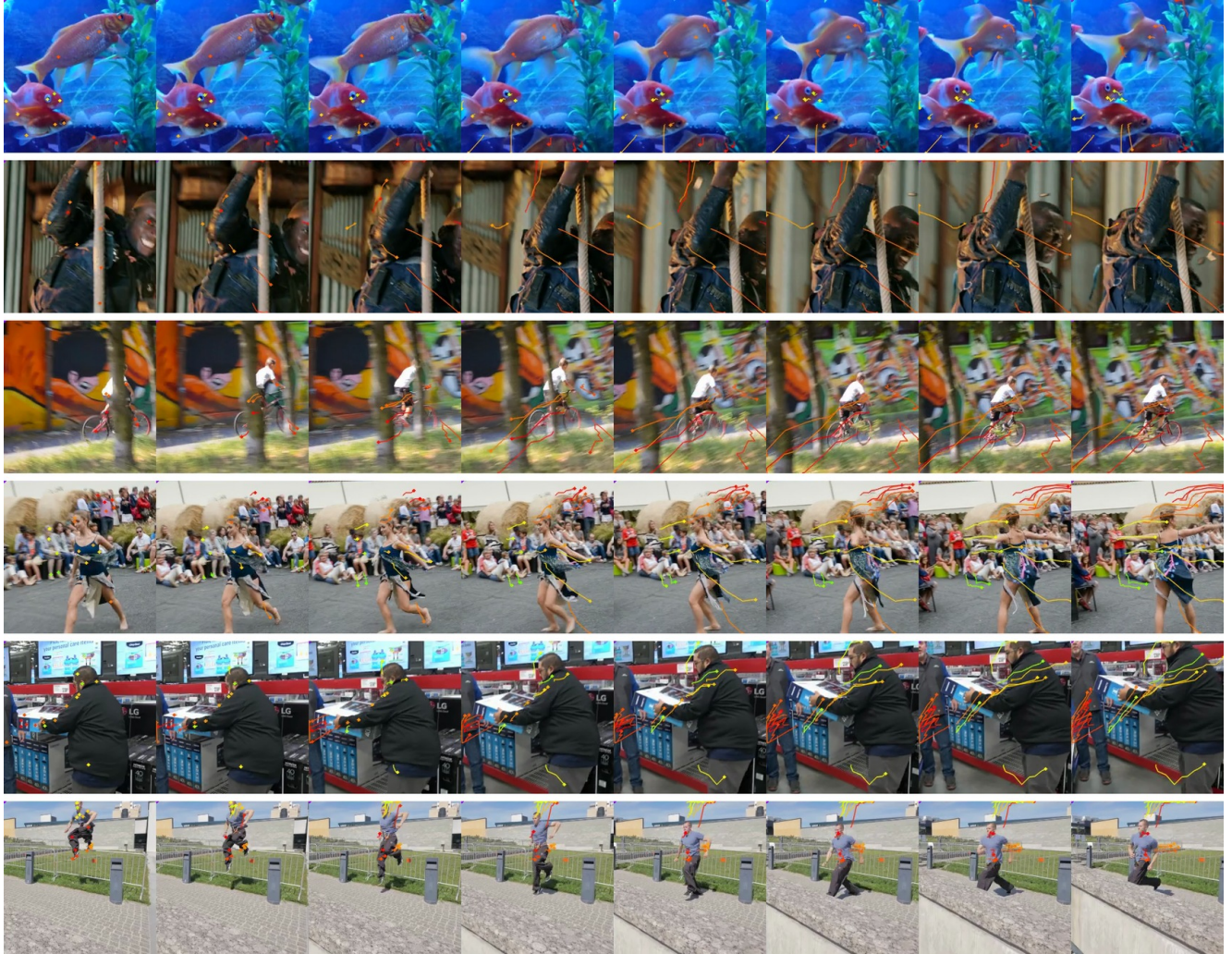


Figure 16. **Zeroshot failure cases.** Examples of challenging zeroshot tracking scenarios where *Video-GMAE* degrades, typically due to complex, high-frequency backgrounds and substantial camera motion. The figure reflects failure modes discussed in the main paper, where the limited 256-Gaussian budget and the static-camera pretraining assumption can reduce tracking accuracy on moving foreground objects.

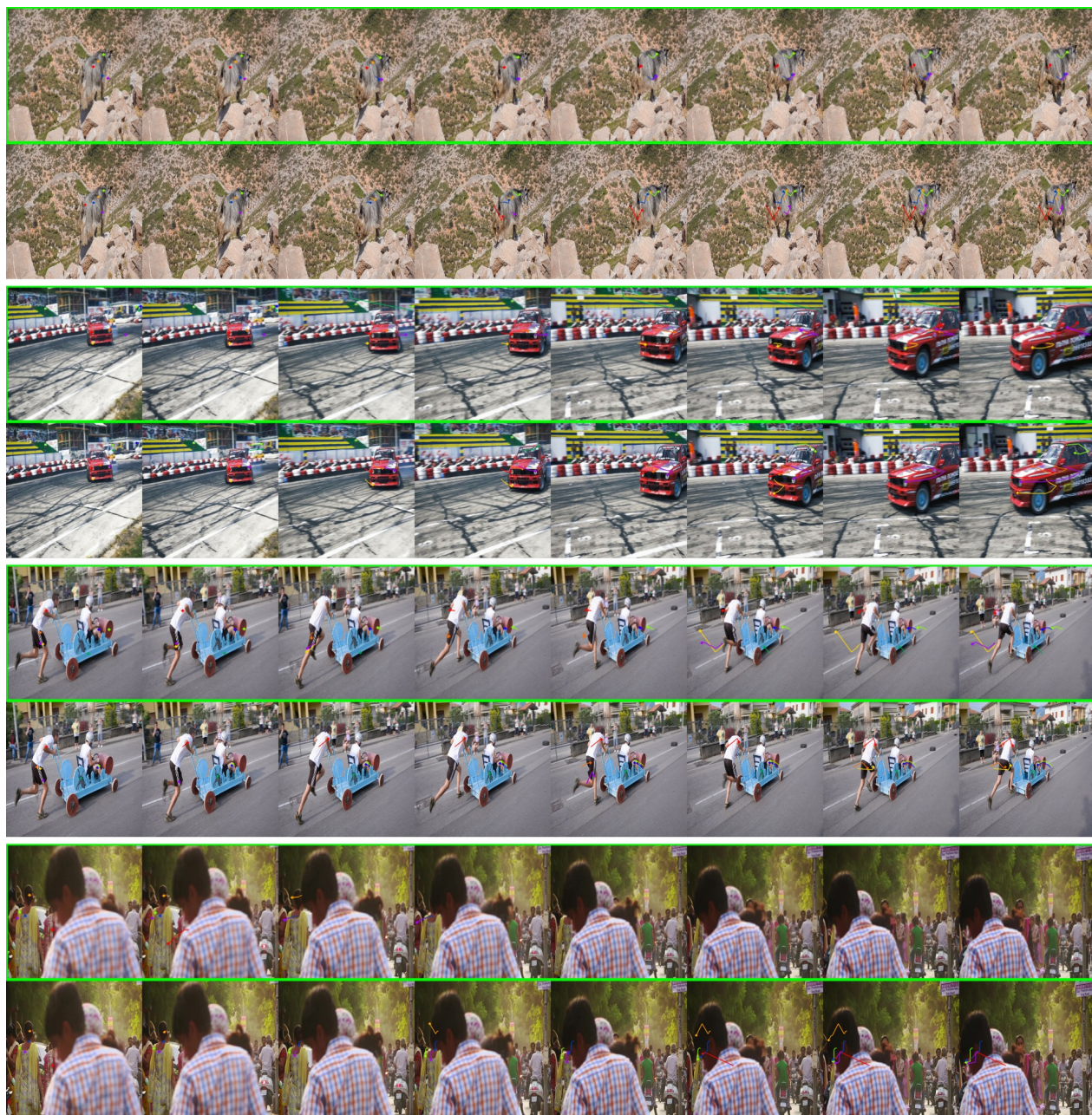


Figure 17. **Qualitative comparison with GMRW-C on TAP-Vid Davis (part 1).** Representative zeroshot tracking sequences comparing *Video-GMAE-zero-shot* (the sequences with the green border) to GMRW-C [1] on TAP-Vid Davis. Each strip shows point trajectories over time for both methods on challenging 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.

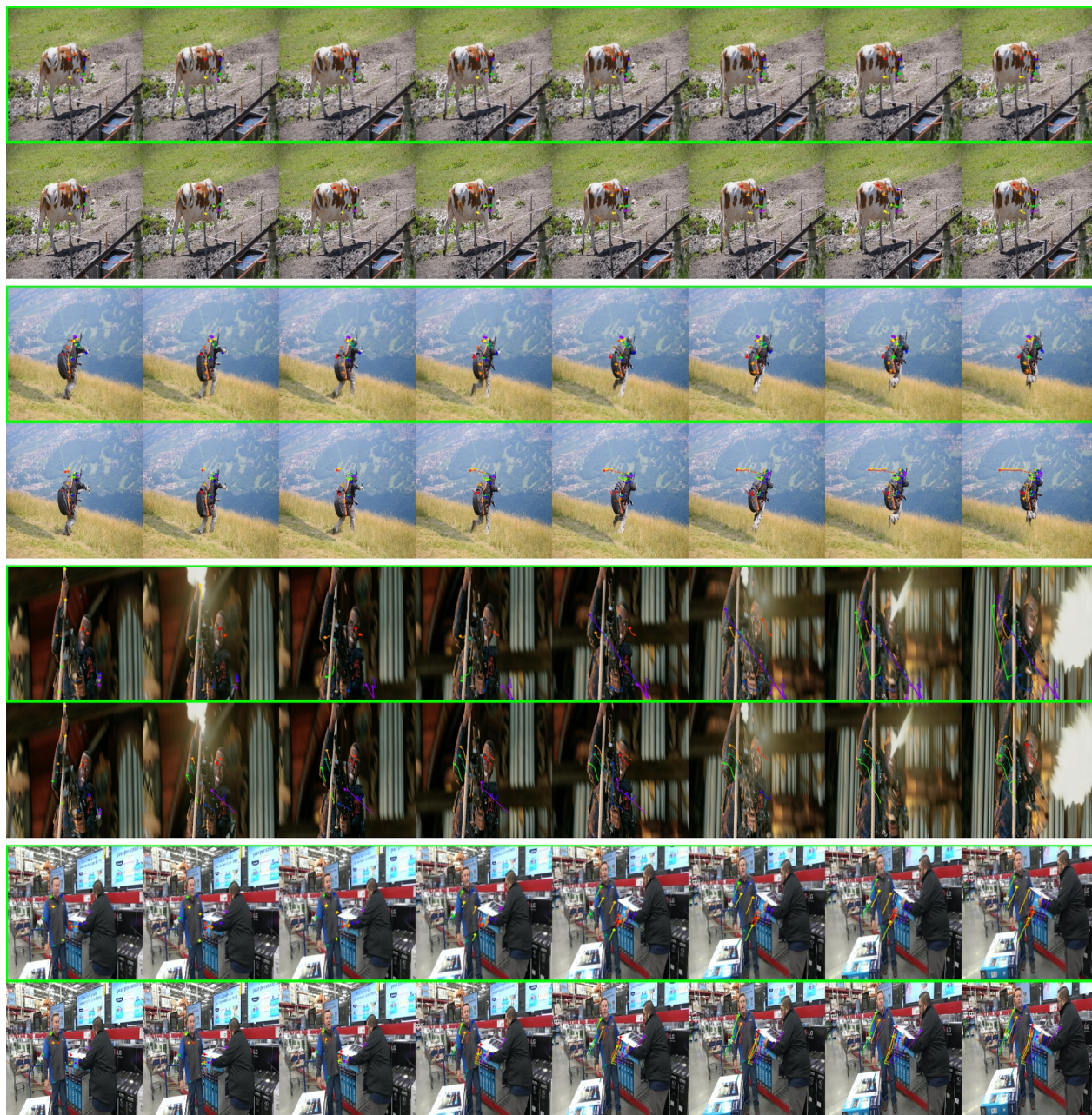


Figure 18. **Qualitative comparison with GMRW-C on TAP-Vid Davis (part 2).** Continuation of the Davis comparison strips in Figure 17.



Figure 19. **Qualitative comparison with GMRW-C on TAP-Vid Kinetics (part 1).** Additional zeroshot tracking comparisons on TAP-Vid Kinetics sequences. While *Video-GMAE-zero-shot* (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.

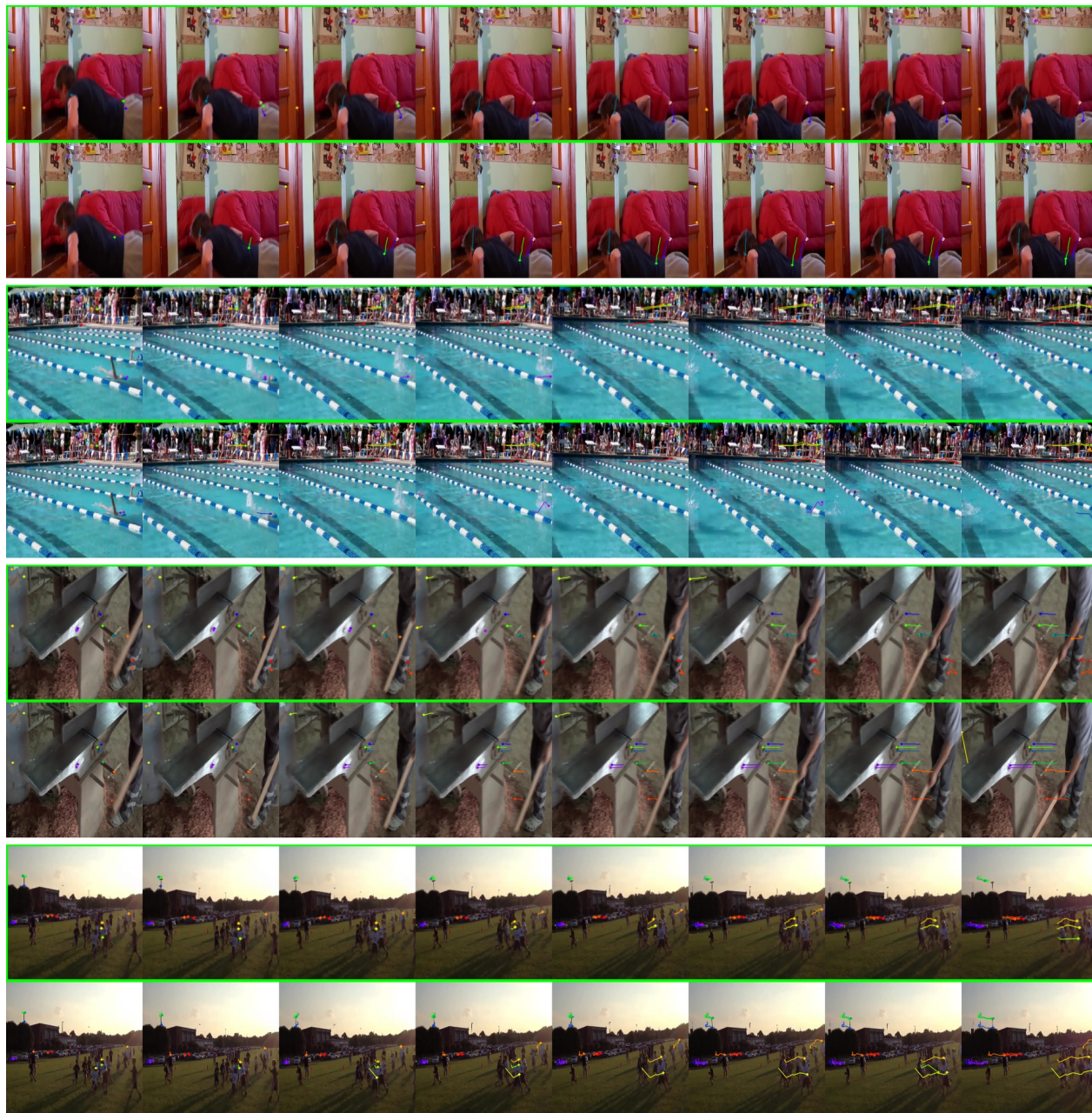


Figure 20. **Qualitative comparison with GMRW-C on TAP-Vid Kinetics (part 2).** Continuation of the Kinetics comparison strips in Figure 19.

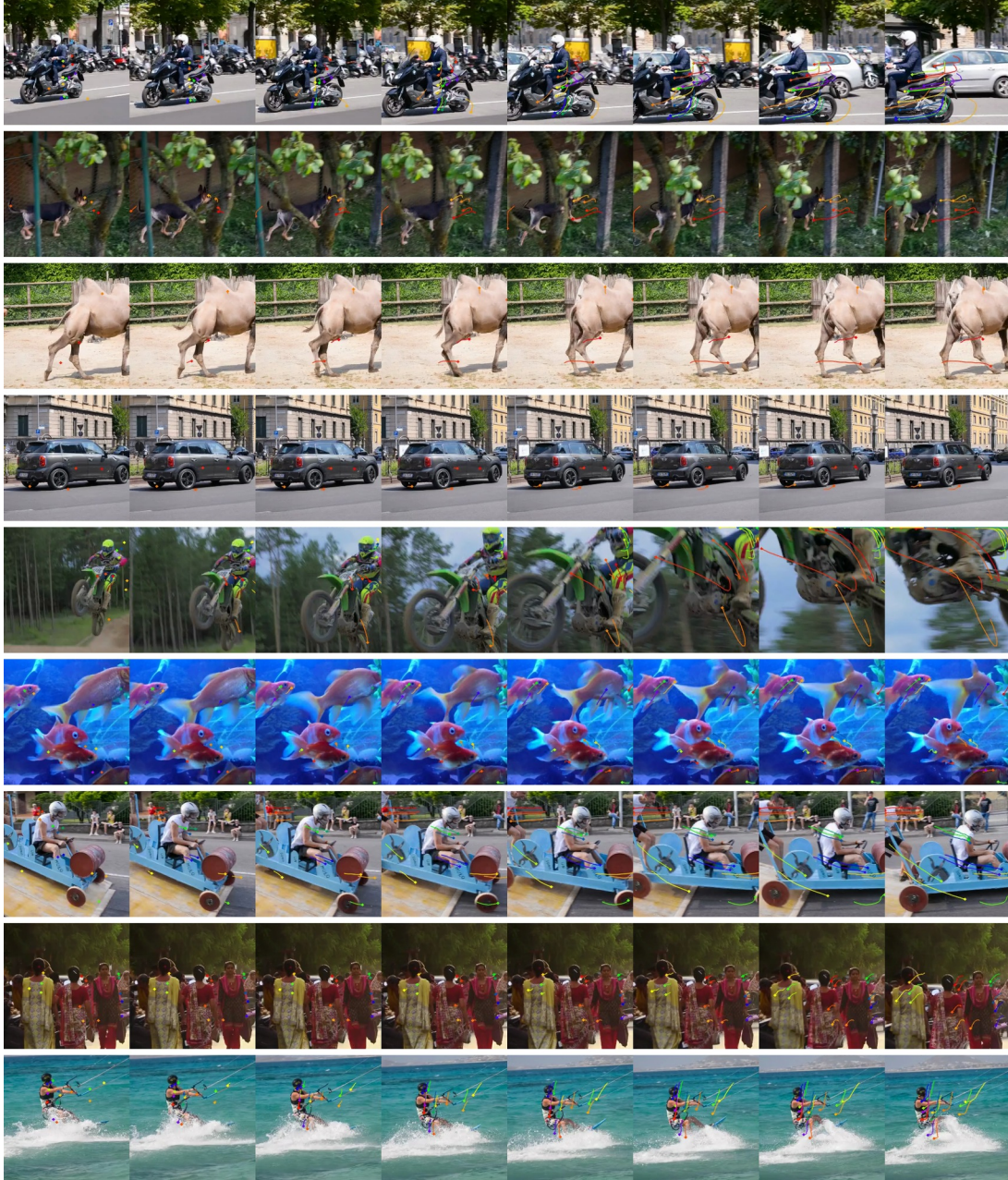


Figure 21. **Fine-tuned tracking on TAP-Vid Davis (part 1).** Still-frame visualization of point tracking with *Video-GMAE* after supervised fine-tuning on TAP-Vid Davis. Compared to the zeroshot sequences in Figures 12, 13, and 16, the fine-tuned model exhibits more precise localization, reduced temporal jitter, and improved handling of occlusions and object interactions.

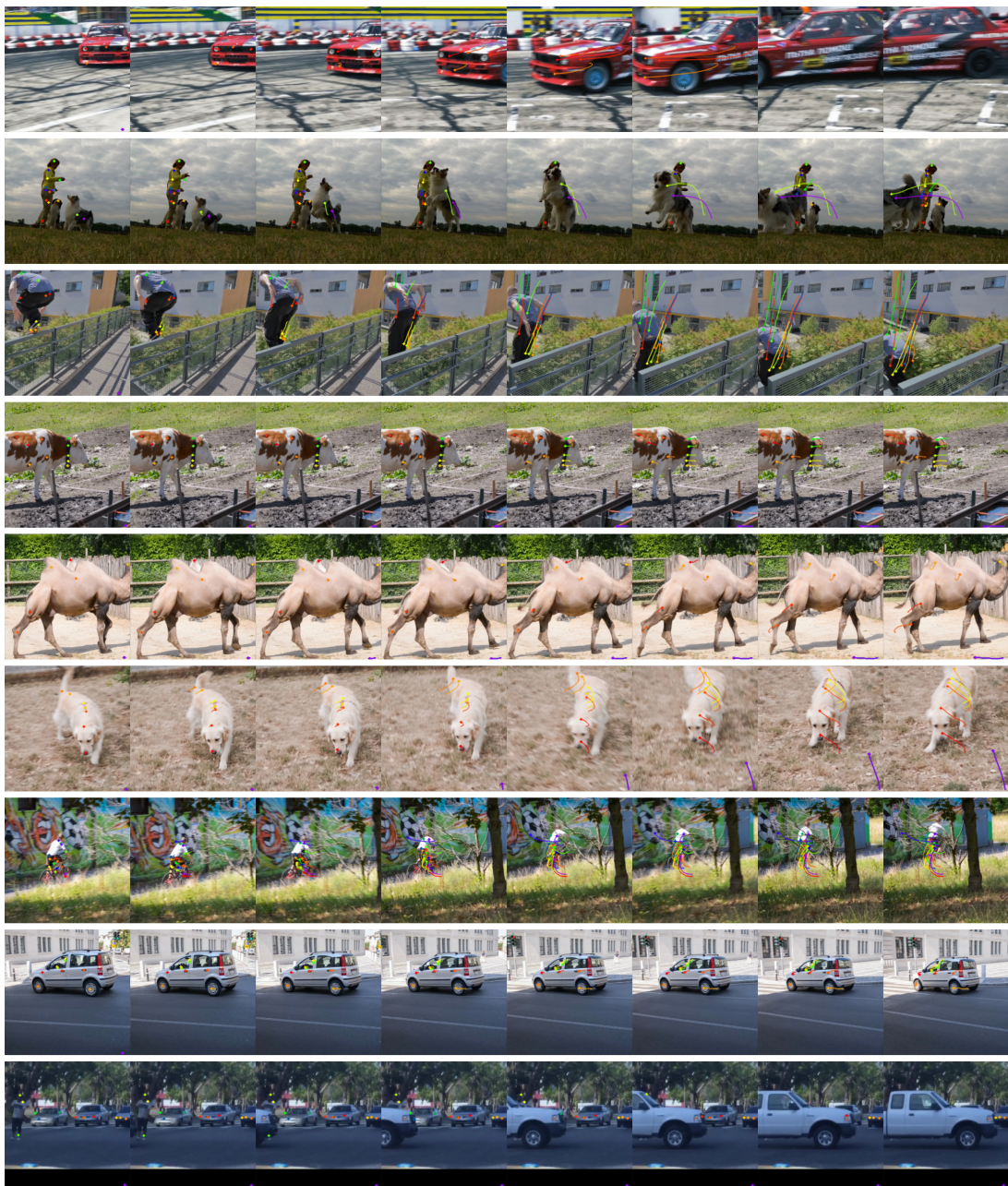


Figure 22. **Fine-tuned tracking on TAP-Vid Davis (part 2).** Continuation of the Davis fine-tuned tracking examples in Figure 21.

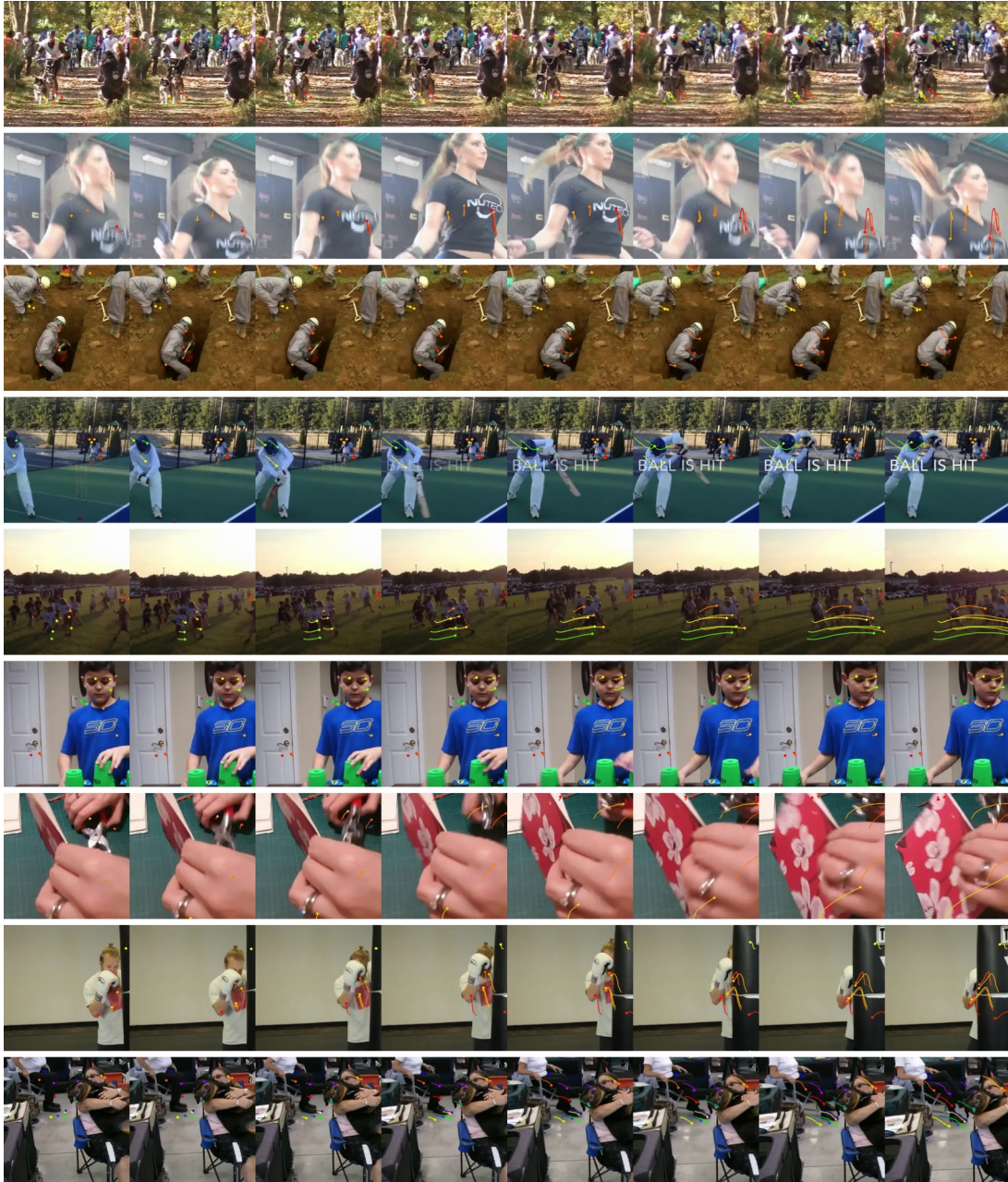


Figure 23. **Fine-tuned tracking on TAP-Vid Kinetics (part 1).** Static example of fine-tuned tracking on TAP-Vid Kinetics. Fine-tuning leads to crisper trajectories and more stable point identities across complex human actions, complementing the qualitative improvements observed in the corresponding videos.

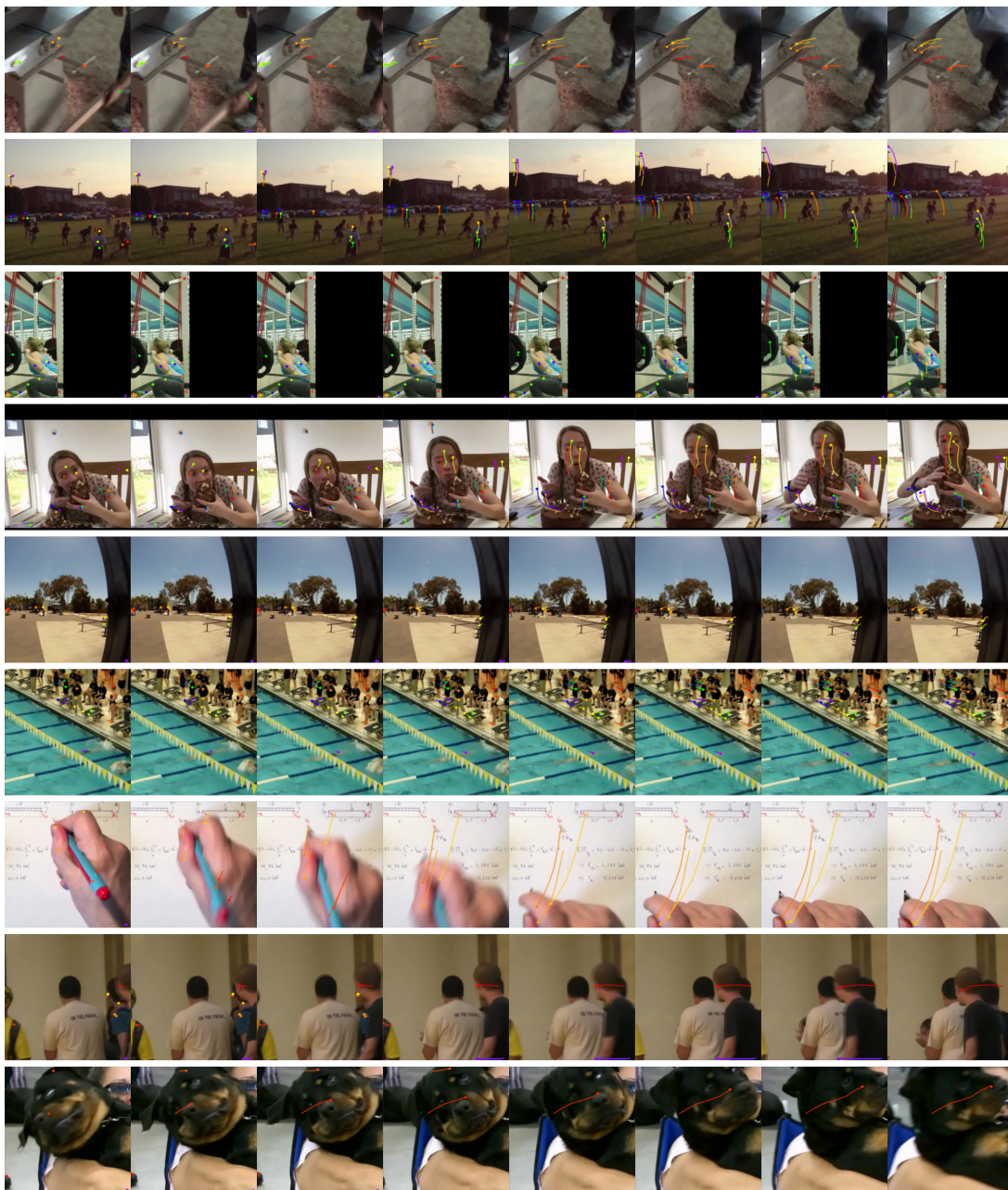


Figure 24. **Fine-tuned tracking on TAP-Vid Kinetics (part 2).** Continuation of the Kinetics fine-tuned tracking examples in Figure 23.

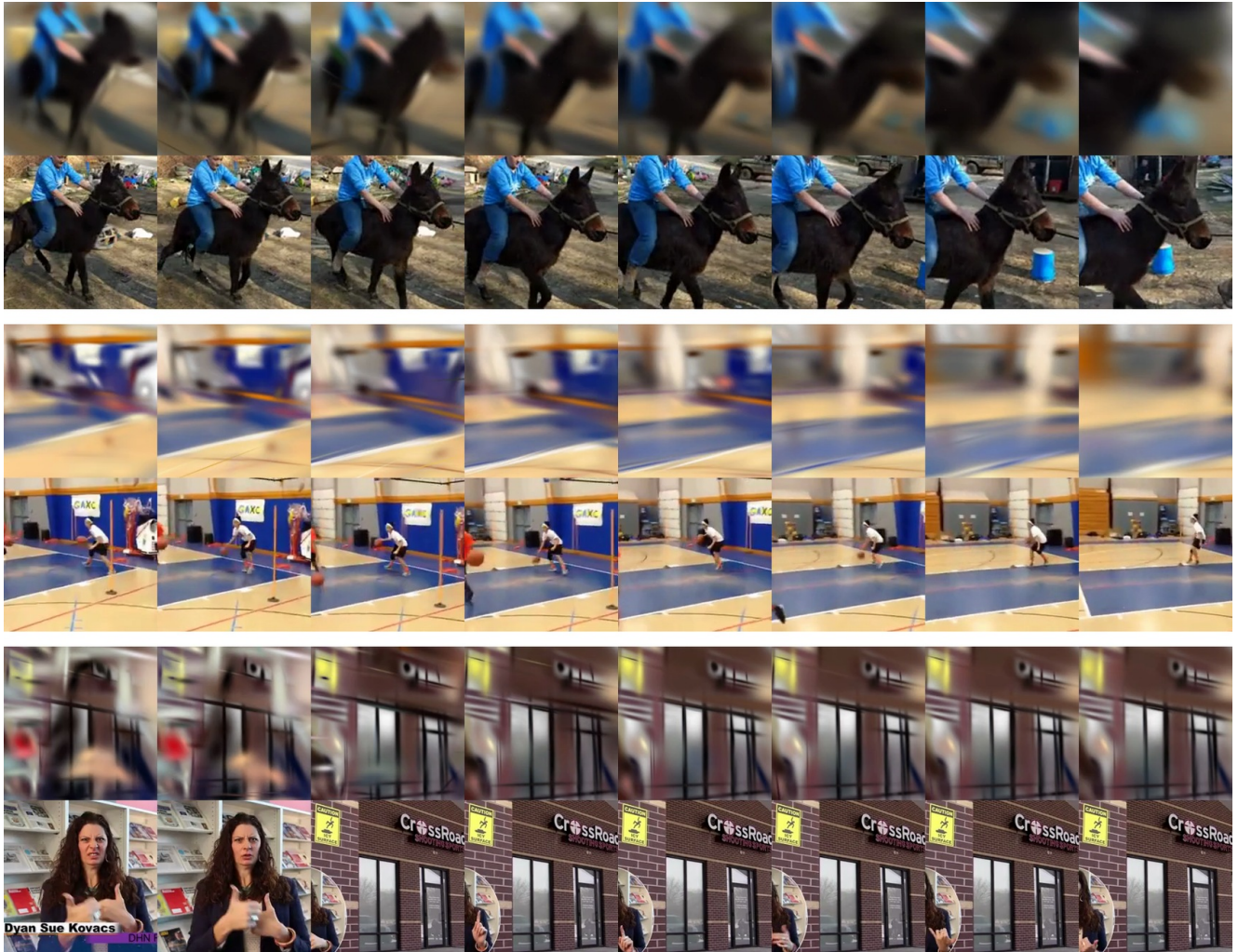


Figure 25. **Video-GMAE** pretraining reconstructions (set 1, part 1). Example frames from the rendered outputs of the *Video-GMAE* decoder during pretraining.

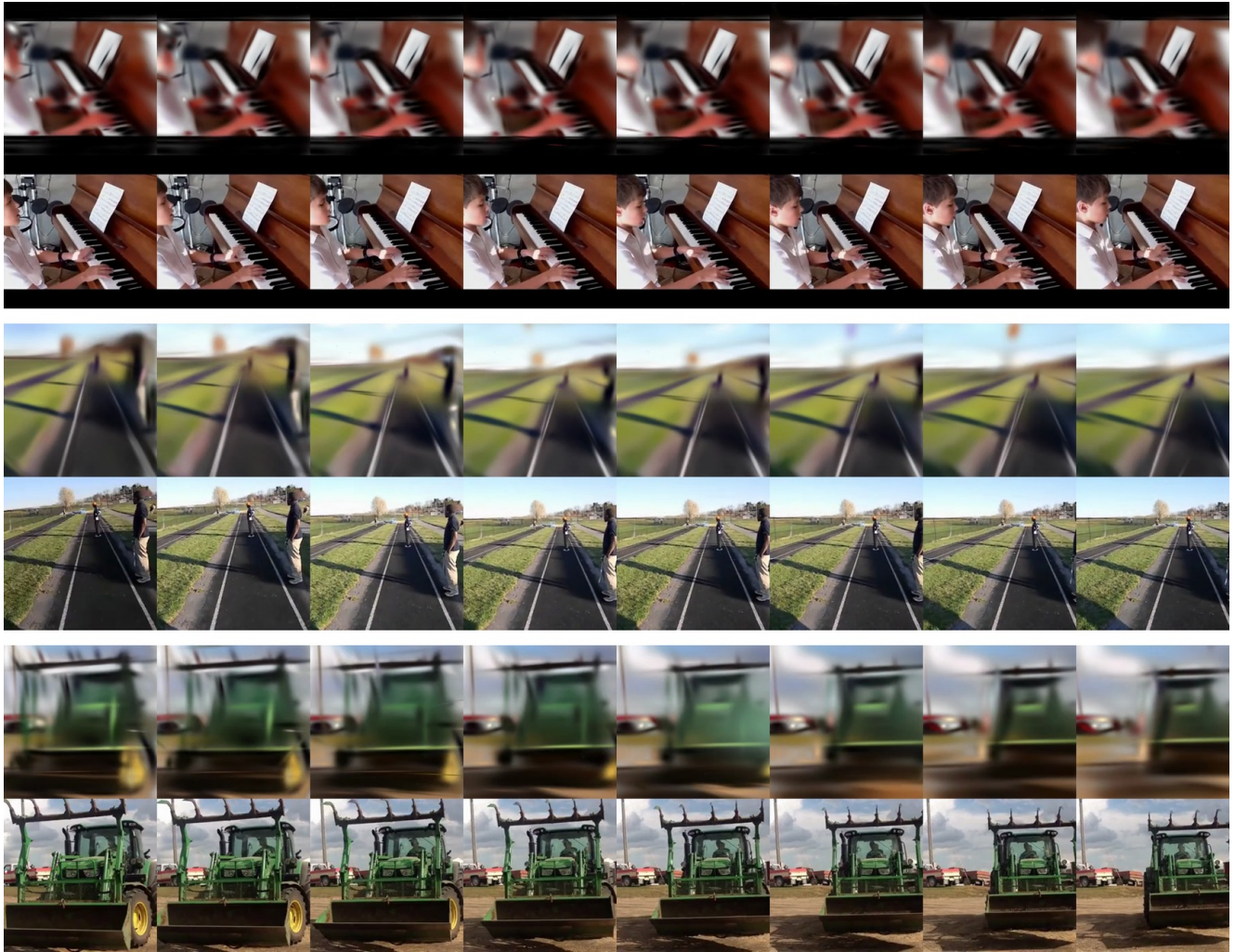


Figure 26. *Video-GMAE* pretraining reconstructions (set 1, part 2). Continuation of the first set of pretraining reconstructions.

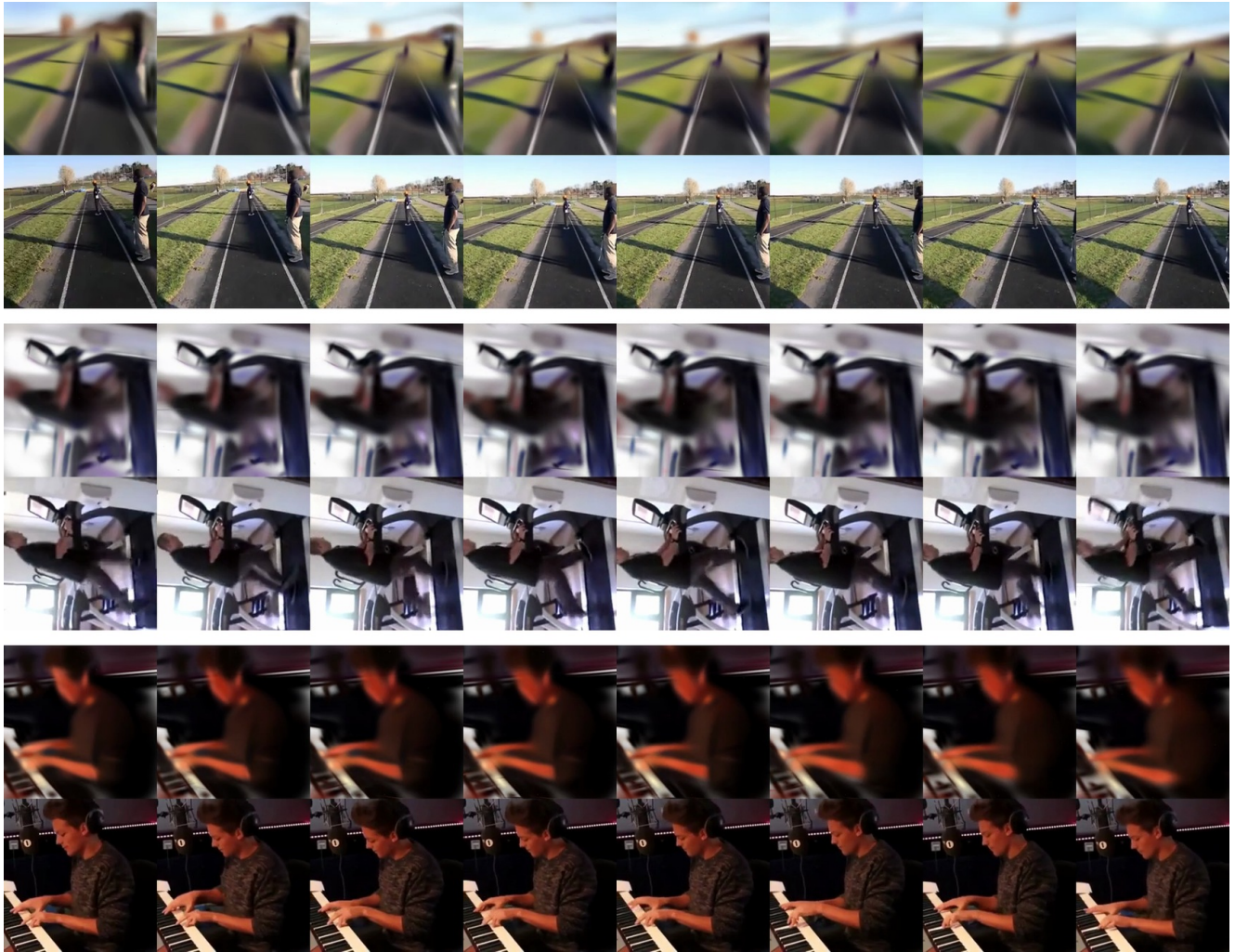


Figure 27. *Video-GMAE* pretraining reconstructions (set 1, part 3). Continuation of the first set of pretraining reconstructions.

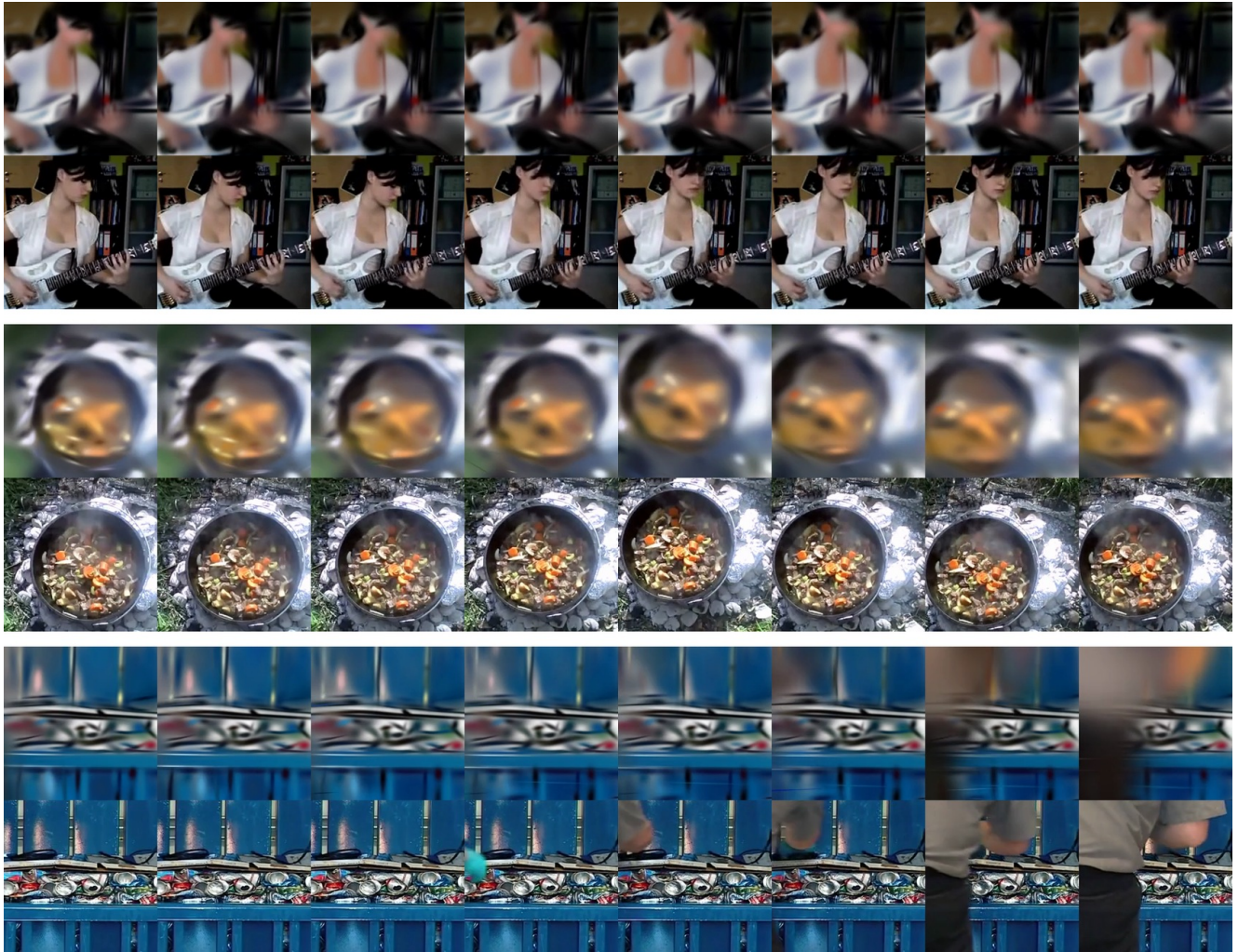


Figure 28. *Video-GMAE* pretraining reconstructions (set 2, part 1). Additional examples of pretraining reconstructions on different sequences.

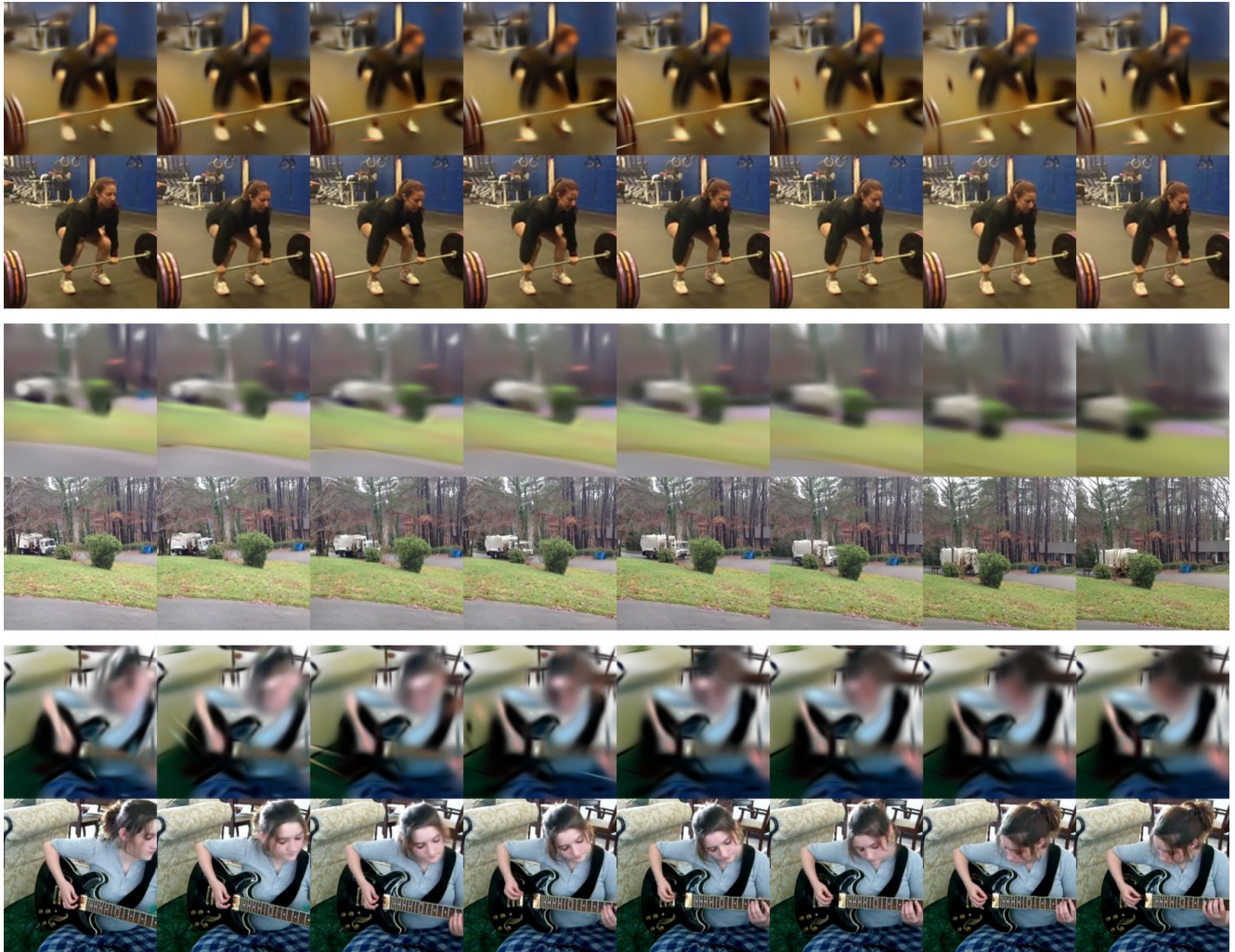


Figure 29. *Video-GMAE* pretraining reconstructions (set 2, part 2). Continuation of the second set of pretraining reconstructions.

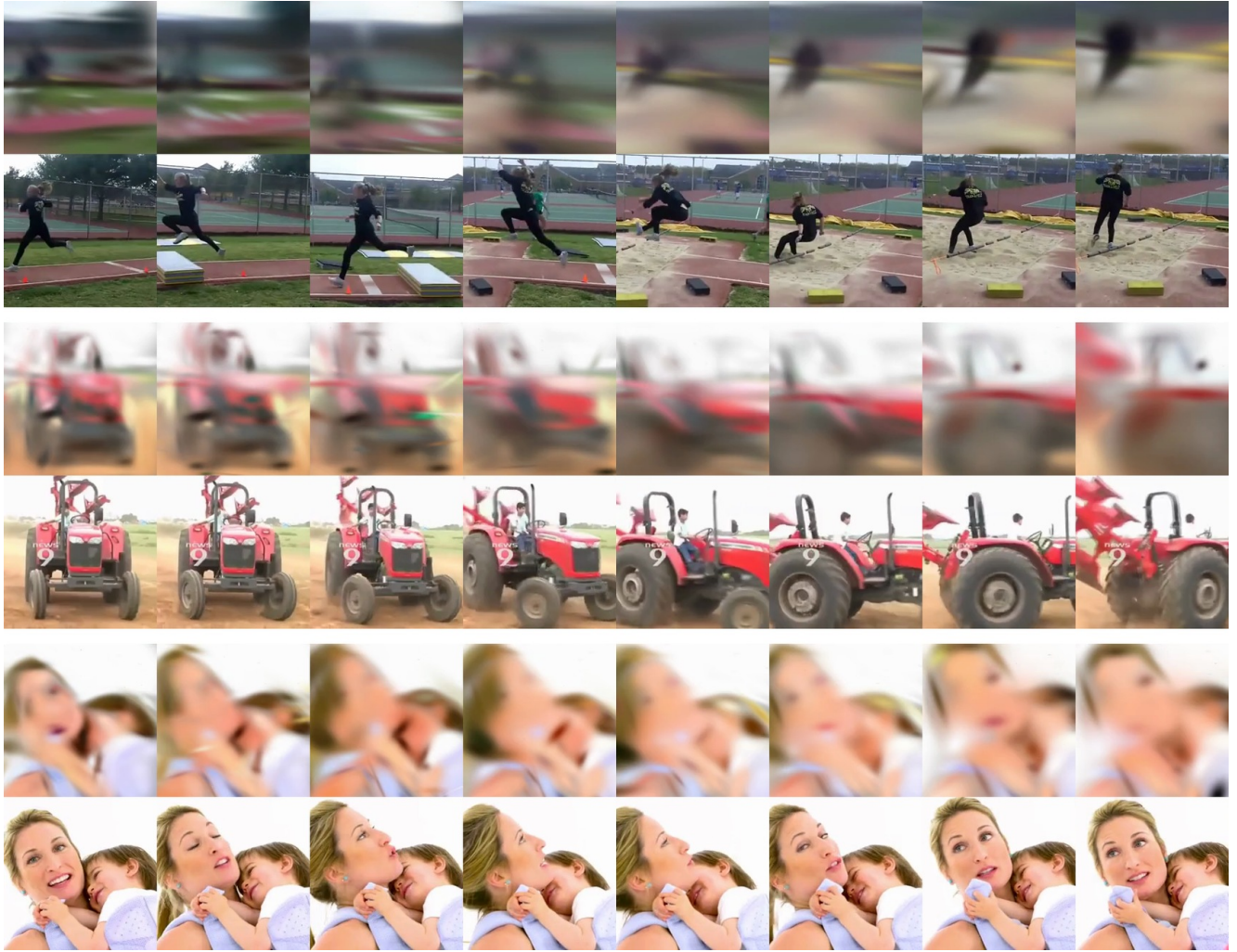


Figure 30. **Video-GMAE pretraining reconstructions (set 2, part 3).** Continuation of the second set of pretraining reconstructions. While the Gaussians capture coarse geometry and dynamics, fine-scale textures and small structures are limited by the 256-Gaussian budget, consistent with the discussion of representation capacity in the main paper.