Is Energy Guidance All You Need? Training-Free Norm Injection for Driving World Models

Xiyan Su, Frank Diermeyer, Markus Lienkamp
Institute of Automotive Technology, Technical University of Munich, Munich Germany

***Abstract*—Driving world models built on large video-diffusion backbones generate realistic scenes but are hard to control: enforcing a traffic norm typically means retraining the backbone or conditioning it on hand-built layouts. We ask whether controllability requires training at all. Our experiment shows that a rectified-flow driving world model, which jointly generates future video and a planned ego trajectory, can have its planned trajectory steered entirely at sampling time by differentiable energy functions that encode driving norms, without knowledge-specific retraining of the diffusion backbone. Concretely, we demonstrate that a world model built on Open-Sora 2.0 MM-DiT backbone can be steered to brake at a counterfactual target by injecting energy guidance at sampling time. However, we find that the generated video does not yet follow the steered trajectory through the backbone's joint self-attention and identify the cross-stream coupling as a crucial requirement for end-to-end-controllable rollouts.**

I. Introduction

Video-diffusion driving world models generate realistic sensor streams and have become increasingly popular for simulation and planning [1, 2, 3]. Their central practical weakness is controllability: a model that produces a beautiful, crisp scene in which the ego vehicle drifts across a solid line, exceeds the speed limit, or collides with another agent is not useful for safety-critical use. However, state-of-the-art mitigation strategies remain costly to scale. One either fine-tunes the backbone whenever a new constraint must be complied with, or one conditions the model on hand-constructed layouts baked in during training, such as maps and agent bounding boxes. Both require high-quality data and couple new driving knowledge to a training run. Neither lets a deployed model add or re-prioritize a rule during sampling time.

Hypothesis: We ask whether controllability requires training at all. Training-free guidance has shown that a frozen diffusion prior can be steered toward samples that satisfy an external objective by perturbing the sampling trajectory with the gradient of a differentiable criterion [4, 5, 6, 7]. We bring this view to world models for autonomous systems: as an application example, we take driving norms as energy functions evaluated on the generated ego trajectory, and inject them only at sampling time into a frozen rectified-flow backbone. The backbone is trained completely without knowledge-specific conditioning (see Fig. 1).

Does video match trajectory? A driving world model is valuable precisely because it generates the video, not just a trajectory. Therefore, the question we center on is whether steering the ego trajectory with an energy term also guides the video to match through the backbone's joint denoising. We find that the trajectory is steered readily, but the video does not yet follow. We trace this back to the model architecture and identify the cross-stream coupling as the main cause.

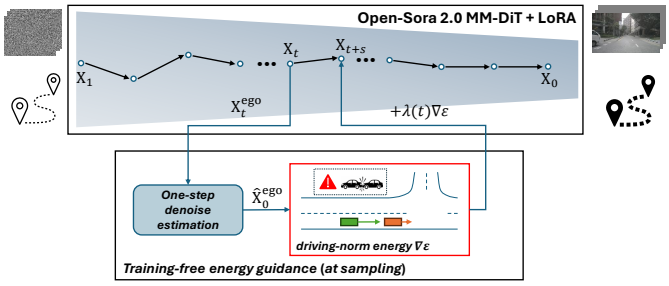

Fig. 1. **Overview.** A frozen video world model (Open-Sora 2.0 MM-DiT + LoRA) jointly denoises future video and an ego trajectory conditioned only on observed past without agent or map at training. Driving norms enter at sampling as energies on the clean ego estimate $\hat{\mathbf{x}}_0$ and are applied as a velocity correction to the ego stream. We study the question of whether the generated video is also steered through the backbone's joint self-attention.

Contributions: Our contributions are three-fold:

- We frame driving-norm compliance as training-free energy guidance on a frozen rectified-flow world model that jointly generates future video and a planned ego trajectory, and we show that all the knowledge can be added only at the sampling time and how.
- We test our hypothesis on a world model built on an Open-Sora 2.0 backbone [8] and show that a training-free energy guidance steers the ego trajectory to a counterfactual braking target at sampling time, simulating a rear-end collision driving norm.
- We identify the central obstacle to end-to-end control: the steered ego trajectory does not propagate to the generated video under the current model architecture, i.e., double-stream rectified flow model aligned through joint self-attention. Based on this finding, we propose a possible solution to that problem.

II. Related Work

Training-free diffusion guidance. A frozen diffusion prior can be steered toward an external objective by perturbing the sampling trajectory with the gradient of a differentiable criterion evaluated on the predicted clean sample $\hat{\mathbf{x}}_0$ [4, 5, 6, 9]. Energy- and composition-based variants combine several criteria [10, 11]. These methods are known to drift off the data manifold and to suffer when $\hat{\mathbf{x}}_0$ averages distinct modes [6, 12]. We apply a similar method to steer the ego plan of a frozen video world model rather than a low-dimensional trajectory model.

Guidance for flow matching. We use a rectified-flow model as our backbone [13, 14, 15], whose deterministic ODE velocity field requires a different guidance treatment than score-based SDEs. Feng et al. [7] formalize guidance for flow matching. We build on this and apply it to a world model. Concurrent work couples flow matching with a trained energy model for ego trajectory planning [16]. We instead inject energies training-free at sampling time without retraining.

Controllable video diffusion. Training-free and conditional control of video diffusion has targeted motion, trajectories, and frame appearance [17, 18, 19, 20]. These control how the video looks or moves. We instead inject semantic driving norms into a video world model.

Robotic World Models for Driving. Video-generation models produce realistic rollouts and, increasingly, actions for autonomous driving and general robotics [1, 2, 3, 21, 22]. In this field, control is often achieved by training-time structured conditioning (maps, layouts; e.g. GAIA-2 [21]) or by generate-then-select over sampled futures (e.g. Drive-WM [3]). Both approaches require additional cost. Recent models jointly generate the ego trajectory alongside the video [23, 24], one on a rectified-flow backbone like ours [24]. Yet the trajectory is produced by learned conditioning, not by being steered at sampling time. We steer the sampler of a general driving world model without knowledge-specific conditioning and retraining.

Rule-aware trajectory generation. Closest to our motivation, CTG and its language-guided successor compile traffic rules into differentiable guidance for trajectory diffusion [25, 26], and Diffusion-Planner applies training-free guidance with collision, comfort, and lane energies to an ego planner [27]. We differ in the model architecture. We use a frozen world model that jointly generates future driving video and planned ego trajectory. Additionally, we study whether the guided trajectory also steers the generated video through joint self-attention between the video and the ego trajectory.

Positioning. Our work mainly differentiates itself on the conjunction: a rectified-flow world model guided by training-free sampling-time energy of its jointly generated ego plan. The closest neighbors fall on either side: training-free energy-guided ego-plan optimization exists, but for trajectory-only models [27]. World models that jointly denoise video and an ego plan exist, but inject knowledge only at training time [23, 24]. None steers a jointly generated ego trajectory at sampling time, and we further examine whether doing so moves the generated video.

III. Method

A. Background: rectified-flow sampling

We use a rectified-flow model as our backbone. For a clean sample $\mathbf{x}_0$ and noise $\mathbf{x}_1 \sim \mathcal{N}(0, I)$, the interpolation and learned velocity field are

$$\begin{aligned} \mathbf{x}_t &= (1-t)\,\mathbf{x}_0 + \big(\sigma_{\min} + (1-\sigma_{\min})t\big)\mathbf{x}_1, \\ \mathbf{v}_\theta(\mathbf{x}_t, t) &\approx (1-\sigma_{\min})\mathbf{x}_1 - \mathbf{x}_0, \end{aligned} \tag{1}$$

and sampling integrates the ODE $\dot{\mathbf{x}}_t = \mathbf{v}_\theta$ from $t{=}1$ to 0. Here, $\sigma_{\min}$ is a small noise floor at the data end ($t{=}0$): flow matching models data as a narrow Gaussian of width $\sigma_{\min}$ rather than a point mass [14]. This keeps the score and the clean-sample inversion well-defined. Pure rectified flow [13] is the $\sigma_{\min}{=}0$ limit, and we use $\sigma_{\min}{=}10^{-5}$. With $\sigma_{\min}{\to}0$ the model admits a one-step estimate of the clean sample,

$$\hat{\mathbf{x}}_0(\mathbf{x}_t, t) \;=\; \mathbf{x}_t - t\,\mathbf{v}_\theta(\mathbf{x}_t, t). \tag{2}$$

B. Training-free energy guidance on the ego trajectory

Our platform is a frozen video-diffusion world model built on a post-trained Open-Sora 2.0 MM-DiT backbone that jointly denoises a future video and a planned ego trajectory $\mathbf{x}_0^{\text{ego}} \in \mathbb{R}^{T\times d}$ conditioned on the observed past without agent and map information (see Fig. 2, full architecture and fine-tuning details in Appendix B). The energy guidance is supplied only at sampling time (see Fig. 1). We inject driving knowledge only through a differentiable energy $\mathcal{E}(\cdot)$ evaluated on the predicted clean ego trajectory. At each sampling step, we form $\hat{\mathbf{x}}_0$ via Eq. (2), take an energy-descent step in clean-trajectory space, and convert it back to a velocity correction:

$$\begin{aligned} \hat{\mathbf{x}}_0' &= \hat{\mathbf{x}}_0 - \lambda(t)\,\nabla_{\hat{\mathbf{x}}_0}\mathcal{E}(\hat{\mathbf{x}}_0), \\ \mathbf{v}' &= \frac{\mathbf{x}_t - \hat{\mathbf{x}}_0'}{\max(t,\, t_{\min})}, \end{aligned} \tag{3}$$

where $\mathcal{E}$ is the driving-norm energy evaluated on the one-step estimate (see Section III-C for an example), and $\lambda(t)$ is an asymmetric schedule that is zero near $t{=}0$ and $t{=}1$. Guidance off near $t{=}1$ lets the prior establish global structure before it is refined, while off near $t{=}0$ avoids the $1/t$ blow-up of Eq. (3).

The one-step estimate $\hat{\mathbf{x}}_0$ is unreliable early in sampling, as the high-variance guess can sit far from the target, so the energy gradient evaluated on it can be large and poorly directed. An unclipped step launches the trajectory off-manifold. We therefore cap the per-step guidance correction at a fixed multiple of the backbone's own velocity (a scale-free clip), in line with standard training-free guidance practice [4, 5].

C. Energy for driving

We integrate the driving norms as differentiable energy $\mathcal{E}(\hat{\mathbf{x}}_0)$ on the predicted clean ego trajectory. The guidance step (3) steers the plan toward low-energy trajectories. In this paper, we use a single energy: an MSE attraction toward a counterfactual braking target in Eq. (4). Similar to other training-free guidance [6, 12], Eq. (3) biases the generative

distribution toward low-energy samples rather than sampling exactly from an energy-tilted target, and it relies on the one-step estimate $\hat{\mathbf{x}}_0$, which is least accurate early in sampling and under multimodality, further discussed in Section V.

We take the notation from Section III-B and Fig. 2: the generated planned ego trajectory $\mathbf{x}_0^{\text{ego}} = (\mathbf{x}, \mathbf{y}, \theta, \mathbf{v}) \in \mathbb{R}^{T \times d}$ $(T = 40, d = 4)$

$$\mathcal{E}(\hat{\mathbf{x}}_0) = \frac{1}{|T|} \sum_{k \in T} \Big[w_{xy} \left\| (x_k, y_k) - (x_k^\star, y_k^\star) \right\|^2 + w_\theta \left(\theta_k - \theta_k^\star\right)^2 + w_v \left(v_k - v_k^\star\right)^2 \Big], \tag{4}$$

where $(x_k, y_k, \theta_k, v_k)$ are the per-frame components of the predicted clean ego trajectory $\hat{\mathbf{x}}_0$ (the text stream channel of Fig. 2), $(\cdot)^\star$ is the counterfactual target trajectory, which in this case simulates a slow lead vehicle in front of the ego, T the valid future frames, and w_{xy}, w_θ, w_v per-channel weights.

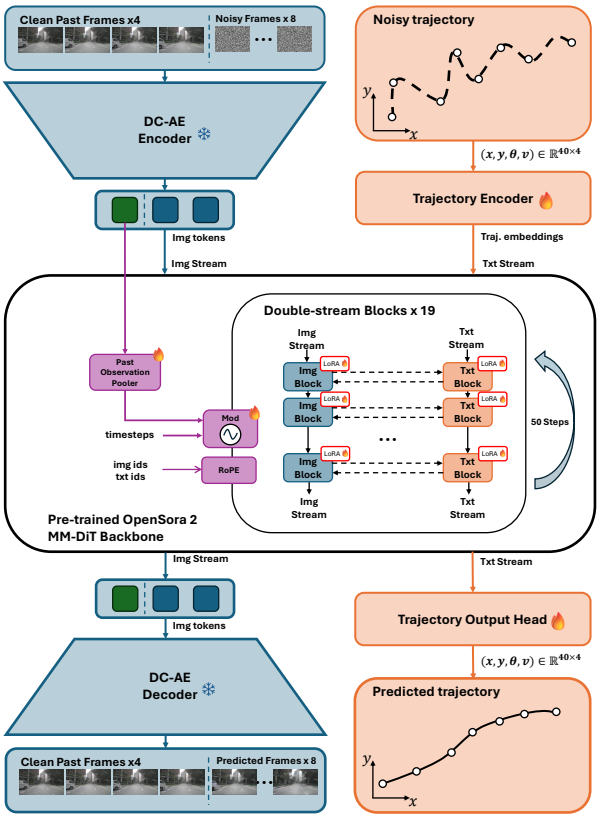

Fig. 2. **The MM-DiT backbone at sampling time without energy guidance**. The observed past and the noised future frames are encoded by the frozen DC-AE into image tokens; the noised ego trajectory (x, y, θ, v) enters the planning (text) stream. The two streams are denoised jointly across the 19 double-stream blocks over 50 steps, modulated by the timestep and the past-observation pooler, with RoPE positional ids, and then decoded into the predicted future frames and the predicted trajectory. The model is post-trained unguided as a driving world model: frozen modules are marked ❄; modules adapted with LoRA or trained from scratch are marked 🔥.

IV. Experiments

We apply this training-free energy guidance to the full video world model, which generates a future driving video and the ego trajectory using a counterfactual target energy that asks the planned trajectory to brake, simulating a rear-end safety driving norm (a slow lead vehicle ahead of the ego). The guidance steers the ego trajectory effectively: across 35 sampled clips, the guided trajectory ends a mean 1.3 m from the brake target while the unguided trajectory ends 13.6 m away of it (Fig. 3, left), showing that a single sampling-time energy steers the trajectory of the full video world model. Detailed results for single clips and per-frame trajectory-video comparison are evaluated and shown in Appendix C (see Fig. 4, Fig. 5, and Fig. 6).

The open question is whether the video follows. Since the video and the trajectory are denoised jointly, the video should follow the steered trajectory through the joint self-attention mechanism from the backbone. However, we discovered that the video does not follow: at the same operating point, the per-frame optical flow of the guided video is statistically indistinguishable from the unguided one (a $\lesssim 2\%$ change, within the across-clip spread; Fig. 3, right), even though the trajectory has braked. We suspect that the problem comes from the fact that the 38 single-stream blocks are currently bypassed in our modified model, where the video and trajectory streams would be processed jointly (see Appendix B1). Our current proposed solution is to re-route the video and trajectory streams back to the 38 single-stream blocks to align the video and trajectory streams.

V. Discussion, Limitations, and Ongoing Work

Limitations of the mechanism. Our guidance in Eq. (3) biases the prior toward low-energy samples, which inherits the known limitations of training-free guidance: the one-step estimate $\hat{\mathbf{x}}_0$ is least accurate early in sampling and under multi-modality [6, 12]. Furthermore, ODE sampling lacks the stochastic self-correction that helps score-SDE samplers return to the data manifold [28]. The asymmetric schedule in Section III-B mitigates but does not eliminate these known limitations completely. We also stabilize the guidance by clipping the per-step correction early in the stage, as the energy gradient is unreliable and $\hat{\mathbf{x}}_0$ can be large (see Appendix A2).

Ongoing work: making the video follow. We established the formulation and showed that sampling-time energy steers the ego trajectory of the world model. The current problem comes from the cross-stream control: the steered planned trajectory does not yet move the video, because the ego interacts with the video only in the double-stream blocks through the joint self-attention. We plan to fix the cross-stream problem by routing both streams through the single-stream blocks from the Open-Sora 2.0 backbone, where the two streams are processed jointly. This could, in turn, re-anchor the ego trajectory to overlap with the generated video. Furthermore, we plan to run further evaluation on the generation realism (e.g., FVD [29]) and rule compliance.

Ongoing work: composing multiple driving norms. A single energy already steers the plan; our final goal is to

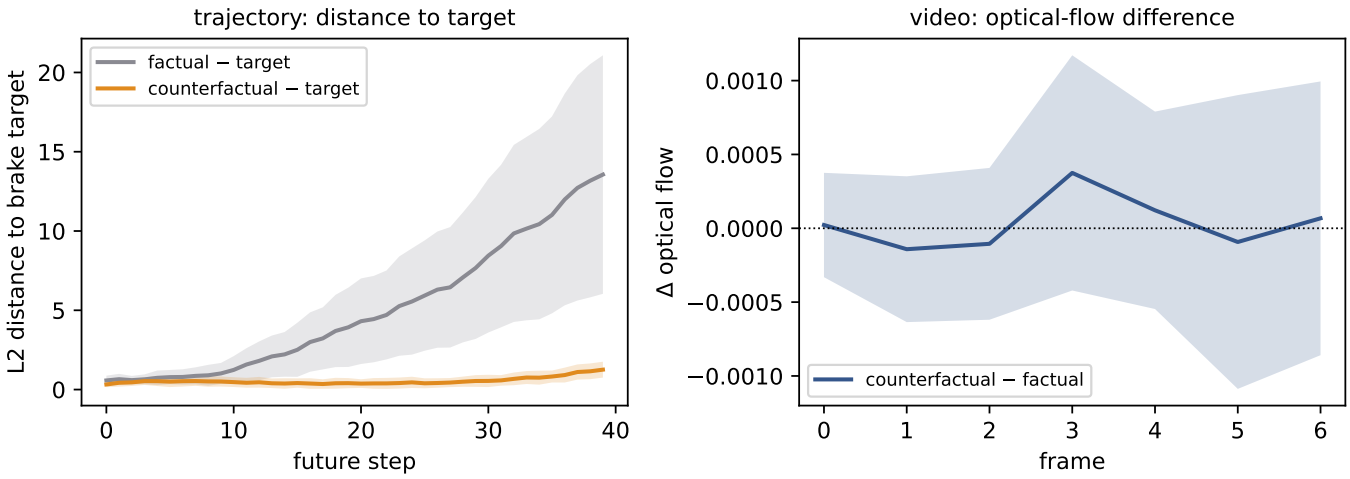

Fig. 3. **The trajectory steers, the video does not.** Across 35 sampled clips at λ=40 (mean $\pm$ std). For each clip, we sample the platform twice from identical noise - unguided (factual) and with a counterfactual brake energy (counterfactual) - and compare. *Left:* L_2 distance of each ego plan to the brake target per step; the counterfactual is driven to the target while the factual diverges. *Right:* per-frame optical-flow difference (counterfactual $-$ factual), which sits at zero within the across-clip spread (shaded). The steered plan does not move the pixels. Detailed results for single clips and per-frame trajectory-video comparison are evaluated and shown in Appendix C.

express the full body of driving knowledge, including collision avoidance, lane keeping, speed limits, and comfort, as a library of differentiable energies injected jointly at sampling time. The difficulty is that such norms routinely conflict: a comfort term that smooths a trajectory may oppose a safety term that demands hard braking. A plain weighted sum of their gradients lets a lower-priority objective override a higher one whenever its magnitude happens to dominate. We therefore plan to impose a priority hierarchy (safety $>$ traffic rules $>$ comfort) at the level of the guidance gradient rather than through hand-tuned weights, which means that lower-priority gradients are projected onto the subspace that does not oppose the active higher-priority constraints at each step. Per-level magnitudes are normalized so that no choice of weights can let, e.g., comfort defeat a safety rule. This would reconcile heterogeneous, possibly conflicting driving rules into a single composed energy field with guaranteed precedence, while preserving the training-free property that a deployed model can add or re-prioritize a rule purely at sampling time.

Conclusion. We show that driving-norm compliance need not be trained in: a frozen rectified-flow world model can have its planned ego trajectory steered at sampling time by differentiable energies. We demonstrate this by steering the trajectory to a counterfactual braking target, which the model is not trained to condition on. The open question fixates on the video stream: under the current architecture, the video does not yet follow the steered trajectory, which we attribute to the cross-stream coupling, and we propose that deep joint denoising of the two streams is the next step toward end-to-end controllable rollouts.

References

[1] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado, "GAIA-1: A generative world model for autonomous driving," *arXiv preprint arXiv:2309.17080*, 2023.

[2] S. Gao, J. Yang, L. Chen, K. Chitta, Y. Qiu, A. Geiger, J. Zhang, and H. Li, "Vista: A generalizable driving world model with high fidelity and versatile controllability," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[3] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. Zhang, "Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[4] J. Yu, Y. Wang, C. Zhao, B. Ghanem, and J. Zhang, "FreeDoM: Training-free energy-guided conditional diffusion model," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[5] A. Bansal, H.-M. Chu, A. Schwarzschild, S. Sengupta, M. Goldblum, J. Geiping, and T. Goldstein, "Universal guidance for diffusion models," *arXiv preprint arXiv:2302.07121*, 2023.

[6] H. Chung, J. Kim, M. T. McCann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems," in *International Conference on Learning Representations (ICLR)*, 2023.

[7] R. Feng, C. Yu, W. Deng, P. Hu, and T. Wu, "On the guidance of flow matching," *arXiv preprint arXiv:2502.02150*, 2025.

[8] X. Peng *et al.*, "Open-sora 2.0: Training a commercial-level video generation model in $200k," *arXiv preprint arXiv:2503.09642*, 2025.

[9] H. Ye, H. Lin, J. Han, M. Xu, S. Liu, Y. Liang, J. Ma, J. Zou, and S. Ermon, "TFG: Unified training-free guidance for diffusion models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[10] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum, "Compositional visual generation with composable diffusion models," in *European Conference on Computer Vision (ECCV)*, 2022.

[11] Y. Du, C. Durkan, R. Strudel, J. B. Tenenbaum, S. Dieleman, R. Fergus, J. Sohl-Dickstein, A. Doucet, and W. Grathwohl, "Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and MCMC," in *International Conference on Machine Learning (ICML)*, 2023.

[12] Y. He, N. Murata, C.-H. Lai, Y. Takida, T. Uesaka, D. Kim, W.-H. Liao, Y. Mitsufuji, J. Z. Kolter, R. Salakhutdinov, and S. Ermon, "Manifold preserving guided diffusion," in *International Conference on Learning Representations (ICLR)*, 2024.

[13] X. Liu, C. Gong, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," in *International Conference on Learning Representations (ICLR)*, 2023.

[14] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *International Conference on Learning Representations (ICLR)*, 2023.

[15] M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden, "Stochastic interpolants: A unifying framework for flows and diffusions," *arXiv preprint arXiv:2303.08797*, 2023.

[16] L. Liu, C. Jia, G. Yu, Z. Song, J. Li, F. Jia, P. Wu, X. Hao, and Y. Luo, "GuideFlow: Constraint-guided flow matching for planning in end-to-end autonomous driving," *arXiv preprint arXiv:2511.18729*, 2025.

[17] S. Jang, T. Ki, J. Jo, J. Yoon, S. Y. Kim, Z. Lin, and S. J. Hwang, "Frame guidance: Training-free guidance for frame-level control in video diffusion models," *arXiv preprint arXiv:2506.07177*, 2025.

[18] Z. Wang, Z. Yuan, X. Wang, Y. Li, T. Chen, M. Xia, P. Luo, and Y. Shan, "MotionCtrl: A unified and flexible motion controller for video generation," in *ACM SIGGRAPH 2024 Conference Papers*, 2024.

[19] S. Yin, C. Wu, J. Liang, J. Shi, H. Li, G. Ming, and N. Duan, "DragNUWA: Fine-grained control in video generation by integrating text, image, and trajectory," *arXiv preprint arXiv:2308.08089*, 2023.

[20] J. Wang, Y. Zhang, J. Zou, Y. Zeng, G. Wei, L. Yuan, and H. Li, "Boximator: Generating rich and controllable motions for video synthesis," *arXiv preprint arXiv:2402.01566*, 2024.

[21] L. Russell, A. Hu, L. Bertoni, G. Fedoseev, J. Shotton, E. Arani, and G. Corrado, "GAIA-2: A controllable multi-view generative world model for autonomous driving," *arXiv preprint arXiv:2503.20523*, 2025.

[22] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "DriveDreamer: Towards real-world-driven world models for autonomous driving," in *European Conference on Computer Vision (ECCV)*, 2024.

[23] K. Zhang, Z. Tang, X. Hu, X. Pan, X. Guo, Y. Liu, J. Huang, L. Yuan, Q. Zhang, X.-X. Long, X. Cao, and W. Yin, "Epona: Autoregressive diffusion world model for autonomous driving," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[24] T. Xia, Y. Li, L. Zhou, J. Yao, K. Xiong, H. Sun, B. Wang, K. Ma, G. Chen, H. Ye, W. Liu, and X. Wang, "DriveLaW: Unifying planning and video generation in a latent driving world," *arXiv preprint arXiv:2512.23421*, 2025.

[25] Z. Zhong, D. Rempe, D. Xu, Y. Chen, S. Veer, T. Che, B. Ray, and M. Pavone, "Guided conditional diffusion for controllable traffic simulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[26] Z. Zhong, D. Rempe, Y. Chen, B. Ivanovic, Y. Cao, D. Xu, M. Pavone, and B. Ray, "Language-guided traffic simulation via scene-level diffusion," in *Conference on Robot Learning (CoRL)*, 2023.

[27] Y. Zheng, R. Liang, K. Zheng, J. Zheng, L. Mao, J. Li, W. Gu, R. Ai, S. E. Li, X. Zhan, and J. Liu, "Diffusion-based planning for autonomous driving with flexible

guidance," in *International Conference on Learning Representations (ICLR)*, 2025.

[28] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations (ICLR)*, 2021.

[29] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv preprint arXiv:1812.01717*, 2018.

[30] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[31] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[32] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

APPENDIX

This appendix derives the energy guidance for rectified flow used in Section III (§ Appendix A), documents the model architecture and fine-tuning details (§ Appendix B), and illustrates the energy guidance on video and ego-trajectory generation (§ Appendix C).

A. Energy Guidance for Rectified Flow

This section first establishes the equivalency between the velocity field in rectified flow and the marginal score from the energy score in Appendix A1, and discusses the implementation and the overhead of the energy guidance in Appendix A2 and Appendix A3.

Our backbone is a rectified-flow model with the interpolant and conditional velocity target

$$\begin{aligned} \mathbf{x}_t &= \alpha_t\,\mathbf{x}_0 + \sigma_t\,\mathbf{x}_1, \quad \mathbf{x}_1 \sim \mathcal{N}(0, I), \\ \alpha_t &= 1 - t, \quad \sigma_t = \sigma_{\min} + (1 - \sigma_{\min})\,t, \end{aligned} \tag{5}$$

with $\mathbf{v}_\theta(\mathbf{x}_t, t) \approx \mathbb{E}[\dot{\mathbf{x}}_t \mid \mathbf{x}_t] = (1 - \sigma_{\min})\mathbf{x}_1 - \mathbf{x}_0$. The floor $\sigma_{\min}$ is the minimum noise at the data end [14]. Open-Sora 2.0 uses $\sigma_{\min}$=10^{-5} with the Movie Gen parameterization [8]. We take the same formulation for the interpolant, the velocity target, this $\sigma_{\min}$, and the timestep schedule at fine-tuning and sampling, so the schedule and $\sigma_{\min}$ match those the backbone was trained with.

1) Score and velocity equivalency: For the Gaussian path, the velocity field and the marginal score $s_t(\mathbf{x}_t) = \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ are affinely equivalent [7, 15]:

$$\begin{aligned} \mathbf{v}_\theta(\mathbf{x}_t, t) &= -\frac{\mathbf{x}_t + \sigma_t\, s_t(\mathbf{x}_t)}{\alpha_t}, \\ s_t(\mathbf{x}_t) &= -\frac{\alpha_t\,\mathbf{v}_\theta(\mathbf{x}_t, t) + \mathbf{x}_t}{\sigma_t}. \end{aligned} \tag{6}$$

The transformation is affine and invertible at every $t \in (0, 1)$, so any score-based guidance transfers to the velocity field exactly, without score network or retraining. The one-step clean-sample estimate is the rectified-flow analogue of Tweedie's posterior mean, $\hat{\mathbf{x}}_0 = \mathbf{x}_t - t\,\mathbf{v}_\theta$ as $\sigma_{\min} \to 0$ (same as Eq. (2)).

2) Guided velocity.: Score-based energy guidance encodes external knowledge as a likelihood $p_t(y \mid \mathbf{x}_t)$. With the standard training-free approximation ($\nabla_{\mathbf{x}_t} \log p_t(y \mid \mathbf{x}_t) \approx -\nabla_{\mathbf{x}_t} \mathcal{E}(\hat{\mathbf{x}}_0)$) and the affine transformation Eq. (6), together with the constant Jacobian $\partial\hat{\mathbf{x}}_0/\partial\mathbf{x}_t = (1 - \sigma_{\min})I$, the guided velocity is a single additive correction in clean-sample space,

$$\begin{aligned} \tilde{\mathbf{v}}(\mathbf{x}_t, t) &= \mathbf{v}_\theta(\mathbf{x}_t, t) + w(t)\,\nabla_{\hat{\mathbf{x}}_0} \mathcal{E}(\hat{\mathbf{x}}_0), \\ w(t) &= \frac{(1 - \sigma_{\min})\,\sigma_t}{\alpha_t}, \end{aligned} \tag{7}$$

followed by the ordinary Euler step. We replace the weight $w(t)$ with the schedule $\lambda(t)$ in Eq. (3), which is capped near t=1 and zeroed near t=0, as weight $w(t)$ grows without bound there. We allow $\lambda(t)$ to be a tunable hyperparameter during sampling in the implementation.

3) Energy guidance overhead during sampling: The correction needs only $\nabla_{\hat{\mathbf{x}}_0} \mathcal{E}$, which we obtain with ordinary automatic differentiation (PyTorch autograd) on the small closed-form energy of the (x, y, θ, v) ego trajectory. Because the energy is evaluated on $\hat{\mathbf{x}}_0$ with the predicted velocity detached, the 11B-parameter backbone never enters the `autograd` graph. Moreover, because the guided quantity is the trajectory rather than the video latents, there is no VAE backpropagation and no pixel-space manifold drift. The cost is one gradient of a small analytic function per step, therefore negligible beside the forward pass.

B. Model Architecture and Fine-tuning Details

This section documents how the video-diffusion driving world model in Section III is obtained. All numbers are for the energy-guided platform evaluated in the main text; the flow-matching convention and timestep schedule are those of Appendix A. The backbone's sampling-time computation graph is shown in Fig. 2.

1) Backbone and Architectural Modifications: The backbone is the Open-Sora 2.0 MM-DiT (DC-AE variant) [8]: 19 double-stream blocks and 38 single-stream blocks, hidden size 3072, 24 heads, with separate $q/k/v$ projections. The image stream predicts the future videos. We repurpose the text stream into ego trajectory prediction. Each stream gets its own timestep embedding. The video reuses the pretrained timestep MLP, while the ego timestep MLP is newly added and zero-initialized, so at initialization (tied timesteps) the model reproduces the pretrained behaviour exactly. A past-observation pooler maps the clean past latent to the conditioning channel. The single-stream blocks are *bypassed* due to architectural modifications for the ego trajectory prediction, because we need to output both video and trajectory prediction at the same time, and the 19 double-stream blocks provide such interfaces.

2) Fine-tuning details: The entire backbone is frozen. We attach LoRA adapters [30] (rank 16, $\alpha = 32$) to the modulation, attention, and MLP projections of both streams (`img_{mod,attn,mlp}`, `txt_{mod,attn,mlp}`, 266 linear layers, 46.7M parameters trained at learning rate $1{\times}10^{-4}$). The newly introduced modules (ego encoder, ego output head, ego timestep MLP, past-obs routing, and pooler) are initialized with zeros and trained from scratch at $3{\times}10^{-4}$. In total 88.1M parameters are trainable out of the $\sim$11B backbone.

3) Objective and optimization: We minimize a joint rectified-flow loss $\mathcal{L} = \mathcal{L}_{\text{video}} + \lambda_{ego}\mathcal{L}_{\text{ego}}$ ($\lambda_{ego} = 1$), each an MSE on the predicted velocity (Appendix A). The ego term is masked to the valid future frames. Optimization uses AdamW ($\beta = (0.9, 0.98)$, weight decay 0, $\epsilon = 10^{-8}$) with gradient clipping at 1.0 and `bf16` autocast. To fit the backbone on a single NVIDIA HGX H100 94GB GPU we keep the frozen weights in `bf16` and apply gradient activation checkpointing to every transformer block. Together, these allow the batch-8 training to fit on the single card.

4) Classifier-free guidance: The platform is also trained for classifier-free guidance [31] over its past conditioning, which sharpens the generated video. With probability 0.15 per

sample we drop the observed past by replacing the pooled past-observation vector with a learned null embedding and zeroing the in-context past latent frames together, so the model additionally learns the unconditional (marginal) future distribution. At sampling we extrapolate $\mathbf{v} = \mathbf{v}_{\varnothing} + w\,(\mathbf{v}_{\mathrm{c}} - \mathbf{v}_{\varnothing})$ on both streams with guidance scale w (we use $w \in [3, 5]$). This is orthogonal to the energy guidance of Section III, which acts only on the ego stream and needs no retraining.

5) Data and autoencoder: We train on nuScenes `v1.0-trainval` [32], front camera, at 256×448, using 12-frame clips at the native sample-data cadence (stride 12, non-overlapping). The frozen DC-AE (`dc-ae-f32t4c128`, scaling factor 0.493, $4\times$ temporal / $32\times$ spatial compression) encodes each clip to a $(3, 8, 14)$ latent, where the first latent frame is the clean past conditioning and the remaining two are the denoised future. The ego target is the 40-step future trajectory.

6) Schedule and checkpoint: Timesteps are sampled logit-normal (in compliance with the Open-Sora 2.0 backbone); the batch size is 8 per GPU. The checkpoint used here is trained for $\sim$22k steps on a single GPU, with validation on a held-out tail of 64 clips and best-on-validation checkpointing. The platform is trained once. Every result in the main text reuses these frozen weights and adds knowledge only through sampling-time energies.

7) Code availability: We release the code required to reproduce the results from the released checkpoint: the training-free energy-guidance sampler, the differentiable energy library and the score-velocity bridge, the counterfactual-brake probe, and the analysis and figure scripts at https://timsu-98.github.io/energy_guidance_for_driving_world_models/. We also release the trained model weights with the LoRA adapters and the additionally introduced modules (88M trainable parameters). Reconstituting the full backbone also requires two checkpoints that we use and do not redistribute: the $\sim$ 11B Open-Sora 2.0 base MM-DiT [8] and its DC-AE autoencoder (`dc-ae-f32t4c128`), both obtained from their original public releases.

C. Counterfactual Probe: Per-Clip Views

Figure 3 aggregates the cross-stream probe. Here, we show its per-clip structure. Figure 4 and Fig. 5 overlay the bird's-eye-view trajectories: the counterfactual plan (orange) tracks the braked target while the unguided plan (grey) runs long, on essentially every clip. Figure 6 aligns one clip's ego speed and the per-frame optical-flow difference on a common frame axis: the plan brakes hard while the optical-flow difference stays around zero, which indicates the steered planned trajectory does not reach the video stream.

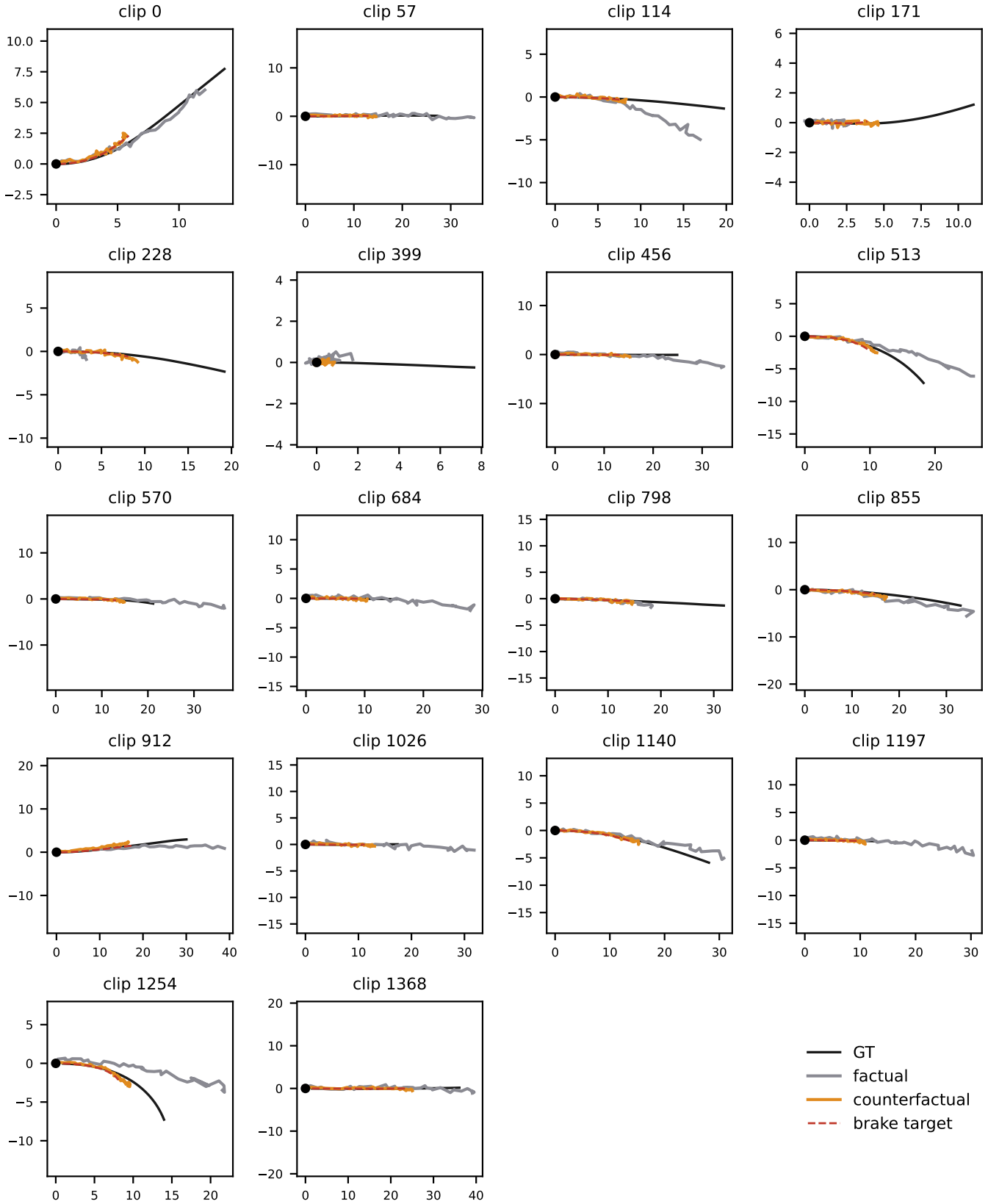

Fig. 4. **Per-clip BEV trajectories (1/2).** All 35 usable sampled clips at λ=40. GT (black), unguided/factual (grey), counterfactual (orange), brake target (red dashed). The counterfactual plan tracks the braked target across clips.

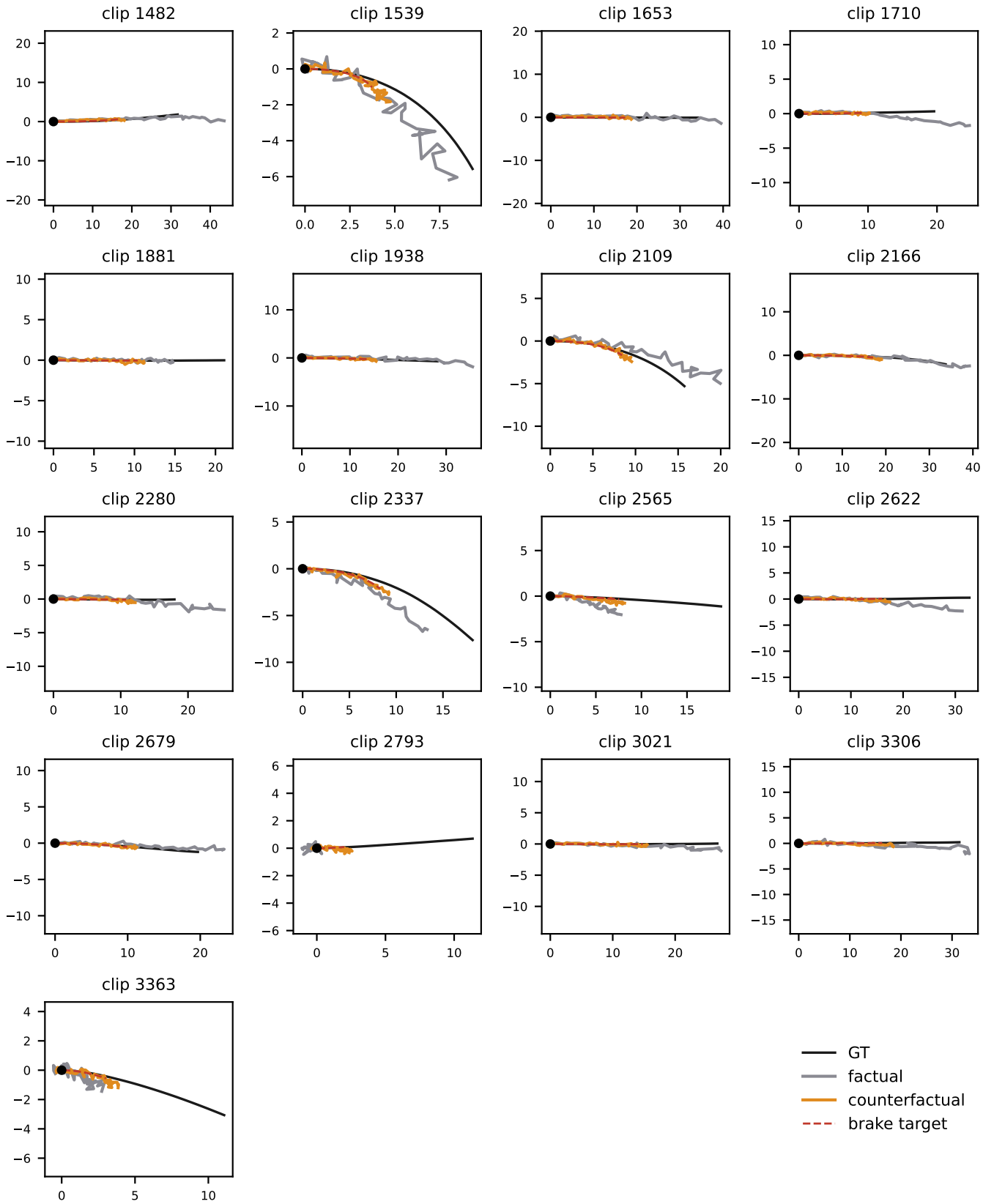

Fig. 5. **Per-clip BEV trajectories (2/2).** Continuation of Fig. 4.

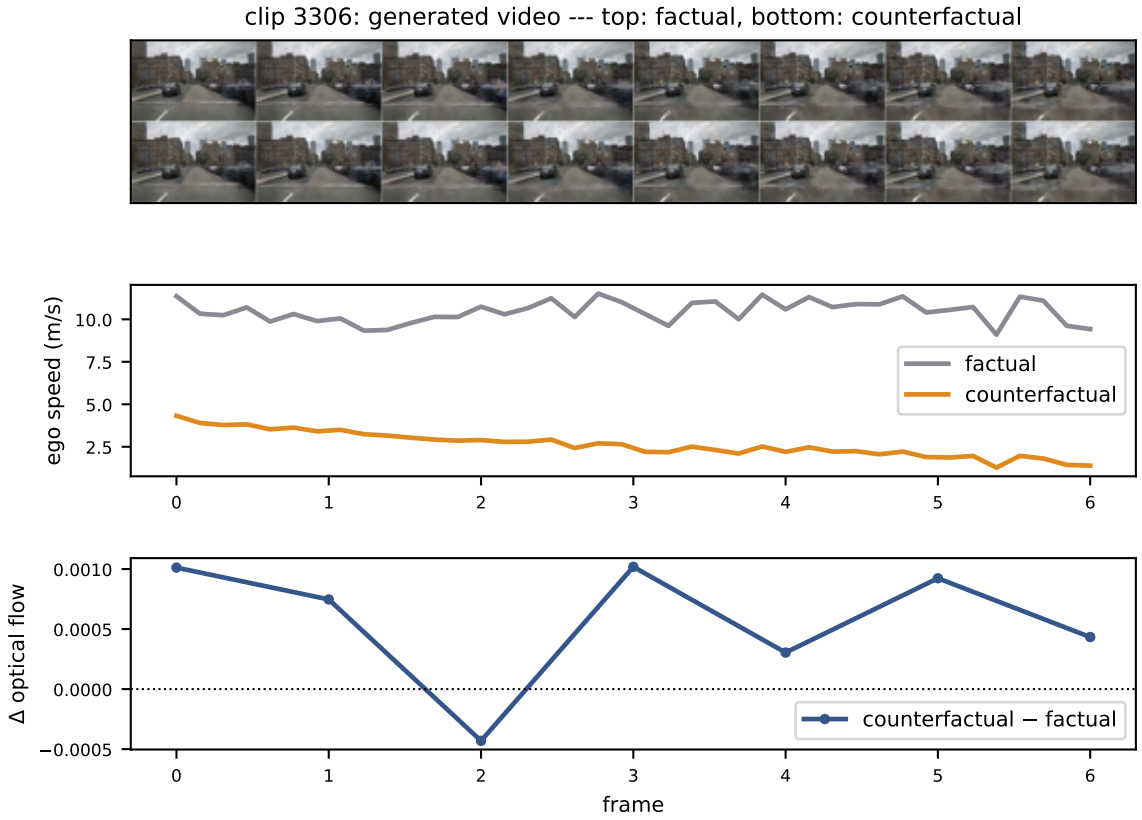

Fig. 6. **One clip, aligned on frames.** *Top:* ego speed - the counterfactual brakes hard while the factual holds speed. *Bottom:* per-frame optical-flow difference (counterfactual − factual), at zero: the trajectory steers but the video does not. (Plan steps are mapped onto the video-frame axis for alignment.)