MAD: Motion Appearance Decoupling for efficient Driving World Models

Supplementary Material

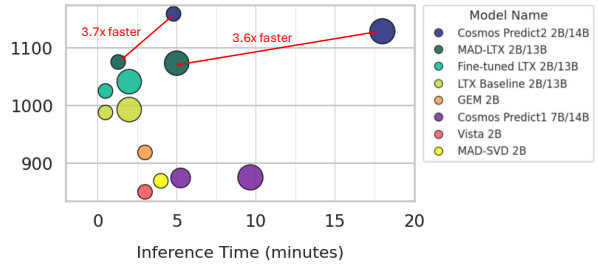

Figure 7. Inference Speed vs. Perceptual Quality. We report the inference time measured on a single GH200 GPU. Performance is quantified using the ELO score computed from our human preference study. **Our MAD-LTX models achieve up to 3.6× faster inference than competitive models of comparable size and performance**

5. Inference Speed

We report the inference latency for 5 seconds video generation in figure 7 on a single NVIDIA GH200 GPU. We exclude functionalities that require additional model loading (making the inference slower) like upsampling or model offloading for Cosmos baselines. All tests use the default resolution and generated frame count specified by each respective model, for models that generate short segments, multiple rollouts are used to reach the 5-second duration used in our evaluation.

Direct measurements for Cosmos-Predict2 on the GH200 were not available. We report its A100 inference time as 41 minutes (14B) and 11 minutes (2B). The A100 inference speed is rescaled to the GH200 for Predict2 using a conversion ratio computed from other baselines.

Our MAD-LTX models achieve up to 3.6× faster inference than competitive models of comparable size and performance

6. Additional Experiments

6.1. Evaluation of Control Capabilities

We evaluate the fidelity of our model's three primary control axes: **ego trajectory control, object control, and textual control**. Experiments are conducted on a subset of N=800 video clips (5 seconds each) from the Waymo Open Dataset. As shown in Tab. 3, we adopt the evaluation protocol from [17] and use our unconditional models ('MAD-LTX-uncond') as a reference baseline to demonstrate the effectiveness of our conditioning signals. The specific evaluation metrics for each modality are defined as follows:

Table 3. Evaluation of control fidelity. We evaluate each control type over 800 generated samples from Waymo Open Dataset [37]. 'MAD-LTX-uncond' (our models without any control inputs) serves as reference baselines to measure the effectiveness of the conditioning. **MAD-LTX effectively supports all three distinct control modalities**

Model	Ego ↓	Obj ↑	Text ↑	
			Action	**Object**
MAD-LTX-2B-uncond	5.2	0.40	38.8%	42.1%
MAD-LTX-2B (ours)	**1.4**	**0.51**	**40.8%**	**45.6%**
MAD-LTX-13B-uncond	3.4	0.45	39.2%	44.6%
MAD-LTX-13B (ours)	**1.5**	**0.55**	**43.9%**	**49.1%**

Ego-Motion Control. We condition the generation on Ground Truth (GT) ego-camera poses from the Waymo Open Dataset. To measure fidelity, we extract the realized ego-trajectory from both the GT video and the generated video using MapAnything [21]. We report the **Average Displacement Error (ADE)** between these two extracted trajectories as the primary metric.

Object Control. We utilize OpenPifPaf [23] to detect and track a single target object in the GT video. We condition the generation on the single object extracted location from the GT video. For evaluation, we run OpenPifPaf [23] on the generated video to locate the corresponding object. We report the **Intersection over Union (IoU)** between the conditioned GT box and the detected box in the generated frame. If the target object fails to appear in the generated video, an IoU of 0 is assigned.

Text Control. We employ an automated Visual Question Answering (VQA) pipeline to assess text adherence across two distinct categories: **Action** (dynamics/maneuvers) and **Object** (entity presence). We use **Qwen2.5** [32] to extract semantic elements from the input prompt and reformulate them into a binary verification question tailored to the specific category—for example, *Is the ego car turning left?* for textual action control, and *Is a cyclist entering the video?* for textual object control. We then feed the generated video and this question into the Vision-Language Model **Qwen2.5-VL** [2] to obtain a Yes or No prediction. We report the **Success Rate** (percentage of "Yes" answers) for each category.

Control Results As detailed in Table 3, we observe consistent quantitative improvements across all metrics and model

sizes when comparing the conditioned MAD-LTX models against their unconditional counterparts. The performance gap is most significant in **Ego-Motion Control amd Object control**. This is expected: while the unconditional model typically forecasts a *plausible* future object/ego trajectory (representing one valid mode of the motion distribution), it is penalized for deviating from the specific Ground Truth realization. The conditioned model, however, effectively utilizes the control signal to lock onto the correct mode. For **Textual Control**, the improvements are consistently positive but less significant. We attribute this to the unconditional model being trained on static captions (describing the first frame), which yields a high starting score, while VLM noise limits the measured gain. These results confirm that—consistent across both parameter scales—**MAD-LTX effectively supports all three distinct control modalities.**

6.2. Standard video quality metrics (FID/FVD)

We report in table 4 the Fréchet Inception Distance [18] (FID) and Fréchet Video Distance [38] (FVD) for the Base LTX [15] model, the standard fine-tuned baseline (Fine-tuned LTX), and our proposed MAD-LTX across both 2B and 13B parameter scales.

All metrics were computed on the OpenDV [42] test set using N=5,000 generated samples per method. As shown in Table 1, while we report these metrics for completeness, we observe—consistent with recent literature[12] —that **FID and FVD scores do not strictly correlate with human perceptual preference in this domain.**

Table 4. **Standard video quality metric on the OpenDV test set.** We report FID and FVD metrics computed using $N = 5,000$ samples. Consistent with prior work[12], **we observe that these distribution based metrics do not necessarily correlate with human perceptual preference.**

Model Size	Method	FID ↓	FVD ↓
2B	Base LTX	4.06	**64.40**
	Fine-tuned LTX	**2.66**	67.17
	MAD-LTX (Ours)	3.72	92.79
13B	Base LTX	2.21	**56.15**
	Fine-tuned LTX	**1.94**	69.56
	MAD-LTX (Ours)	2.39	59.61

6.3. Comparison with external Baselines

Photorealistic world models, such as MAD-LTX, VISTA, and GEM, are primarily designed as *controllable simulators* rather than unconditional motion planners. Therefore, the key benchmark for these systems is their ability to strictly respect input control conditions while maintaining high visual fidelity.

We compare MAD-LTX against two controllable baselines, VISTA and GEM. To ensure a fair evaluation aligned with their established protocols, we measure ego-motion control fidelity via the Average Displacement Error (ADE) on a shorter 2.5-second horizon using $N = 800$ validation samples from the Waymo dataset.

As shown in Tab. 5, MAD-LTX at both the 2B and 13B parameter scales significantly outperforms existing baselines in control fidelity, achieving an ADE of 0.89 for the 13B model compared to GEM's 1.47. Furthermore, while distribution-based metrics like FID and FVD can sometimes be noisy for fine-grained comparisons (as discussed in Sec. 6.2), the massive margin of improvement observed here (e.g., an FVD of 59.61 vs. GEM's 186.19) reflects a clear leap in generation quality. **MAD-LTX models, delivers both superior visual quality and adherence to ego-motion conditioning.**

Model	FID ↓	FVD ↓	ADE ↓
VISTA	10.05	215.41	1.77
GEM	9.77	186.19	1.47
MAD-LTX 2B	3.72	92.79	1.06
MAD-LTX 13B	**2.39**	**59.61**	**0.89**

FID/FVD: OpenDV val (5s, 5k). ADE: Waymo val (2.5s, 800).

Table 5. Comparison with external baselines. **MAD-LTX significantly outperforms previous state-of-the-art models in both visual quality and control precision.**

7. Human Evaluation

7.1. Protocol

We conducted a blinded pairwise comparison study hosted on a custom Hugging Face interface. The specific layout of the pairwise comparison interface is illustrated in fugure 8. To ensure high-quality feedback, the participant pool consisted exclusively of domain experts—computer vision researchers specializing in autonomous driving.

For each baseline, we generated $N = 100$ video samples (5 seconds duration). In every comparison, annotators were presented with two anonymized videos and asked to rate their preference (Strongly Prefer, Prefer, or No Preference) based on the following specific prompts:

- **General Quality:** *"Overall which video do you prefer? In other words which video is harder to distinguish from real video?"* (Results reported in Figure 5 of the main manuscript).
- **Motion Quality and Realistic Dynamics:** *"Which video has more realistic, fluid and coherent motion, and is physically and socially plausible? Stopping suddenly without reason, collisions between objects, or cars driving in wrong direction are example of poor motion quality."*

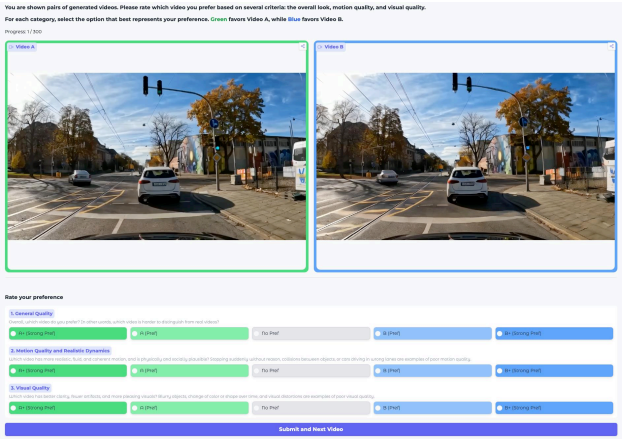

Figure 8. **User Study Interface:** We employ a side-by-side pairwise comparison protocol to evaluate generation performance. Participants are asked to rate their preference between two anonymized videos across three specific criteria: General Quality, Motion Quality and Realistic Dynamics, and Visual Quality. For each criterion, annotators select from a 5-point scale ranging from "Strong Preference" or "Preference" for a specific model, to "No Preference."

- **Visual Quality:** *"Which video has better clarity, fewer artifacts, and more pleasing visual? Blurry object, change of color or shape over time, and visual distortions are example of poor visual quality."*

7.2. Additional results

In this section, we report the detailed results for **Motion Quality** and **Visual Quality**.

The detailed pairwise win rates for Motion Quality and Visual Quality are presented in figures 9, 10, 11 and 12. We observe that user preferences on these two axes strongly correlate with the General Quality results reported in the main manuscript. Consistent across both the 2B and 13B parameter scales, **MAD-LTX outperforms all evaluated open-source baselines in terms of both motion realism and visual fidelity.** In alignment with the global preference trends, the proprietary Cosmos-Predict2 is the only model that retains a higher preference rate than our method on both criteria.

8. Additional Hyperparameters

We share a comprehensive list of all the hyper parameters we used for training our Motion Forecaster and our Motion Synthesizer in Tab. 6. All codes with the corresponding config files will be released for reproducibility.

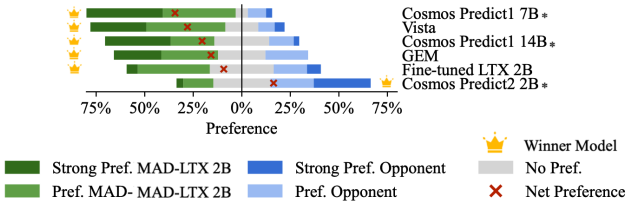

Figure 9. **Motion Realism Preference (MAD-LTX 2B).** Human preference evaluation specific on motion plausibility (e.g., fluidity, absence of collisions/sudden stops). '*' denotes proprietary mod- els which use private datasets and significant computational resource **MAD-LTX-2B outperforms all open-source models, ranking second only to the proprietary Cosmos-Predict2.**

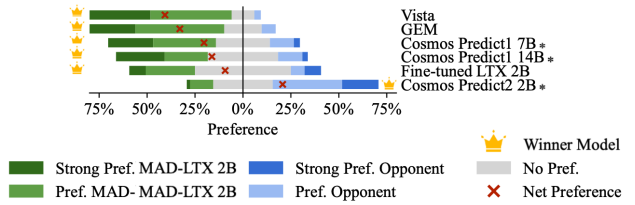

Figure 10. **Visual Quality Preference (MAD-LTX 2B).** Human preference evaluation specific on visual quality (e.g., temporal consistency, distortions). '*' denotes proprietary models which use private datasets and significant computational resources. **MAD-LTX-2B outperforms all open-source models, ranking second only to the proprietary Cosmos-Predict2.**

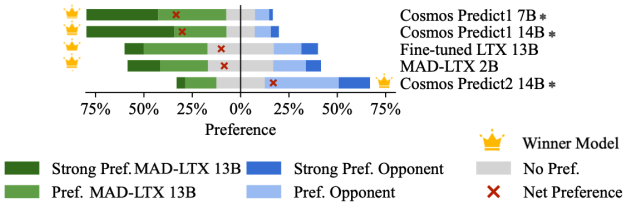

Figure 11. **Motion Realism Preference (MAD-LTX 13B).** Human preference evaluation specific on motion plausibility (e.g., fluidity, absence of collisions/sudden stops). '*' denotes proprietary models which use private datasets and significant computational resources. **MAD-LTX-13B outperforms all open-source models, ranking second only to the proprietary Cosmos-Predict2.**

9. Limitations and Future Work

While our intermediate motion representation has proved to be useful, it currently omits contextual driving signalization for precise prediction, such as traffic light states or traffic signs. Furthermore, we observed that while the overall pseudo-labeling strategy was successful, the generated lane keypoints are inherently noisy, hindering accurate road layout modeling. An ablation showed that replacing these with road segmentation masks could offer a more ro-

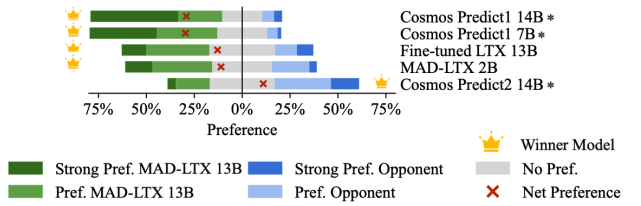

Figure 12. **Visual Quality Preference (MAD-LTX 13B).** Human preference evaluation specific on visual quality (e.g., temporal consistency, distortions). '*' denotes proprietary models which use private datasets and significant computational resources. **MAD-LTX-13B outperforms all open-source models, ranking second only to the proprietary Cosmos-Predict2.**

Table 6. LTX Training Hyperparameters

Category	Parameter	Value	Description
Model Strategy	Model Source	`LTX_2B_0.9.6_DEV` and `LTX_13B_097_DEV`	Base 2B parameter model and 13B parameter model.
	Training Mode	`lora`	Low-Rank Adaptation fine-tuning.
	Precision	`bf16`	Brain Floating Point 16 mixed precision.
LoRA Config	Rank (r)	512	High rank for high expressivity.
	Alpha (α)	512	Scaling factor (1:1 ratio with rank).
	Targets	`q, k, v, out, ff`	Applies to Attention and Feed-Forward blocks.
Optimization	Learning Rate	2×10^{-4}	Step size for the optimizer.
	Optimizer	`adamw`	Standard AdamW optimizer.
	Batch Size	32	Small batch size due to VRAM constraints. But used 32 GPUs, so an effective batch size of 32.
	Scheduler	`linear`	Linear learning rate decay.

bust structural representation. For future work, we will focus on incorporating these driving elements and exploring segmentation-based road representations.

Beyond the driving domain, the core MAD principle could be extended to other areas for which pose is a natural representation, such as human-centric video generation, an area where industry currently focuses on extensive training over large datasets, and to ego-view object manipulation for robotics, which directly relates to the world model problematics.