SEEING RISK BEFORE IT HAPPENS: RISK-AUGMENTED WORLD MODEL FOR ANTICIPATORY DRIVING RISK UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing driving risk perception methods identify the safety-critical object only after the risk has become salient, leaving little time to react. We study anticipatory driving risk understanding, which localizes and describes a potential risk object as it will appear at a future risk moment, before the hazard becomes imminent. We address the task with a generate-then-perceive paradigm, in which a driving world model generates the future scene and risk understanding is performed within it. However, existing world models have no notion of safety criticality, making the generated risk object prone to distortion and blurring. We propose the Risk-Augmented World Model (RAWM), which augments the generation of the risk object with its appearance and semantic cues through a dual-branch conditioning module built on decoupled cross-attention. Moreover, a learnable risk prompt module estimates both cues from the early observation at inference without requiring risk annotations. On the DRAMA benchmark, RAWM consistently improves all localization and captioning metrics over the baseline world model and approaches the performance of an oracle model that observes the ground-truth future risk scene. Ablations on generation fidelity and downstream risk understanding attribute these gains to a more faithful generation of the future risk object. Compared to alternatives that either directly predict risks from the early observation or extrapolate risk-object trajectories, RAWM achieves increasingly larger gains as the anticipation horizon grows.

1 INTRODUCTION

Safe driving hinges on recognizing a developing risk early enough to act. A skilled driver mentally rolls the scene forward to single out an emerging risk and respond in time. However, existing risk perception methods identify the safety-critical object only after the risk has become salient or even an accident is imminent (Malla et al., 2023; Artham et al., 2026), leaving little margin to maneuver. We instead study *anticipatory driving risk understanding*, which proactively localizes and describes the risk object as it will appear at a future risk moment, before the risk turns critical (Fig. 1).

A straightforward approach is to predict the future risk directly from the early observation. Yet such direct prediction requires the model to infer scene evolution and identify the safety-critical object simultaneously, and sparse risk annotations can hardly support learning driving scene dynamics. Alternatively, the trajectory of the risk object could be extrapolated from its observed motion into the future. However, risks often emerge from complex interactions or unexpected events, where such extrapolation becomes unreliable.

Driving world models generate future scenarios from past observations (Hu et al., 2023; Gao et al., 2024b), providing a natural foundation for anticipatory risk understanding. In this work, we leverage them in a *generate-then-perceive* paradigm: the world model first generates a plausible future scenario, after which a task-specific head performs risk understanding within it. In this way, the world model can learn scene dynamics from large-scale driving videos, while risk understanding reduces to a perception task over concrete future states. The generated future scenario also provides direct visual evidence, making the basis of risk understanding transparent and interpretable, and is itself a richer anticipatory signal than abstract prediction outputs.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

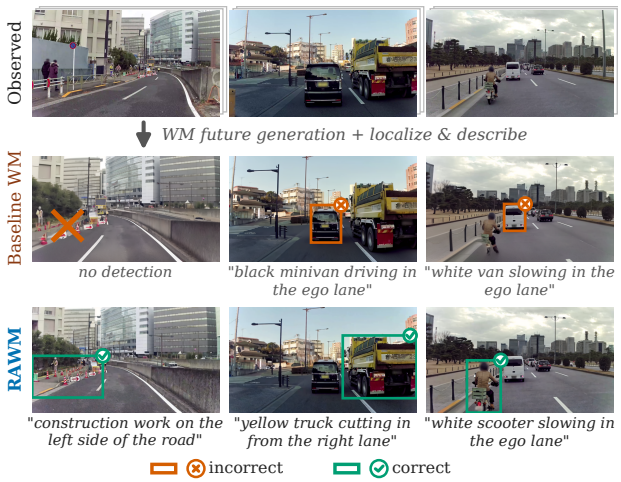


Figure 1: Anticipatory driving risk understanding via world-model generation. From a few observed frames (top), the world model generates the future scene, and a risk head localizes and describes the risk object within it before the risk moment actually arrives. The baseline world model often generates the risk object at low fidelity, so the risk head may target a wrong object or detect nothing, while RAWM renders it faithfully enough for correct risk understanding.

The key challenge of the paradigm lies in preserving the fidelity of the generated risk object. A world model is trained to reconstruct the whole scene, but it lacks explicit awareness of which object is safety-critical. Thus, risk objects, which are often small, sparse, and associated with highly uncertain future behaviors, are prone to distortion and blurring in the generated future scenarios (Fig. 1). When that happens, localization and captioning in the generated frame inevitably deteriorate and may even fail when the target disappears. Hence, the central problem is to make the generated future faithful to the object that will become safety-critical.

We address this problem with the Risk-Augmented World Model (RAWM), which augments the generation of the risk object in the future scenario. RAWM conditions generation on two complementary cues of the object, its *appearance* and its *semantics*, through a dual-branch conditioning module, so that the generated future renders it more faithfully. At inference, a learnable risk prompt module estimates both cues from the observed frames without requiring risk annotations. Our contributions are threefold:

- We formulate anticipatory driving risk understanding as a forecasting task, shifting risk understanding from perceiving salient risks toward anticipating emerging ones and securing reaction time for the ego vehicle.
- We propose a generate-then-perceive paradigm for this task and realize it with RAWM, which augments future generation toward faithfully rendering the risk object, turning the generated scene into a reliable basis for the task.
- On the DRAMA benchmark (Malla et al., 2023), RAWM consistently improves all localization and captioning metrics over a baseline world model and approaches an oracle that observes the ground-truth future risk frame. We further provide the first characterization of how driving risk understanding varies with the anticipation horizon.

2 Related Work

Driving world models. Generative world models learn to imagine how a scene evolves (Ha & Schmidhuber, 2018; Hafner et al., 2020). In driving, one line of work uses them as neural simulators for high-fidelity scene generation. GAIA-1 (Hu et al., 2023), Vista (Gao et al., 2024b), and GAIA-2 (Russell et al., 2025) generate realistic driving futures under action, text, or multi-camera control, and GenAD (Yang et al., 2024a) scales video prediction to web-scale driving data. DriveDreamer (Wang et al., 2024a) and MagicDrive (Gao et al., 2024a) condition generation on structured layouts

such as road maps and 3D boxes, and DriveDreamer-2 (Zhao et al., 2025) adds language-guided customization. A complementary line synthesizes safety-critical scenarios as training or evaluation data (Zhou et al., 2025; Stoler et al., 2025; Peng et al., 2025).

Another line uses world models to assist downstream perception and planning. ViDAR (Yang et al., 2024b) and DriveWorld (Min et al., 2024) pre-train perception backbones by forecasting future point clouds and 4D occupancy. OccWorld (Zheng et al., 2024) and Drive-OccWorld (Yang et al., 2025) forecast occupancy dynamics for planning, and Drive-WM (Wang et al., 2024b) plans by evaluating multiview visual forecasts. A further step couples them directly with action: MILE (Hu et al., 2022) and Think2Drive (Li et al., 2024) learn driving policies by imitation or reinforcement inside latent world models, and LAW (Li et al., 2025) self-supervises end-to-end driving by predicting future latent scene features. However, existing driving world models are not designed to preserve task-specific entities or attributes during future generation, which may limit their applicability to tasks requiring fine-grained object-level fidelity.

Driving risk understanding. DRAMA (Malla et al., 2023) formulates joint risk localization and captioning and provides the benchmark we build on. Subsequent work tackles the task with multimodal LLMs on the same benchmark (Ding et al., 2025; Fan et al., 2025; Artham et al., 2026), localizing and describing the risk object while further generating intention predictions or safety suggestions for the ego vehicle. All of these methods perform *reactive* risk understanding, in which the risk is identified only after it has become salient. We instead target anticipatory driving risk understanding and address it by localizing and describing the risk in a generated future scene, so the understanding is available before the risk moment actually arrives. Methods for reactive risk understanding are therefore complementary rather than competing: they see the very risk frame we must generate, and their reported numbers are not directly comparable to ours.

3 Risk-Augmented World Model

3.1 Problem Formulation and Overview

We consider driving clips in which one safety-critical object (henceforth the *risk object*) is annotated at a designated *risk moment*. Let $x_{1:T}$ denote the frames of a clip, with the risk moment at the final timestep T, where the risk object carries a bounding box b_T and a free-form caption c_T. Anticipatory driving risk understanding is a forecasting task that observes only the first k frames $O = x_{1:k}$ and predicts (b_T, c_T). In other words, the task is to localize and describe the risk object as it will appear at the future risk moment, beyond the observed window. The *anticipation horizon* h measures how far ahead the risk moment lies. Relative to reactive risk understanding, the horizon is the lead time of the understanding and the reaction time gained for the ego vehicle.

We address the task with a *generate-then-perceive* paradigm. A generative world model G first synthesizes the future $\hat{x}_{k+1:T} = G(O)$. A risk head then localizes and describes the risk object in the generated risk frame $\hat{x}_T$, producing $(\hat{b}_T, \hat{c}_T)$. We instantiate G with Vista (Gao et al., 2024b), a latent video diffusion model (Rombach et al., 2022; Blattmann et al., 2023) for driving that denoises future frames conditioned on the observed ones. Whether this pipeline succeeds depends on the fidelity of the risk object in the generated $\hat{x}_T$, which a world model trained for generic future prediction does not guarantee. RAWM therefore augments future generation toward the risk object (Fig. 2). A dual-branch conditioning module conveys the object to G through two conditions, its *appearance* and its *semantics* (Sec. 3.2), injected via decoupled cross-attention (Sec. 3.3). During training, the conditions are extracted from the annotated risk frame x_T. At inference, a learnable risk prompt module estimates them from the observed frames alone, so no risk annotation is required (Sec. 3.4). Features of the generated risk frame finally serve as the input to the risk head (Sec. 3.5).

3.2 Dual-Branch Risk Conditioning

The most direct way to convey the risk object is to specify its location, as in layout-conditioned generation, where boxes or maps are given as inputs (Wang et al., 2024a; Gao et al., 2024a). In our setting, however, there is no usable location. The risk object's observed location quickly becomes stale, as both the object and the ego vehicle keep moving, while its location at the future risk moment is exactly what we aim to predict. We therefore drop location entirely: RAWM tells the world model

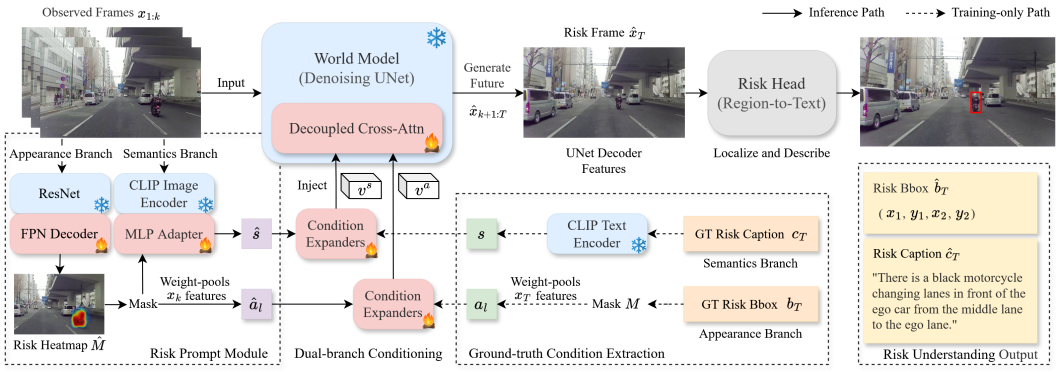


Figure 2: RAWM overview. Dual-branch risk conditions (appearance and semantics) are injected into the frozen world model through decoupled cross-attention. They are extracted from the annotated risk frame during training and estimated from the observed frames by the risk prompt module at inference. The risk head localizes and describes the risk object on the multi-scale UNet decoder features of the generated risk frame.

what object to preserve, not where it will be. The risk object is conveyed only through its appearance and semantics, the two location-free conditions that its annotations readily provide.

Appearance. The appearance condition v^a captures how the risk object looks and is extracted from a reference frame paired with a soft risk mask M. During training, the reference is the annotated risk frame x_T, and the mask comes from its annotated box b_T. At inference, the last observed frame serves as the reference and the mask is estimated (Sec. 3.4). A frozen convolutional encoder (ResNet-50; He et al., 2016) computes multi-scale feature maps $\{F_l\}$ of the reference, and each scale is reduced by mask-weighted average pooling:

$$a_l = \frac{\sum_p M_l(p)\, F_l(p)}{\sum_p M_l(p)}, \tag{1}$$

where p indexes spatial positions and M_l is M resampled to the resolution of F_l. Pooling deliberately collapses the spatial layout. The resulting appearance vector a_l encodes how the risk object looks (color, shape, and scale-specific texture) but not where it lies. Each a_l is then expanded into a set of condition tokens by an MLP, which we call a *condition expander*, and the tokens from all scales are concatenated into the appearance condition v^a.

Semantics. The semantic condition v^s captures what the risk object means for driving (its category and behavior). During training, the annotated risk caption c_T is encoded by a frozen CLIP text encoder (Radford et al., 2021) into a semantic vector s. At inference no caption exists, so a CLIP image pathway produces its counterpart $\hat{s}$ in the shared embedding space (Sec. 3.4). A condition expander of the same MLP form, shared by the text-encoded s and the image-estimated $\hat{s}$, expands the semantic vector into condition tokens v^s. The two conditions are thus complementary: appearance preserves the risk object's instance-specific visual detail that language abstracts away, and semantics supplies its identity and state that pixels alone do not convey.

3.3 Decoupled Cross-Attention Injection

Risk conditions must improve the generation fidelity of the risk object without eroding the world model's pretrained prior (Zhang et al., 2023; Ye et al., 2023). We therefore follow the decoupled design of IP-Adapter (Ye et al., 2023) and attach the conditions to every spatial cross-attention layer of the world model's denoising UNet as parallel branches. Let Q be the query tokens of a layer, and let (K, V) be the projections of its original context. We add condition-specific projections and sum the attention outputs:

$$\mathrm{Attn}^{\star}(Q) = \mathrm{Attn}(Q, K, V) + \mathrm{Attn}\big(Q, W_K^a v^a,\, W_V^a v^a\big) + \mathrm{Attn}\big(Q, W_K^s v^s,\, W_V^s v^s\big). \tag{2}$$

The query and the original projections are reused and kept frozen, as is the entire world-model backbone, so only the new projections and the condition expanders are trained. The additive form injects the conditions only where a query attends to them, modulating the features that render the risk object, while the frozen pathway preserves the pretrained generation behavior. We leave the temporal attention layers untouched, since the conditions are time-invariant descriptors shared by all future frames and the risk object's motion is left to the world model's learned scene dynamics.

3.4 Learnable Risk Prompt Module

At inference no risk annotation exists, so a learnable risk prompt module estimates the conditions from the observed frames and supplies them to the world model as a risk prompt. A ResNet encoder processes each of the k observed frames, and an FPN-style decoder (Lin et al., 2017) fuses them to estimate a soft risk heatmap $\hat{M}$. The heatmap highlights the prospective risk region and plays the role of the mask in Sec. 3.2 (examples in App. D). The appearance vectors $\hat{a}_l$ are pooled from the multi-scale features of the last observed frame under $\hat{M}$ (Eq. 1). For semantics, an adapter fuses the CLIP global embedding of the last observed frame with its patch tokens pooled under $\hat{M}$, and maps the fusion into the caption embedding space, yielding the estimated $\hat{s}$. Because the conditions carry no explicit location information, the risk prompt module never forecasts where the risk object will be. It only estimates which object is becoming critical and distills its appearance and semantics.

3.5 Reading Risk from the Generated Future

The generated risk frame is consumed not as pixels but through the world model's internal representation. At the final denoising step, we take multi-scale feature maps of $\hat{x}_T$ from the decoder of the denoising UNet. A region-to-text risk head takes these features and outputs the box $\hat{b}_T$ and the caption $\hat{c}_T$: it first detects the risk object and then describes it, so every caption rests on a detection (details in App. A). Under this design, generation fidelity translates into risk understanding performance: a risk object generated more faithfully is more discernible in the feature maps, enabling more reliable localization and captioning, as Sec. 4.3 and Sec. 4.4 confirm empirically.

3.6 Training Strategy

Training of RAWM proceeds in three stages: the world model learns to use the risk conditions, the risk prompt module learns to estimate them, and the module is finally refined in the joint pipeline.

Stage I: Risk-Conditioned Adaptation. This stage injects the ground-truth conditions extracted from risk annotations into the frozen world model and trains the new attention projections and the condition expanders. The training objective is the world model's original diffusion loss, spatially reweighted toward the risk region:

$$\mathcal{L}_{\text{gen}} = \mathbb{E}\big[\, W \odot \ell_\theta \,\big], \tag{3}$$

where ℓ_θ denotes the per-location denoising error computed with the conditions injected, and W amplifies the error inside the risk box. The reweighting concentrates adaptation on the risk object, otherwise dominated by the background.

Stage II: Condition Estimation. This stage trains the prompt module alone to reproduce ground-truth conditions:

$$\mathcal{L}_{\text{pred}} = \sum_l \big(1 - \cos(\hat{a}_l, a_l)\big) + \big(1 - \cos(\hat{s}, s)\big) + \mathcal{R}(\hat{M}), \tag{4}$$

where the first two terms align the estimated appearance and semantic vectors with their ground-truth counterparts via cosine similarity, and $\mathcal{R}(\hat{M})$ regularizes the estimated risk heatmap to match the total mass of the ground-truth risk mask M and stay concentrated.

Stage III: Generation-Guided Refinement. This stage injects the estimated conditions and refines the risk prompt module under the generation objective in Eq. 3, retaining the alignment and regularization terms of Eq. 4 at reduced weights (App. A), with gradients flowing through the frozen world model. The prompt module therefore learns not merely to match the ground-truth conditions but to supply conditions that best support faithful generation of the risk object.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Dataset and protocol. We evaluate on DRAMA (Malla et al., 2023), the benchmark that defines the joint risk localization-and-captioning task, which we extend to the anticipatory setting. It contains 17,785 interactive driving clips collected in Tokyo, each annotated at its final frame, the risk moment, with the bounding box and free-form caption of the risk object. A native clip contains 10 frames at about 3 Hz, which we temporally interpolate to 10 Hz with RIFE (Huang et al., 2022). The world model observes $k{=}3$ consecutive frames of the clip and generates the future up to the risk moment. Our main results use a 1.5 s anticipation horizon, and Sec. 4.5 sweeps it from 0.7 s to 2.5 s.

Metrics. Localization is measured by Mean-IoU and Acc@0.5 (Malla et al., 2023). Captioning is measured by BLEU-1–4 (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), ROUGE-L (Lin, 2004), and SPICE (Anderson et al., 2016). The fidelity of the generated future (Sec. 4.3) is measured by PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018), computed over the full frame and within the risk region. All metrics are formally defined in App. B.

Compared methods. All compared methods except the trajectory extrapolation baselines use a risk head of identical architecture and training recipe, trained per method on its own input features, so differences in risk understanding performance can be attributed to the input features rather than the head. The compared methods are as follows. (i) *Direct-prediction baseline*: this method predicts the risk directly from the early observation, using a ViT (Dosovitskiy et al., 2021) to encode the last observed frame, with no generation. Since the observed frame does not contain the risk object's future position, we report the method on captioning only. (ii) *Trajectory extrapolation baselines*: these methods forecast the risk box from its observed motion, extrapolating its past trajectory rather than generating any future scene: Constant Velocity (CV; Schöller et al., 2020) extrapolates the box from its most recent displacement, and Future Object Localization (FOL; Yao et al., 2019) predicts it with a recurrent network from observed boxes and optical flow (App. C). Both receive the ground-truth past trajectory and are reported on localization only, as they produce no caption. (iii) *Baseline WM*: the same world model, which generates the future risk frame without the risk conditions. (iv) *RAWM (ours)*: the full pipeline, which generates the future risk frame under the conditions estimated by the prompt module from the observed frames. (v) *GT-future oracle*: a ViT encodes the ground-truth future frame at the risk moment, and risk understanding is performed directly on it. As this frame is never available as input to our task, the method serves as an oracle upper bound rather than a competitor (App. D reports a ResNet variant).

Implementation. The world model is Vista (Gao et al., 2024b) with its backbone frozen throughout. The new attention projections and condition expanders amount to 83 M trainable parameters, and the risk prompt module to 31 M. The risk head follows GRiT (Wu et al., 2024), a region-to-text model that jointly detects and describes objects, and consumes three feature scales.

4.2 MAIN RESULTS: ANTICIPATORY DRIVING RISK UNDERSTANDING

Table 1 yields three observations. First, world-model generation leads the non-generative alternatives on both halves of the task, localization and captioning. On captioning, even the baseline world model with no risk conditioning surpasses direct prediction (SPICE 47.22 vs 45.96), and RAWM extends the margin (50.24). On localization, RAWM leads both trajectory extrapolation baselines (Mean-IoU 62.89 vs 36.51 for CV and 54.40 for FOL) even though they receive the ground-truth past trajectory of the risk box. Sec. 4.5 shows both margins widening with the horizon. Second, risk conditioning brings a further gain on top of world-model generation. RAWM consistently lifts all metrics over the baseline world model (Mean-IoU 59.25 $\rightarrow$ 62.89, SPICE 47.22 $\rightarrow$ 50.24). Since the two differ only in the risk conditions, steering generation toward the risk object translates into better risk understanding. Moreover, the estimated conditions sacrifice little: the pipeline nearly matches the same model under ground-truth conditioning (62.89 vs 63.56 Mean-IoU; App. D). Third, the remaining gap to the GT-future oracle concentrates in localization. RAWM approaches the oracle on captioning (SPICE 50.24 vs 52.41), while localization lags further behind (Mean-IoU 62.89 vs

Table 1: Anticipatory risk understanding on DRAMA at the 1.5 s horizon (Mean-IoU, Acc@0.5, BLEU-1/4, METEOR, ROUGE-L, SPICE). The direct-prediction baseline (captioning-only) and the trajectory extrapolation baselines (localization-only) are the non-generative alternatives. GT-future is an oracle that observes the real future risk frame.

Method	IoU	Acc	B-1	B-4	MET.	R-L	SP.
Direct pred.	–	–	71.29	48.35	37.04	65.32	45.96
CV (extrap.)	36.51	18.73	–	–	–	–	–
FOL (extrap.)	54.40	57.29	–	–	–	–	–
Baseline WM	59.25	72.94	71.24	49.10	37.21	66.15	47.22
RAWM	**62.89**	**75.48**	**72.43**	**51.66**	**38.45**	**68.39**	**50.24**
GT-future	73.17	81.40	74.08	54.25	39.79	70.37	52.41

Table 2: Reconstruction fidelity of the generated risk frame, over the full frame and the risk region.

	Full frame			Risk region		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Baseline WM	18.37	0.861	0.218	16.03	0.674	0.222
RAWM	**18.64**	**0.870**	**0.215**	**18.17**	**0.754**	**0.187**

Table 3: Ablation of the appearance and semantic conditions.

Conditioning	IoU	Acc	B-4	MET.	R-L	SP.
None (baseline)	59.25	72.94	49.10	37.21	66.15	47.22
Semantics only	60.82	74.03	50.35	37.59	67.46	48.13
Appearance only	62.05	74.80	51.12	38.06	67.90	49.58
Both (RAWM)	**62.89**	**75.48**	**51.66**	**38.45**	**68.39**	**50.24**

73.17). The remaining gap thus lies mainly in predicting the object's future position rather than in rendering the object itself, consistent with the positional drift analyzed in Sec. 4.6.

4.3 Fidelity of the Generated Future

Compared with the baseline world model, RAWM improves all six reconstruction fidelity measures, each computed between the generated risk frame and its ground-truth counterpart (Table 2). The gain is large in the risk region (PSNR +2.14, SSIM +0.080, LPIPS -0.035) and small but consistently positive over the full frame (PSNR +0.27, SSIM +0.009, LPIPS -0.003). Conditioning therefore concentrates its benefit on the risk object without degrading the rest of the scene. This fidelity gain in the risk region is the mechanism behind the risk understanding gains over the baseline world model in Table 1, as the ablation in Sec. 4.4 supports.

4.4 Ablation Study

We ablate the two risk conditions at the 1.5 s horizon, conditioning generation on appearance alone or semantics alone, and compare with the baseline world model and full RAWM (Table 3). Each condition alone improves over the baseline on all metrics, and the two together perform best, confirming that both conditions are effective and complementary. Appearance brings the larger gain on localization and captioning: relative to the full gain of RAWM over the baseline, it alone recovers 77% on Mean-IoU and 79% on B-4, while semantics alone recovers 43% and 49%. Appearance contributes more because it supplies the visual detail that makes the object recognizable in the generated frame, while adding semantics on top of appearance brings a further gain (62.05 $\rightarrow$ 62.89 Mean-IoU), as semantics conveys the object's identity and state that appearance lacks. On the generation side, the ablated models are scored by reconstruction fidelity within the risk region, and the two conditions show the same relative strengths: appearance recovers 74–80% of the full fidelity gain (PSNR, SSIM, and LPIPS) and semantics 31–53% (per-metric results in App. E). The two ablations

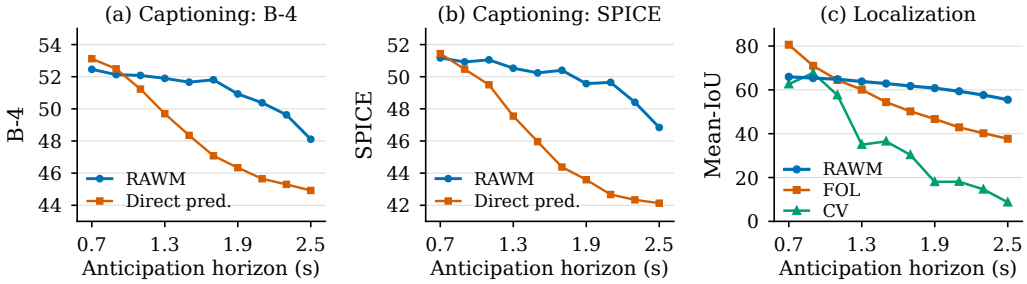


Figure 3: Risk understanding versus anticipation horizon (0.7–2.5 s). (a), (b) Captioning: the margin of RAWM over the direct-prediction baseline widens as the horizon grows. (c) Localization: the extrapolation baselines collapse at long horizons, while RAWM degrades gracefully.

show a consistent pattern, upstream on generation fidelity and downstream on risk understanding, confirming that both conditions act through the same route, a more faithfully generated risk object.

4.5 Effect of the Anticipation Horizon

Sliding the observation window, we vary the anticipation horizon over ten values, from 0.7 s to 2.5 s in 0.2 s steps, comparing RAWM with the direct-prediction baseline on captioning and the trajectory extrapolation baselines on localization (Fig. 3; full tables in App. F). Across this range, RAWM itself degrades gracefully, without sharp drops: Mean-IoU falls from 65.96 at 0.7 s to 55.52 at 2.5 s, while captioning declines more slowly (SPICE 51.17 $\rightarrow$ 46.84).

Against the direct-prediction baseline, the advantage of world-model generation on captioning grows with the horizon (Fig. 3a,b). At 0.7 s direct prediction even leads marginally, since its observed frame is still a faithful proxy for the near future, but it falls away as the horizon lengthens. The SPICE margin of RAWM grows from half a point at 0.9 s to about seven points around 2 s. The reason is that captioning needs a view of the future risk moment: the observed frame drifts ever further from it as the horizon grows, whereas RAWM re-synthesizes that moment.

Against the trajectory extrapolation baselines, the advantage of world-model generation on localization emerges at long horizons: the curves form a scissors pattern with a crossover at about 1.1 s (Fig. 3c). Before the crossover the baselines lead, as the setup deliberately favors them: they are handed the recent positions of the risk box and simply extrapolate the future box, while RAWM must both generate the future and detect the risk within it, bearing detection errors that the baselines never face. Beyond the crossover, both baselines fall below RAWM and then collapse, reaching 8.7 (CV) and 37.7 (FOL) Mean-IoU at 2.5 s against 55.5 for RAWM. Extrapolation is blind to the scene dynamics that redirect an agent, so its error compounds, whereas RAWM inherits these dynamics from large-scale driving videos and places the risk object at a plausible location.

The two comparisons cover the two halves of the task and point to the same conclusion. Direct prediction sustains captioning, and trajectory extrapolation sustains localization, only at short horizons. Generating the future with RAWM sustains both, with widening margins. Anticipatory risk understanding therefore calls for a generative world model that imagines the future, rather than a predictor that extrapolates into it. To our knowledge, this is the first characterization of how driving risk understanding varies with the anticipation horizon. These curves quantify what each additional second of reaction time costs in understanding accuracy, a trade-off any driving system must weigh.

4.6 Qualitative Results

Fig. 4 compares the baseline world model and RAWM on the generated risk frames and the resulting risk understanding. On the baseline's frames, the low generation fidelity of the risk object leads to all three failure modes: the risk head localizes a wrong object (the motorcycle and sedan scenes), misdescribes the risk object (the van and pedestrian scenes), or detects nothing (the hatchback scene). RAWM renders the same objects faithfully enough that the risk head localizes and describes them

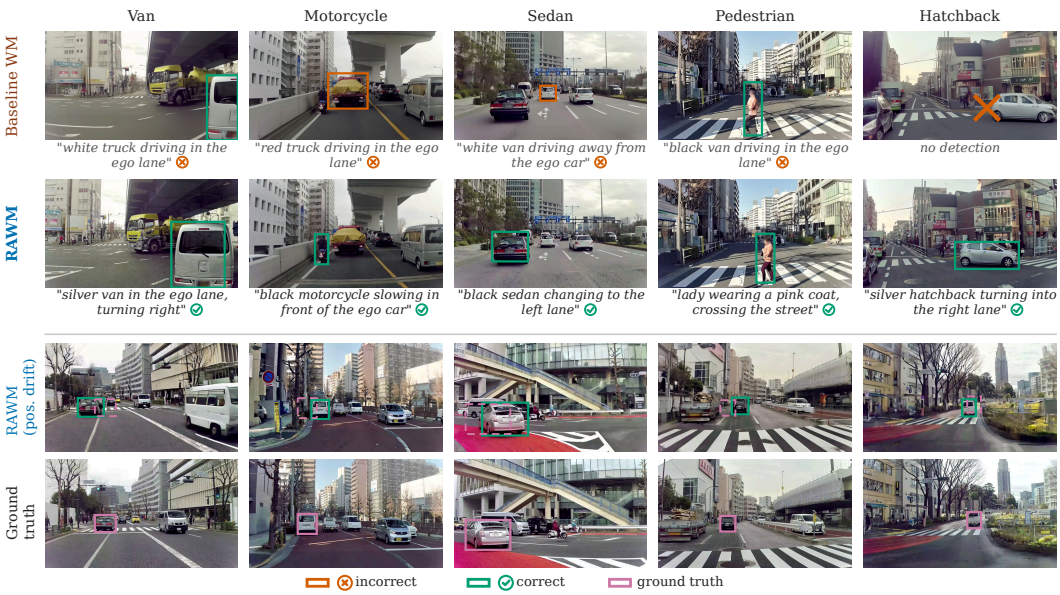


Figure 4: Qualitative comparison between the baseline world model and RAWM. The top two rows show generated risk frames with predicted boxes and captions: the risk head fails on the baseline's frames but localizes and describes the risk object correctly on RAWM's. The bottom two rows illustrate positional drift, the dominant source of RAWM's localization gap, with deliberately pronounced cases: the risk object is localized correctly in the generated frame (green box) but offset from the ground truth, marked by dashed and solid pink boxes on the generated and real risk frames.

correctly. The bottom two rows focus on positional drift, with deliberately pronounced cases. The risk object is rendered well and localized correctly in the generated frame, but the generated scene drifts from the real future, so the box scores poorly against the ground-truth location. This is why the gap to the GT-future oracle in Table 1 concentrates in localization. Tracing the localized object back to its real counterpart in the observed frames and scoring there would factor the offset out. We nonetheless follow the forecasting convention and evaluate on the generated frame at the future moment, so the reported numbers include the unavoidable cost of generation-induced drift.

5 Conclusion

We introduce anticipatory driving risk understanding, a new forecasting task that proactively localizes and describes a potential risk object as it will appear at a future risk moment, before the risk turns critical, and address it with a generate-then-perceive paradigm. Our Risk-Augmented World Model strengthens the fidelity of the risk object in the generated future by conditioning generation on its appearance and semantics through a dual-branch conditioning module, with both estimated by the risk prompt module at inference. RAWM improves all localization and captioning metrics over the baseline world model, and ablations trace these gains to the improved fidelity of the generated risk object. It also increasingly surpasses the direct-prediction and trajectory-extrapolation alternatives as the anticipation horizon grows.

Several limitations point to concrete next steps. The remaining gap to the GT-future oracle concentrates in localization and mainly comes from positional drift of a faithfully rendered risk object, which calls for stronger modeling of agent motion in the world model. Our protocol also follows the benchmark in fixing the risk moment and annotating a single risk object, leaving free-running horizons and multi-object risk open. Finally, the generated risk frame that RAWM delivers depicts the impending situation in full visual detail, beyond what boxes and captions can convey. A natural next step is to couple it to downstream planning, so that the ego vehicle can reshape its trajectory against a risk foreseen before it actually arrives.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing, such as checking grammar and refining wording to make the manuscript clearer and easier to read. We have not used generative AI tools to help develop conceptual frameworks, propose or refine hypotheses, design experiments, implement methods, clean or reformat datasets, or interpret results, and generating synthetic datasets, formulating or proving mathematical claims, writing proofs, translation, and qualitative or thematic data analysis are not applicable to this work. We have reviewed all AI-assisted work, and every suggested edit was checked by the authors to ensure that it preserves the intended technical meaning. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work uses the DRAMA dataset under its academic, non-commercial license and involves no new data collection or human subjects. All pretrained models are used under their public licenses (App. H). The task aims to improve driving safety by anticipating risks before they become critical. Generated future scenes are model predictions rather than observations: they can drift from the real future (Sec. 4.6) and should support, not replace, a driver's or a driving system's own perception.

REPRODUCIBILITY STATEMENT

We have made the following efforts to ensure the reproducibility of our work. Sec. 3 describes the complete method, and Sec. 4.1 specifies the dataset, the evaluation protocol, the metrics, and the compared methods. App. A provides the implementation details of every module and training stage, including the hyperparameters, the computing environment, and the random seeds. App. B formally defines all metrics, and App. C details the trajectory extrapolation baselines. App. H lists the dataset and pretrained models used, together with their licenses. The supplementary material contains the code for all components of the pipeline, together with the training and inference configurations.

REFERENCES

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

Sainithin Artham, Shankar Gangisetty, Avijit Dasgupta, and C. V. Jawahar. DriveSafe: A framework for risk detection and safety suggestions in driving scenarios. *arXiv preprint arXiv:2605.16892*, 2026.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. HiLM-D: Enhancing MLLMs with multi-scale high-resolution details for autonomous driving. *International Journal of Computer Vision*, 133:5379–5395, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Jiaqi Fan, Rui Zhang, Yafei Wang, Hongqing Chu, Bingzhao Gao, and Qingwei Liu. MLLM-SUL: Multimodal large language model for semantic scene understanding and localization in traffic scenarios. In *2025 7th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI)*, pp. 321–326, 2025.

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. MagicDrive: Street view generation with diverse 3d geometry control. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024a.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

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 (CVPR)*, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Anthony Hu, Gianluca Corrado, Nicolas Griffiths, Zak Murez, Corina Gurau, Hudson Yeo, Alex Kendall, Roberto Cipolla, and Jamie Shotton. Model-based imitation learning for urban driving. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Qifeng Li, Xiaosong Jia, Shaobo Wang, and Junchi Yan. Think2drive: Efficient reinforcement learning by thinking with latent world model for autonomous driving (in CARLA-V2). In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. DRAMA: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1043–1052, 2023.

Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, Liping Jing, Yiming Nie, and Bin Dai. DriveWorld: 4D pre-trained scene understanding via world models for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, 2002.

Mingxing Peng, Yuting Xie, Xusen Guo, Ruoyu Yao, Hai Yang, and Jun Ma. LD-Scene: LLM-guided diffusion for controllable generation of adversarial safety-critical driving scenarios. *arXiv preprint arXiv:2505.11247*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. GAIA-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

Christoph Schöller, Vincent Aravantinos, Florian Lay, and Alois Knoll. What the constant velocity model can teach us about pedestrian motion prediction. *IEEE Robotics and Automation Letters*, 5(2):1696–1703, 2020.

Benjamin Stoler, Juliet Yang, Jonathan Francis, and Jean Oh. RCG: Safety-critical scenario generation for robust autonomous driving via real-world crash grounding. *arXiv preprint arXiv:2507.10749*, 2025.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. DriveDreamer: Towards real-world-driven world models for autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024a.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. GRiT: A generative region-to-text transformer for object understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Yu Yang, Jianbiao Mei, Yukai Ma, Siliang Du, Wenqing Chen, Yijie Qian, Yuxiang Feng, and Yong Liu. Driving in the occupancy world: Vision-centric 4D occupancy forecasting and planning via world models for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9327–9335, 2025.

Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual point cloud forecasting enables scalable autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Yu Yao, Mingze Xu, Chiho Choi, David J. Crandall, Ella M. Atkins, and Behzad Dariush. Egocentric vision-based future vehicle localization for intelligent driving assistance systems. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. DriveDreamer-2: LLM-enhanced world models for diverse driving video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10412–10420, 2025.

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. OccWorld: Learning a 3D occupancy world model for autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Jiawei Zhou, Linye Lyu, Zhuotao Tian, Cheng Zhuo, and Yu Li. SafeMVDrive: Multi-view safety-critical driving video synthesis in the real world domain. *arXiv preprint arXiv:2505.17727*, 2025.

APPENDIX

App. A gives implementation details for each component of RAWM (Sec. 3). App. B defines the evaluation metrics of Sec. 4.1, and App. C details the trajectory extrapolation baselines compared there. App. D extends the experiments of Secs. 4.2–4.4 at the main horizon (an oracle encoder variant, ground-truth versus estimated conditions, complete ablation metrics, and the estimated risk heatmaps), and App. E adds generation-quality results (the fidelity ablation and distributional scores). App. F reports the full anticipation-horizon results (Sec. 4.5), App. G discusses the scope and limitations of our study, and App. H lists assets and licenses. The code in the supplementary material implements all components described here.

A IMPLEMENTATION DETAILS

A.1 DUAL-BRANCH CONDITIONING MODULE

Following the design in Secs. 3.2 and 3.3, the appearance condition uses 4 tokens for each of the three feature scales of the appearance encoder (12 tokens in total), the semantic condition uses 8 tokens, and each token is 1024-dimensional. The conditions are injected into all 16 spatial cross-attention layers of the denoising UNet, while the temporal attention layers are unchanged. The appearance encoder is a frozen ResNet-50 (He et al., 2016) whose three intermediate feature maps, of sizes 72×128, 36×64, and 18×32 for the 576×1024 input, provide the three scales. The semantic branch uses a frozen CLIP ViT-L/14 (Radford et al., 2021): its text encoder embeds the annotated caption during training, and its image encoder supplies the embedding from which the semantic condition is estimated at inference (Sec. 3.4). The risk masks are defined on the 72×128 grid: the ground-truth risk mask M is the annotated box rasterized with fractional coverage at the boundary, and the estimated mask is the heatmap $\hat{M}$ that the risk prompt module outputs on the same grid. Multi-scale masks are obtained by area downsampling, so the pooling in Eq. 1 follows the same convention for ground-truth and estimated masks.

A.2 RISK PROMPT MODULE

The ResNet encoder of Sec. 3.4 is the frozen ResNet-50 of Sec. A.1, and the FPN-style decoder (Lin et al., 2017) that fuses its per-frame feature maps into the risk heatmap $\hat{M}$ uses 256 channels per level and outputs on the 72×128 grid. The appearance branch reuses the same encoder, so the estimated vectors $\hat{a}_l$ and their ground-truth counterparts lie in one feature space, which is what the alignment terms of Eq. 4 compare. For semantics, the adapter is a two-layer MLP of width 2048 with a residual connection from the CLIP global embedding, and the caption embedding space it maps into is the 768-dimensional output space of the CLIP text encoder.

A.3 TRAINING

Stage I trains the new attention projections and the condition expanders (82.9 M trainable parameters) with the risk-weighted diffusion loss (Eq. 3). The weight W is 5 inside the annotated risk box and 1 at all other locations, normalized per sample to unit mean, which keeps the loss scale unchanged. In this stage and in Stage III, each condition is independently dropped to zero with probability 0.15, so the model also learns the unconditional branch that classifier-free guidance uses at inference (Sec. A.4). The learning rate is 2×10^{-4} with a per-GPU batch size of 1, and training runs 55 epochs.

Stage II trains the risk prompt module (31.4 M trainable parameters) under the objective in Eq. 4, with a learning rate of 1×10^{-4} and a per-GPU batch size of 4, and runs for 30 epochs, in parallel with Stage I, since the two stages share no trainable parameters. Within $\mathcal{R}(\hat{M})$, the total-mass term is weighted 1.0 and the concentration term 0.5. Because the ground-truth and the estimated vectors pass through the same condition expanders, the vector-level alignment carries over to the injected condition tokens.

Stage III injects the estimated conditions, initializes the dual-branch conditioning module from the final Stage I checkpoint and the risk prompt module from the final Stage II checkpoint, and refines

the prompt module under the Stage I generation objective for 16 epochs with gradients through the frozen world model. The learning rate is 3×10^{-5} with a per-GPU batch size of 1, and the alignment and regularization terms of Eq. 4 are kept at $0.1\times$ their original weights.

Training uses 4 NVIDIA H20 GPUs, and the three stages take roughly 19 days, 8 hours, and 6 days, respectively. All experiments run on Ubuntu 22.04 with Python 3.10 and PyTorch 2.4.1 (CUDA 11.8, cuDNN 9.1). Training and sampling fix random seeds via `seed_everything` (seed 23), including the initial noise of the diffusion sampler, and all reported results are obtained from a single training and evaluation run per configuration.

A.4 Inference

Sampling uses the Euler-EDM sampler with 50 steps and classifier-free guidance whose scale increases linearly across the generated frames from 1.0 to 2.5, generating from the 3 observed frames to the risk moment. Features for the risk head are taken at the final denoising step from the decoder blocks immediately after each upsampling-and-skip fusion, giving three scales (72×128, 36×64, 18×32) with 640, 1280, and 1280 channels.

A.5 Risk Head

The head follows GRiT (Wu et al., 2024), which comprises a visual encoder, a foreground object extractor, and a text decoder. We remove the visual encoder and instead align the three input feature scales to the extractor's feature-pyramid interface with a lightweight adapter. The foreground object extractor, a two-stage design with a proposal generator and an RoI head, localizes candidate risk regions on the aligned features, and a 6-layer transformer text decoder autoregressively generates the caption of a detected region from its features, so every caption rests on a detection. The extractor and the decoder are trained jointly end to end. At inference, the top-scoring detection is taken as the risk object and reported together with its caption.

One head is trained for each compared method except the trajectory extrapolation baselines, with identical architecture and training recipe, which enables the controlled comparison in Sec. 4.1. RAWM and the Baseline WM use the UNet decoder features described in Sec. A.4, while the direct-prediction baseline and the GT-future oracle use ViT-B (Dosovitskiy et al., 2021) features of their respective input frames (the last observed frame and the ground-truth risk frame). In every case the feature provider stays frozen while the head is trained, so differences in risk understanding performance trace back to the quality of the input features rather than to the head.

B Evaluation Metrics

Localization follows the DRAMA protocol (Malla et al., 2023). Mean-IoU averages the intersection-over-union between the predicted and the ground-truth box of the risk object at the risk moment, and Acc@0.5 is the fraction of samples whose IoU exceeds 0.5. The analyses in the main text lead with Mean-IoU because it is continuous in the box overlap, whereas the thresholded Acc@0.5 becomes erratic when the IoU distribution sits near the threshold (Sec. F.1). Captioning uses the standard captioning metrics: BLEU-n (Papineni et al., 2002) measures n-gram precision (the main text reports BLEU-1 and BLEU-4; the tables here add BLEU-2 and BLEU-3), METEOR (Banerjee & Lavie, 2005) aligns unigrams with stem and synonym matching, ROUGE-L (Lin, 2004) measures the longest common subsequence, and SPICE (Anderson et al., 2016) compares semantic propositional content extracted from scene graphs. The analyses emphasize SPICE as the most semantic of the suite, closest to whether the caption conveys the right risk.

Generation quality is measured by two complementary families. PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) are computed between the generated risk frame and its ground-truth counterpart, over the full frame and within the risk region, and measure reconstruction fidelity. FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018) compare the feature distributions of generated frames and videos, respectively, with those of real ones, and measure overall realism (Sec. E.2).

C Trajectory Extrapolation Baselines

C.1 Recovering Risk-Object Trajectories

DRAMA annotates the risk box only at the risk moment, so the trajectory extrapolation baselines cannot be run from the raw data. We hand them a trajectory recovered by propagating the annotated box backward through the clip with RAFT optical flow (Teed & Deng, 2020). Grid points inside the box are warped frame by frame, and a box is refit at each step from the 5th to 95th percentiles of the warped point coordinates, which recovers both translation and scale. We additionally extract a $5\times5\times2$ RoI-pooled flow feature per frame for FOL. The recovered trajectories are anchored to the ground-truth box and temporally smooth, which favors the baselines. The baselines nonetheless collapse at long anticipation horizons (Sec. 4.5).

C.2 Constant Velocity

CV estimates the velocity of the risk-box center as its displacement between the last two observed steps and extrapolates its future position linearly. This treatment of the center follows the strict definition of Schöller et al. (2020), which predicts point positions only. Because the IoU protocol scores a full box rather than a point position, our adaptation applies the same constant-velocity assumption to the box width and height, extrapolating them linearly in the same way.

C.3 FOL Re-Implementation

Future Object Localization (FOL; Yao et al., 2019) is re-implemented following the official code and the equations of the paper. We keep the original design in the following respects. The box stream (4-dimensional normalized boxes) and the flow stream (50-dimensional RoI flow) are encoded by separate GRUs. The decoder state is initialized from the average of the two final hidden states passed through a projection with ReLU. Each decoding step receives a learned transform of the previous hidden state and predicts the displacement of the future box relative to the last observed box. The loss is the per-step Euclidean distance summed over all future steps. We deviate in two ways. First, the ego-motion stream is removed, since DRAMA provides no odometry. Second, the observed history is three frames, matching the observation window of the world model. We therefore describe the baseline as an adapted re-implementation rather than the official model.

D Extended Results at the Main Horizon

D.1 A ResNet Variant of the GT-Future Oracle

The last two rows of Table 4 run the GT-future oracle with a ResNet-50 encoder in place of the ViT. The main text reports the ViT variant because it is the stronger of the two and matches the encoder of the direct-prediction baseline, keeping them directly comparable. The ViT-B and ResNet-50 oracles score closely on all metrics, and the ResNet oracle also remains well above the full RAWM pipeline on localization (71.53 vs 62.89 Mean-IoU). The oracle's localization margin is therefore not an artifact of one particular encoder.

D.2 Ground-Truth versus Estimated Conditions

The Stage I rows of Table 4 evaluate risk understanding when the ground-truth conditions, extracted from the annotated risk frame, are injected in place of the estimated ones. The evaluated models are Stage I checkpoints (Sec. 3.6) at epochs 15, 30, and 54. Two observations follow. First, understanding improves monotonically as Stage I training proceeds (Mean-IoU 61.60 $\rightarrow$ 62.14 $\rightarrow$ 63.56), and every checkpoint exceeds the baseline world model on all metrics, so the conditioning mechanism is effective on its own. Second, the full pipeline with estimated conditions reaches 62.89 Mean-IoU and 50.24 SPICE, within 0.67 and 0.16 of the epoch-54 row, which serves as the natural reference because Stage III training starts from that checkpoint (Sec. A.3). Relative to the baseline, ground-truth conditioning improves Mean-IoU by 4.31 and SPICE by 3.18. Estimating the conditions preserves 84% of the former gain and 95% of the latter. Therefore, the estimated conditions sacrifice little, as claimed in Sec. 4.2.

Table 4: Extended results at the 1.5 s horizon. The first block is the condition ablation of Sec. 4.4 on all metrics, the second replaces the estimated conditions with the ground-truth ones at three Stage I checkpoints, and the third is the GT-future oracle under two encoders. The Baseline WM and RAWM rows are those of Table 1.

	Localization		Captioning						
Method	IoU	Acc	B-1	B-2	B-3	B-4	MET.	R-L	SP.
Baseline WM (no conditions)	59.25	72.94	71.24	63.78	56.48	49.10	37.21	66.15	47.22
Semantics only	60.82	74.03	71.38	64.29	57.32	50.35	37.59	67.46	48.13
Appearance only	62.05	74.80	72.07	64.87	57.98	51.12	38.06	67.90	49.58
RAWM (both conditions)	62.89	75.48	72.43	65.31	58.50	51.66	38.45	68.39	50.24
Stage I ep. 15 (GT conditions)	61.60	74.66	71.69	64.54	57.39	50.16	38.11	68.13	47.55
Stage I ep. 30 (GT conditions)	62.14	75.07	72.15	64.96	57.92	50.81	38.17	68.24	48.36
Stage I ep. 54 (GT conditions)	63.56	76.11	72.51	65.38	58.62	51.83	38.52	68.49	50.40
GT-future oracle (ViT-B)	73.17	81.40	74.08	67.40	60.95	54.25	39.79	70.37	52.41
GT-future oracle (ResNet-50)	71.53	80.32	73.71	66.93	60.39	53.58	39.44	69.83	51.76

D.3 Complete Metrics for the Condition Ablation

The first block of Table 4 reports the condition ablation in Sec. 4.4 on the full metric suite. Table 3 omits BLEU-1, BLEU-2, and BLEU-3 to fit the column width. The complete results show the same ordering on all metrics: each condition alone improves over the baseline, appearance contributes more than semantics, and the two together perform best.

D.4 Estimated Risk Heatmaps

Complementing the comparison of ground-truth and estimated conditions above, Fig. 5 visualizes estimated risk heatmaps $\hat{M}$ on example clips. From the early observed frames alone, the risk prompt module concentrates the heatmap mass on the object that is about to become critical, across object categories, scales, and distances (large construction and utility vehicles, sedans, crossing and roadside pedestrians, and traffic lights), which supports the heatmap's role as a pooling weight over the observed scene (Sec. 3.4). This ability originates in the appearance alignment of Stage II rather than in any location supervision. The ground-truth appearance vectors are pooled from the features of the risk frame under the mask of its annotated box and therefore encode how the risk object looks at the risk moment, while their estimated counterparts are pooled from the features of the last observed frame under $\hat{M}$ (Eq. 1). Because an object's appearance changes little over the anticipation horizon, the estimated vectors can match their ground-truth targets only when $\hat{M}$ concentrates its mass on the same object in the observed scene. Minimizing the alignment terms of Eq. 4 therefore drives the heatmap onto the emerging risk object.

E Additional Generation-Quality Results

E.1 Condition Ablation on Risk-Region Fidelity

Table 5 applies the condition ablation of Sec. 4.4 to the generation side, with the same four configurations at the 1.5 s horizon: no risk conditions, semantics alone, appearance alone, and both. Each configuration is scored by reconstruction fidelity within the risk region, measured by PSNR, SSIM, and LPIPS between the generated risk frame and its ground-truth counterpart (App. B). Each condition alone improves risk-region fidelity over the baseline world model on all three measures (PSNR 16.03 $\rightarrow$ 17.16 with semantics alone and 16.03 $\rightarrow$ 17.74 with appearance alone), and conditioning on both is best on all three (PSNR 18.17, SSIM 0.754, LPIPS 0.187). Relative to the full gain from no conditioning to both, appearance alone recovers 74–80% (79.9% on PSNR, 77.5% on SSIM, and 74.3% on LPIPS) and semantics alone 31–53% (52.8%, 47.5%, and 31.4%).

The fidelity ablation thus parallels the understanding ablation in Table 3, where appearance alone recovers 77% of the Mean-IoU gain and semantics alone 43%: the two conditions show the same



Figure 5: Estimated risk heatmaps $\hat{M}$ overlaid on the observed scene. The risk prompt module highlights the emerging risk object before the risk moment, across object categories and scales.

Table 5: Risk-region reconstruction fidelity under the condition ablation (1.5 s horizon).

Conditioning	PSNR↑	SSIM↑	LPIPS↓
None (baseline)	16.03	0.674	0.222
Semantics only	17.16	0.712	0.211
Appearance only	17.74	0.736	0.196
Both (RAWM)	**18.17**	**0.754**	**0.187**

relative strengths. Together, the two ablations support the conclusion of Sec. 4.4 that both conditions improve risk understanding through the same route, raising the fidelity of the generated risk object, on which the risk head bases the box and the caption.

E.2 Distributional Metrics

Table 6 reports FID and FVD for the baseline world model and RAWM. Both distributional scores regress under conditioning. The two metric families answer different questions. PSNR, SSIM, and LPIPS (Table 2) are computed against the ground-truth future frame of the same clip, so they measure how faithfully the generated risk frame, and in particular its risk object, is reconstructed. FID and FVD instead compare the feature distributions of the generated frames and videos with those of the real ones over the whole evaluation set, without matching each generated frame to its ground-truth counterpart, and therefore measure overall realism. Risk conditioning steers generation toward rendering the risk object faithfully: reconstruction fidelity improves on all measures, most strongly inside the risk region, while the conditioned frames deviate further from the feature distribution of

Table 6: Distributional scores (FID and FVD) at the 1.5 s horizon.

Method	FID↓	FVD↓
Baseline WM	5.94	54.31
RAWM	11.99	78.82

Table 7: The direct-prediction baseline across anticipation horizons (captioning only; the observed frame does not contain the risk object's future position, so localization is not comparable).

h (s)	B-1	B-2	B-3	B-4	MET.	R-L	SP.
0.7	73.08	66.28	59.45	53.12	39.05	69.35	51.44
0.9	72.70	65.94	58.82	52.49	38.54	68.62	50.47
1.1	72.39	65.07	58.02	51.23	38.26	67.93	49.50
1.3	71.74	63.91	56.64	49.70	37.57	66.52	47.55
1.5	71.29	62.90	55.39	48.35	37.04	65.32	45.96
1.7	70.77	62.03	54.36	47.08	36.49	64.18	44.38
1.9	70.42	61.46	53.62	46.34	36.19	63.56	43.59
2.1	70.14	60.95	53.16	45.65	35.87	62.94	42.67
2.3	69.97	60.68	52.73	45.30	35.75	62.63	42.34
2.5	69.84	60.43	52.41	44.92	35.58	62.30	42.13

Table 8: Localization across horizons for the trajectory extrapolation baselines and RAWM (Mean-IoU and Acc@0.5).

	Mean-IoU			Acc@0.5		
h (s)	CV	FOL	RAWM	CV	FOL	RAWM
0.7	62.6	80.6	66.0	73.2	96.6	77.3
0.9	67.7	71.0	65.4	87.9	86.7	77.0
1.1	57.6	64.5	64.9	78.2	72.4	76.7
1.3	35.0	60.1	63.8	19.8	65.9	76.1
1.5	36.5	54.4	62.9	18.7	57.3	75.5
1.7	30.3	50.2	61.8	8.1	51.4	74.7
1.9	18.0	46.7	60.8	1.0	46.2	73.7
2.1	18.1	42.9	59.4	1.4	42.7	72.2
2.3	14.6	40.2	57.6	0.6	35.9	70.0
2.5	8.7	37.7	55.5	0.2	32.3	67.4

the real frames, which FID and FVD register as a regression. Because the risk head predicts the box and the caption from the risk object as rendered in the generated frame, reconstruction fidelity is the quantity our task depends on, and we report the distributional scores for completeness.

F Full Anticipation-Horizon Results

Tables 7–9 give the numbers behind Fig. 3, at ten horizons from 0.7 s to 2.5 s in 0.2 s steps. The 1.5 s rows are identical to the main-horizon results in Table 1. Fig. 6 plots Mean-IoU and every captioning metric of Tables 7–9 against the horizon: the localization panel compares RAWM with the trajectory extrapolation baselines, and the seven captioning panels compare RAWM with the direct-prediction baseline. On every captioning metric, the margin of RAWM over the direct-prediction baseline widens up to about 2 s and remains clearly positive at 2.5 s.

F.1 Mean-IoU versus Acc@0.5

Table 8 lists both Mean-IoU and Acc@0.5 for CV, FOL, and RAWM at every horizon, whereas Fig. 3c and Fig. 6 plot Mean-IoU. The reason is that Acc@0.5 thresholds each IoU at 0.5 and becomes erratic when the IoU distribution sits near the threshold. The CV baseline illustrates this: between 1.1 s and 1.3 s its Mean-IoU declines from 57.6 to 35.0, yet its Acc@0.5 collapses from 78.2 to 19.8, because the bulk of its IoU distribution slides from just above 0.5 to just below it. Mean-IoU is continuous and therefore the more informative curve. Separately, CV starts non-monotonically

Table 9: RAWM across anticipation horizons on all risk understanding metrics.

	Localization		Captioning						
h (s)	IoU	Acc	B-1	B-2	B-3	B-4	MET.	R-L	SP.
0.7	65.96	77.33	72.93	65.88	59.15	52.46	38.80	69.12	51.17
0.9	65.37	77.01	72.81	65.67	59.18	52.13	38.71	68.87	50.92
1.1	64.85	76.74	72.87	65.63	58.89	52.08	38.73	68.95	51.05
1.3	63.82	76.06	72.56	65.70	58.69	51.90	38.58	68.54	50.53
1.5	62.89	75.48	72.43	65.31	58.50	51.66	38.45	68.39	50.24
1.7	61.76	74.66	72.26	65.14	58.29	51.81	38.34	68.47	50.40
1.9	60.79	73.71	71.94	64.78	57.84	50.93	38.07	67.85	49.57
2.1	59.35	72.17	71.68	64.40	57.42	50.38	37.76	67.36	49.65
2.3	57.64	69.95	71.06	63.72	56.61	49.63	37.27	66.75	48.41
2.5	55.52	67.42	70.28	62.65	55.30	48.11	36.59	65.43	46.84

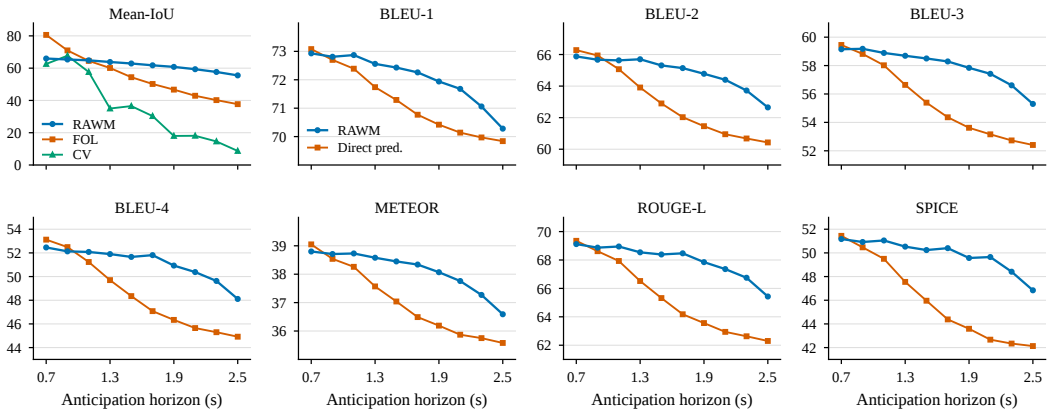


Figure 6: Mean-IoU and the captioning metrics versus anticipation horizon (0.7–2.5 s). The localization panel shows RAWM and the trajectory extrapolation baselines. The captioning panels show RAWM and the direct-prediction baseline.

(Mean-IoU 62.6, 67.7, 57.6 over the first three horizons) because its velocity estimate comes from the displacement between the last two observed boxes, making it sensitive to frame-level noise.

G Scope and Limitations

No-risk scenes and false alarms. Every DRAMA clip is annotated with exactly one risk object at its risk moment, so our protocol measures where and what the risk object will be, not whether a risk is present at all. The task as formulated therefore does not include a no-risk output, and we do not measure false alarms on ordinary driving. Adding an explicit no-risk decision and evaluating it on logs that mix risky and uneventful clips is a necessary step toward deployment, and the present numbers should be read as anticipation quality on clips that do contain a risk.

Baseline comparisons and attribution. The direct-prediction baseline shares the risk head and its training recipe with RAWM, but not its inputs or its capacity: it encodes the last observed frame with a ViT, whereas RAWM observes three frames and carries a pretrained world model together with the conditioning and prompt modules (Sec. 4.1). This comparison therefore shows that generating the future helps under a fixed risk head, but it does not isolate how much of the gain requires generation itself. Two controls would strengthen the claim: an input- and capacity-matched temporal forecaster that predicts the future risk directly from the same observed frames, and a prompt-only variant that feeds the estimated conditions to the risk head without generating the future.

Inference cost and effective lead time. RAWM inherits the cost of diffusion sampling. The world model runs 50 denoising steps over the frames up to the risk moment before the risk head reads the generated risk frame (Sec. A.4), and this stage dominates the end-to-end cost, with the risk head and the prompt module adding comparatively little. This study targets the anticipation quality of the generate-then-perceive paradigm rather than a real-time system, and we report no latency measurements. Whether the effective lead time, the anticipation horizon minus the time to generate and read the future, remains positive depends on the sampler and the hardware. Reducing the number of denoising steps and distilling the sampler would both shorten the generation time, and we leave this to future work.

H Assets and Licenses

DRAMA is used under its academic, non-commercial license. The pretrained components are used under their respective public licenses: Apache-2.0 for Vista (Gao et al., 2024b) and BERT, MIT for CLIP, RIFE (Huang et al., 2022), and GRiT, and BSD-3-Clause for the torchvision ResNet-50 weights and RAFT.