GAIA-2: A Controllable Multi-View Generative World Model for Autonomous Driving

Lloyd Russell*   **Anthony Hu***   **Lorenzo Bertoni***   **George Fedoseev***

Jamie Shotton   **Elahe Arani**   **Gianluca Corrado***

Wayve
`research@wayve.ai`

Abstract

Generative models offer a scalable and flexible paradigm for simulating complex environments, yet current approaches fall short in addressing the domain-specific requirements of autonomous driving—such as multi-agent interactions, fine-grained control, and multi-camera consistency. We introduce GAIA-2, *Generative AI for Autonomy*, a latent diffusion world model that unifies these capabilities within a single generative framework. GAIA-2 supports controllable video generation conditioned on a rich set of structured inputs: ego-vehicle dynamics, agent configurations, environmental factors, and road semantics. It generates high-resolution, spatiotemporally consistent multi-camera videos across geographically diverse driving environments (UK, US, Germany). The model integrates both structured conditioning and external latent embeddings (e.g., from a proprietary driving model) to facilitate flexible and semantically grounded scene synthesis. Through this integration, GAIA-2 enables scalable simulation of both common and rare driving scenarios, advancing the use of generative world models as a core tool in the development of autonomous systems. Videos are available at https://wayve.ai/thinking/gaia-2.

1 Introduction

Realistic simulation of driving scenarios is a foundational requirement for the development, training, and evaluation of autonomous driving systems. Generative world models enable scalable and diverse synthetic data creation, reducing reliance on expensive real-world data collection and facilitating robust evaluation in safe and repeatable environments. Unlike general-purpose text-to-video or image-to-video models, which primarily focus on visual realism and temporal coherence, autonomous driving applications demand fine-grained control over domain-specific aspects of the scene.

Generative models for autonomous driving must accurately simulate factors such as the actions of the ego-vehicle, the locations and movements of other agents (e.g., vehicles, pedestrians, cyclists), and their interactions. Moreover, these models must allow conditional generation based on contextual attributes, such as geographic location, weather, time of day, road configuration (e.g., speed limits, number of lanes, pedestrian crossings, traffic lights, intersections), and rare but critical edge-case scenarios. The ability to generate consistent multi-camera video streams is also essential, as autonomous vehicles rely on spatially and temporally coherent input from multiple perspectives for perception and decision-making. Meeting these domain-specific constraints presents unique challenges that are not adequately addressed by existing video generative models.

---

*Equal contribution, random order.

BACK LEFT   FRONT LEFT   FRONT FORWARD   FRONT RIGHT   BACK RIGHT

Figure 1: A selection of '*from scratch*' generations demonstrating the diversity of synthetic scenarios possible with GAIA-2.

While recent advancements in latent video generation, particularly those leveraging continuous [1, 2] or quantized [2, 3, 4] latent spaces, have led to improvements in efficiency and visual quality, current approaches remain limited in scope. In the context of autonomous driving, several models have introduced task-specific conditioning mechanisms [5, 6, 7, 8, 9], enabling some control over scenario elements. However, existing models typically support only a subset of the required capabilities, such as limited conditioning types, lack of multi-camera generation, or poor spatial-temporal coherence. As such, a comprehensive and unified generative framework remains an open challenge in the field.

To address this gap, we introduce *GAIA-2*, a domain-specialized latent diffusion model that represents a significant advancement in video-generative world modelling for autonomous driving. GAIA-2 advances prior work by supporting a conditional generation of high-resolution, multi-camera driving scenes, with fine-grained control over ego-vehicle actions, agent behavior, scene geometry, and environmental factors. The model can generate up to five temporally and spatially consistent camera streams at a resolution of 448×960, accommodating various multi-camera rig configurations used in real-world autonomous vehicle systems.

GAIA-2 supports conditioning on a wide range of scene attributes, including ego-vehicle kinematics (e.g., speed, curvature), geographic regions (UK, US, Germany), time of day, weather, and a rich taxonomy of road layout features—such as number and type of lanes (e.g., drivable, bus, cycle), presence of pedestrian crossings, traffic lights, and intersection topology (e.g., one-way roads, roundabouts). It also allows direct control over the locations, orientations, and dimensions of dynamic agents within the scene. This extensive set of conditioning parameters allows for precise control over the generated scenarios, enabling simulation of both typical driving conditions and edge cases critical for robust system evaluation.
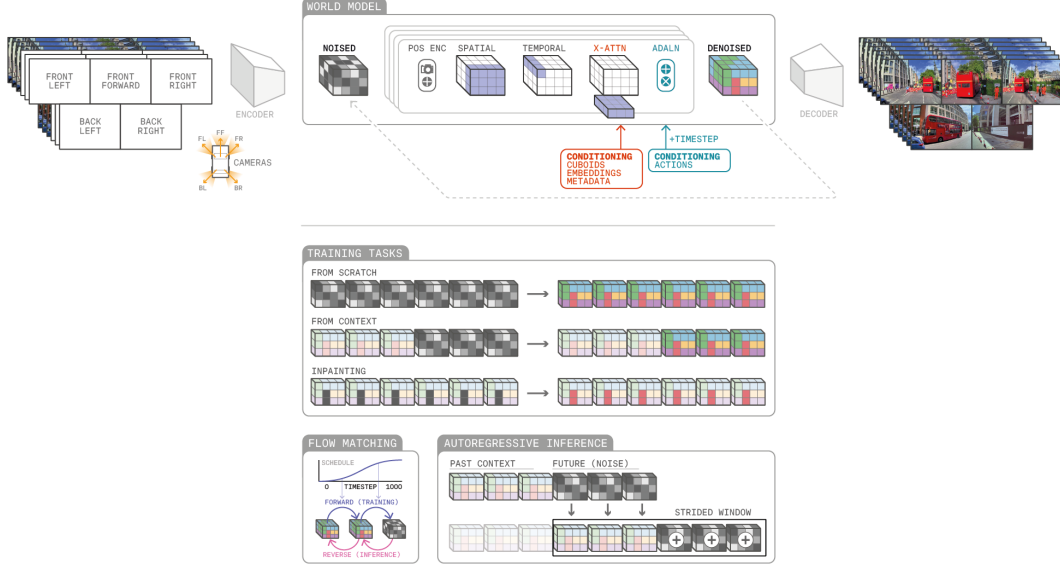
Figure 2: GAIA-2 world model. The architecture schematic shows full surround camera views independently encoded by a video tokenizer. The combined multi-camera latents are used as past context and linearly interpolated with sampled noise. We add camera parameters along with positional encodings before passing the latents through a space-time factorized transformer. Conditioning is implemented via cross-attention and adaptive layer norm. The latents are denoised into future latents, conditioned on various inputs, including actions, 3D bounding boxes, metadata, and scenario embeddings. The denoised latents are then decoded back to pixel space by the video tokenizer. Below, we depict the various tasks GAIA-2 is trained on, including from scratch generations, context rollouts, and spatial inpainting. At inference, we can generate beyond the window duration GAIA-2 was trained on by taking overlapping strides to generate new frames conditioned on previously generated frames.

To ensure flexibility and interoperability, GAIA-2 enables conditioning on external latent spaces, including CLIP embeddings and a proprietary model trained to produce driving-specific latent representations. This capability allows semantic control over scene content and facilitates integration with downstream planning or perception modules. GAIA-2 further supports multiple video generation modes, including generation from scratch, prediction from past context, and selective content editing through inpainting.

By integrating these capabilities into a single framework, GAIA-2 sets a new benchmark in video-generative world models for autonomous driving. It enables detailed, controllable, and realistic simulation across diverse conditions, providing a powerful tool for scalable training and robust evaluation of autonomous driving systems.

2 Model

GAIA-2 is a surround-view video generative world model with structured conditioning, multi-camera coherence, and high spatial-temporal resolution. The architecture (Figure 2) is composed of two primary components: a video tokenizer and a latent world model. Together, these modules enable GAIA-2 to generate realistic and semantically coherent video across multiple viewpoints with rich conditional control.

The video tokenizer compresses the raw high-resolution video into a compact, continuous latent space, preserving semantic and temporal structure. This compact representation enables efficient learning and generation at scale. The world model then learns to predict future latent states, conditioned on past latent states, actions, and a diverse set of domain-specific control signals. It can also be used to generate novel states completely from scratch and modify video content through inpainting. The predicted latent states are subsequently decoded back into pixel space using the video tokenizer decoder.
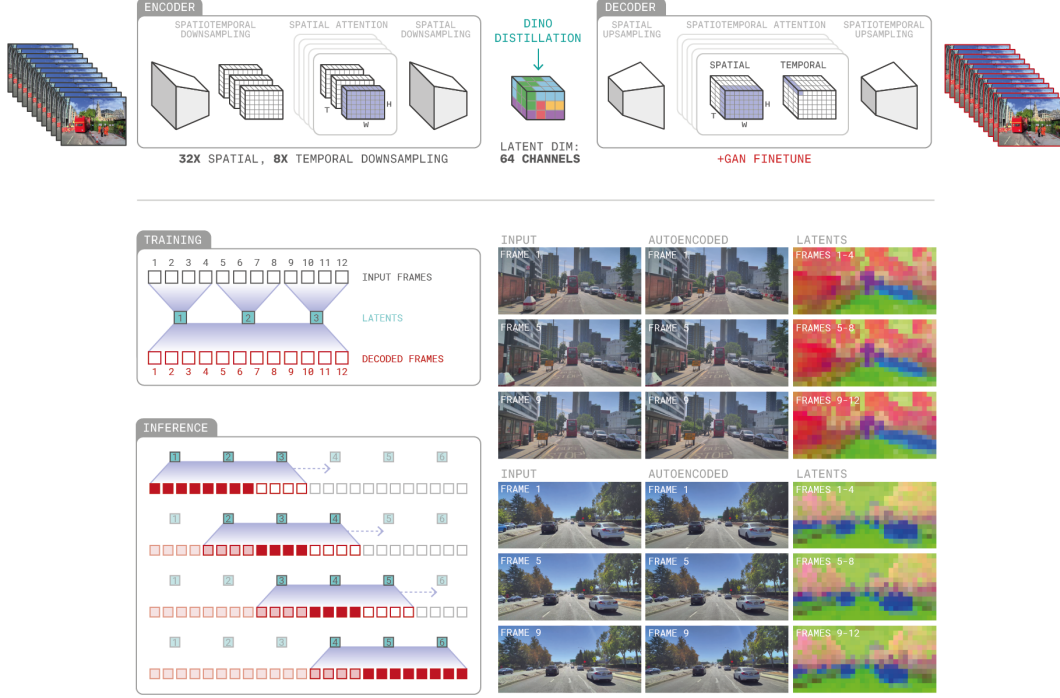
3

Figure 3: GAIA-2 video tokenizer. The architecture schematic depicts input frames spatiotemporally downsampled into temporally independent latents. A high spatial compression rate is compensated for by an increased latent channel dimension. Latents are decoded back to video frames, leveraging full spatiotemporal attention to ensure temporal consistency. To autoencode long sequences, we employ a rolling inference mechanism where current latents are decoded using past and future context in a sliding window fashion. The examples show input frames, decoded frames, and a visualization of the latent space. The first 3 principal components of the latent features are mapped to RGB values. Notably, the latent space is semantically stable across frames and across samples.

In contrast to most latent diffusion models, GAIA-2 employs a much higher spatial compression rate (e.g., $32\times$ vs. the more typical $8\times$), which is offset by increasing the channel dimension of the latent space (e.g., 64 channels instead of 16). This yields fewer but semantically richer latent tokens. The resulting advantages are twofold: (1) shorter latent sequences enable faster inference and better memory efficiency, and (2) the model demonstrates improved ability to capture video content and temporal dynamics. This parameterization strategy is inspired by prior work on compact latent representations [10, 11].

Unlike its predecessor GAIA-1 [5], which relied on discrete latent variables, GAIA-2 employs a continuous latent space, improving temporal smoothness and reconstruction fidelity. Furthermore, GAIA-2 introduces a flexible conditioning interface that supports ego-vehicle actions, dynamic agent states (e.g., 3D bounding boxes), structured metadata, CLIP and scenario embeddings, and camera geometry. This design enables robust control over scene semantics and agent behavior during generation, while ensuring cross-view consistency and temporal coherence.

2.1 Video Tokenizer

The video tokenizer compresses pixel-space video into a compact latent space that is continuous and semantically structured. It is composed of a space-time factorized transformer with an asymmetric encoder-decoder architecture (with 85M and 200M parameters, respectively). The encoder extracts spatiotemporally downsampled latents that are temporally independent; the decoder reconstructs full-frame video from these latents using temporal context for temporal consistency.

2.1.1 Encoder

Given an input video $(\mathbf{i}_1, ..., \mathbf{i}_{T_v})$, the encoder e_ϕ computes latent tokens $(\mathbf{z}_1, ..., \mathbf{z}_{T_L}) = e_\phi(\mathbf{i}_1, ..., \mathbf{i}_{T_v})$, where T_v corresponds to the number of video frames and T_L the number of latents. Let us denote by $H_v \times W_v$ the spatial resolution of the video frames and $H \times W$ the spatial resolution of the latent tokens. The encoder downsamples spatially by a factor $\frac{H_v}{H} = 32$ and temporally by a factor $\frac{T_v}{T_L} = 8$. The latent dimension is $L = 64$, which results in a total compression of $(\frac{T_v \times H_v \times W_v \times 3}{T_L \times H \times W \times L} = 384)$ with $(T_v, H_v, W_v) = (24, 448, 960)$ and $(T_L, H, W) = (3, 14, 30)$.

This downsampling is achieved by temporally striding the input frames by $2\times$ and the following modules:

1. A downsampling convolutional block with stride $2 \times 8 \times 8$ (time, height, width) followed by another downsampling convolutional block with stride $2 \times 2 \times 2$ (both operating at embedding dimension 512).

2. A series of 24 spatial transformer blocks with dimension 512 and 16 heads.

3. A final convolution with stride $1 \times 2 \times 2$, followed by a linear projection to $2L$ channels to model a Gaussian distribution over latents. Note that the latent dimension is doubled as the encoder predicts the mean and standard deviation of a Gaussian distribution. During both training and inference, the latents are sampled from this resulting distribution.

2.1.2 Decoder

The decoder architecture is:

1. A linear projection from the latent dimension to the embedding dimension followed by a first upsampling convolutional block with stride $1 \times 2 \times 2$ (upsampling is achieved with a depth-to-space module [12]).

2. A series of 16 space-time factorized transformer blocks with dimension 512 and 16 heads.

3. An upsampling convolutional block with stride $2 \times 2 \times 2$ followed by 8 space-time factorized transformer blocks with dimension 512 and 16 heads.

4. A final upsampling convolutional block with stride $2 \times 8 \times 8$ and dimension 3 corresponding to the pixel RGB channels.

A key difference between the encoder and decoder is that the encoder independently maps 8 consecutive video frames to a single temporal latent, while the decoder jointly decodes the $T_L = 3$ temporal latents to $T_v = 24$ video frames for temporal consistency. During inference, the video frames are decoded with a sliding window. The logic is shown in diagrams '*Training*' and '*Inference*' of Figure 3.

2.1.3 Training Losses

The video tokenizer is trained with a combination of pixel reconstruction and latent space losses:

- Image reconstruction with L_1, L_2, and perceptual losses [13].
- DINO [14] distillation on the latent features through a cosine similarity loss, encouraging semantic alignment with pre-trained representations. The video frames are encoded with the DINO model and temporally downsampled with linear interpolation to match the dimensionality of the latent features.
- Kullback-Leibler divergence loss [1] with respect to a standard Gaussian distribution to regularize the latent space.

To improve visual quality, we further fine-tune the decoder with a GAN loss [4] using a 3D convolutional discriminator and the image reconstruction losses, while keeping the encoder frozen. The discriminator is a series of residual 3D convolutional blocks with base channel 64, stride 2 in time and space with 3D blur pooling [15], channel multipliers $[2, 4, 8, 8]$, 3D instance normalization, and LeakyReLU with slope 0.2. It uses spectral normalization [16] and the original GAN loss implemented with a softplus activation function.

2.2 World Model

The latent world model predicts future latent states conditioned on past latents, actions, and a rich set of conditioning inputs. It is implemented as a space-time factorized transformer with 8.4B parameters and is trained using flow matching [17] for stability and sample efficiency.

Let $\mathbf{x}_{1:T} \in \mathbb{R}^{T \times N \times H \times W \times L}$ be the input latents, where T is the temporal window and N is the number of cameras. The input latents are obtained by independently encoding each camera view with the encoder e_ϕ. At each timestep t, we also provide an action vector $\mathbf{a}_t$ and a conditioning vector $\mathbf{c}_t$.

2.2.1 Architecture

The world model is a space-time factorized transformer with hidden dimension C. Each action $\mathbf{a}_t$ is embedded to $\mathbb{R}^C$, and the conditioning vector $\mathbf{c}_t$ embedded to $\mathbb{R}^{K \times C}$, where K corresponds to the number of conditioning variables. The flow matching time $\tau \in [0, 1]$ is also mapped to $\mathbb{R}^C$ using a sinusoidal encoding [18].

Flow matching time τ and action $\mathbf{a}_t$ are injected into each transformer block through the adaptive layer norm [19], while we use cross-attention for the other conditioning variables $\mathbf{c}_t$. We have found that action conditioning was more accurate when using adaptive layer norm over cross-attention. As the action affects every spatial token, the adaptive layer norm provides an explicit information gateway instead of having to rely on a learnable attention mechanism.

Regarding positional encoding, we independently encode: (i) spatial token position with a sinusoidal embedding, (ii) camera timestamp with a sinusoidal embedding followed by a small MLP, and (iii) camera geometry (distortion, intrinsic, and extrinsics) with learnable linear layers. All these positional encodings are added to the input latents at the beginning of each transformer block, similar to [20].

The world model contains 22 space-time factorized transformer blocks with hidden dimension $C = 4096$ and 32 heads. Each transformer block contains a spatial attention (over space and cameras), a temporal attention, a cross-attention, and an MLP layer with an adaptive layer norm. For increased training stability, we use query-key normalization [21] before each attention layer.

2.2.2 Losses

At training time, we randomly sample the number of context frames $t \in \{0, ..., T - 1\}$ where $t = 0$ corresponds to from scratch generation. We also sample a flow matching time $\tau \in [0, 1]$ according to a pre-defined distribution (see Section 2.2.4 for more details). The context latents $\mathbf{x}_{1:t}$ are unchanged, while the future latents $\mathbf{x}_{t+1:T}$ are linearly interpolated with random Gaussian noise $\boldsymbol{\epsilon}_{t+1:T} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

$$\mathbf{x}_{t+1:T}^\tau = \tau \mathbf{x}_{t+1:T} + (1 - \tau)\boldsymbol{\epsilon}_{t+1:T} \tag{1}$$

The velocity target vector $\mathbf{v}_{t+1:T}$ is the difference between target latents and random noise:

$$\mathbf{v}_{t+1:T} = \mathbf{x}_{t+1:T} - \boldsymbol{\epsilon}_{t+1:T} \tag{2}$$

The world model f_θ predicts the target velocity $\mathbf{v}_{t+1:T}$ conditioned on context latents $\mathbf{x}_{1:t}$, actions $\mathbf{a}_{1:T}$ and conditioning variables $\mathbf{c}_{1:T}$.

$$\hat{\mathbf{v}}_{t+1:T} = f_\theta(\mathbf{x}_{t+1:T}^\tau | \mathbf{x}_{1:t}, \mathbf{a}_{1:T}, \mathbf{c}_{1:T}, \tau) \tag{3}$$

The model is trained with an L_2 loss between the predicted and the target velocity.

$$\mathcal{L}_{\text{world-model}} = \mathbb{E}_{t \sim P(t), \tau \sim p(\tau)}[||\mathbf{v}_{t+1:T} - \hat{\mathbf{v}}_{t+1:T}||^2] \tag{4}$$

where $t \sim P(t)$ denotes the distribution of context frames and $\tau \sim p(\tau)$ the distribution of flow matching time.

2.2.3 Conditioning

GAIA-2 supports rich and structured conditioning inputs that enable fine-grained control over the generated scenes. These inputs include ego-vehicle actions, dynamic agent properties, scene-level metadata, camera configurations, timestamp embedding, and external latent representations such as CLIP or proprietary scenario embeddings. The conditioning mechanisms are integrated into the world model through a combination of adaptive layer normalization (for action), additive modules (for camera geometry and timestamp), and cross-attention (for all other variables).

Camera Parameters. We compute separate embeddings for intrinsics, extrinsics, and distortion, which are then summed to form a unified camera encoding. For intrinsics, we extract focal lengths and principal point coordinates from the intrinsic matrix, normalize them, and project them into a shared latent space. Extrinsics and distortion coefficients are similarly processed through their respective encoders to yield compact representations. This configuration allows the model to effectively incorporate real-world camera variations. Figure 5 illustrates the top three most common camera rig configurations in the training dataset.

Video Frequency. To account for variable video frame rates, GAIA-2 uses timestamp conditioning. Each timestamp is: (*i*) Normalized relative to the present time and scaled to the range $[-1, 1]$, (*ii*) Transformed using sinusoidal functions (Fourier feature encoding), and (*iii*) Passed through an MLP to produce a vector in the shared latent space. This encoding captures both low- and high-frequency temporal variation and enables the model to reason effectively over videos recorded at different rates.

Action. The ego-vehicle action is parameterized by speed and curvature. Since these quantities span multiple orders of magnitude, we use a symmetric logarithmic transformation *symlog* [22] for normalization:

$$\text{symlog}(y) = \text{sign}(y) \frac{\log(1 + s|y|)}{\log(1 + s|y_{\max}|)} \tag{5}$$

Here, y represents speed or curvature, and s is a scale factor: For curvature (units: m^{-1}, range: 0.0001 to 0.1), we use $s = 1000$ to amplify low values. For speed (units: m/s, range: 0 to 75), we use $s = 3.6$ to express values in km/h. The result is a compact representation scaled to $[-1, 1]$, improving training stability.

Dynamic Agents. To represent surrounding agents, we use 3D bounding boxes predicted by a 3D object detector [23] re-trained on our dataset. Each box encodes the 3D location, orientation, dimensions, and category of an agent. The 3D boxes are projected into the 2D image plane and normalized, yielding $f_i \in \mathbb{R}^{T \times N \times B \times 13}$ conditioning features where T denotes the number of temporal latents, N the number of cameras, and B the maximum number of 3D bounding boxes (zero-padding as needed). Each feature dimension is embedded independently and aggregated via a single-layer MLP.

To enhance model robustness and generalizability, we implement dropout at both the feature dimension and instance levels during training. Specifically, feature dimensions are dropped out with a probability of $p = 0.3$, allowing the model to operate under incomplete information at inference. This setup allows, for example, conditioning on 2D projected boxes without specifying the 3D locations of instances, or omitting orientations to let the model predict the most plausible orientation based on other conditions.

At the instance level, for each camera, we sample a frame $t \in \{1, ..., T\}$ and calculate the number of detected instances $N_{\text{instances}}$. We then sample the number of instances to condition on $n \in \{0, ..., \min(B, N_{\text{instances}})\}$, and apply dropout to the instances exceeding this sample size. This allows the model to adapt to a variable number of dynamic agents during inference. Note that while keeping n constant across time, we do not use instance tracking, allowing the model to independently determine whether the conditioning features across frames belong to the same or different instances.

Metadata. Metadata features are categorical and embedded using dedicated learnable embedding layers. These include: Country, weather, time of day; Speed limits; Number and types of lanes (e.g., bus, cycle); Pedestrian crossings, traffic lights and their states; One-way road indicators and intersection types. These embeddings allow GAIA-2 to learn nuanced relationships between scene-level features and their effects on behavior, enabling simulation of both typical and rare scenarios.

CLIP Embedding. To enable semantic scene conditioning, GAIA-2 supports conditioning on CLIP embeddings [24]. During training, we extract CLIP features using the image encoder on video frames. At inference, these can be replaced with CLIP text encoder outputs from natural language prompts. All CLIP embeddings are projected into the model's latent space using a learnable linear projection. This enables zero-shot control over scene semantics through natural language or visual similarity.

Scenario Embedding. GAIA-2 can also be conditioned on scenario embeddings obtained from an internal proprietary model trained to encode driving-specific information. These embeddings compactly capture ego-action and scene context, such as road layout and agent configurations. The scenario vectors are projected via a learnable linear layer into the latent space before integration into the transformer. This allows high-level scenario generation from a compact abstract representation.

2.2.4 Flow matching Time Distribution

A critical factor for training the world model under the flow matching framework is the distribution of the flow matching time τ. This distribution determines how frequently the model sees latent inputs that are close to real latents versus heavily perturbed.

We use a bimodal logit-normal distribution with two modes:

- A primary mode centered at $\mu = 0.5$, $\sigma = 1.4$, sampled with probability $p = 0.8$. This biases the model towards learning with low-to-moderate noise levels. Empirically, this encourages learning useful gradients, as even small amounts of noise can significantly perturb high-capacity latents.
- A secondary mode centered at $\mu = -3.0$, $\sigma = 1.0$, sampled with probability $p = 0.2$. This concentrates training on nearly pure noise regions around $\tau = 0$, helping the model learn spatial structures and low-level dynamics, such as ego-motion or object trajectories.

This bimodal strategy ensures that training is effective across both low and high-noise regimes, improving generalization and sample quality.

In addition, the input latents $\mathbf{x}_t$ are normalized by their mean μ_x and standard deviation σ_x, following [25], to ensure their magnitude matches that of the added Gaussian noise. This avoids scale mismatch between the signal and the perturbation, which can otherwise degrade training dynamics.

3 Data

GAIA-2 is trained on a large-scale internal dataset specifically curated to support the diverse demands of video generation for autonomous driving. The dataset comprises approximately 25 million video sequences, each spanning 2 seconds, collected between 2019 and 2024. Recordings were obtained across three countries—the United Kingdom, the United States, and Germany—to ensure coverage of geographically and environmentally diverse driving conditions.

To capture the complexity of real-world autonomous driving, data collection involved multiple vehicle platforms, including three different car models and two van types. Each vehicle was outfitted with either five or six cameras, configured to provide comprehensive 360-degree surround-view coverage. The camera systems varied in their capture frequencies—20 Hz, 25 Hz, and 30 Hz—introducing a range of temporal resolutions. This variability reflects the heterogeneous nature of sensor setups deployed in actual autonomous vehicles and supports GAIA-2's ability to generalize across different input rates and hardware specifications.

An important characteristic of the dataset is the variability in camera placement throughout the data collection period. Over time, the positions and calibrations of the cameras were adjusted across platforms, introducing a broad spectrum of spatial configurations. This diversity provides a strong training signal for generalizing across different camera rigs, a key requirement for scalable synthetic data generation in the autonomous driving domain.

The dataset also encompasses a wide range of driving scenarios, including diverse weather conditions, times of day, road types, and traffic environments. To ensure coverage across this complexity, we explicitly balance the training data not just on individual features, but on their joint probability distribution. This approach enables a more representative learning signal by modeling realistic co-occurrences—e.g., specific lighting and weather conditions in certain geographies or behaviors unique to particular road types. To prevent redundant training samples, we enforce a minimum temporal stride between selected sequences, reducing the risk of duplication while maintaining a natural distribution.

For evaluation, we implement a geographically held-out validation strategy. Specific validation geofences are defined to exclude certain regions from the training set entirely. This ensures that model

evaluation is performed on unseen locations, providing a more rigorous assessment of generalization performance across different environments.

In summary, this dataset provides a robust foundation for training GAIA-2. Its extensive temporal and spatial coverage, diversity in vehicle and camera configurations, and principled validation setup make it uniquely well-suited for developing generative world models capable of producing realistic and controllable driving video across varied real-world conditions.

4 Training Procedure

This section describes the training procedures for both components of GAIA-2: the video tokenizer and the world model. Each component is trained independently using large-scale compute infrastructure and tailored loss configurations to optimize their respective objectives.

Video Tokenizer. The video tokenizer was trained for $300{,}000$ steps with a batch size of 128 using 128 H100 GPUs. Input sequences consisted of $T_v = 24$ video frames sampled at their native capture frequencies (20, 25, or 30 Hz). Random spatial crops of size 448×960 were extracted from the frames. For each training sample, a camera view was randomly selected from the available $N = 5$ perspectives. Notably, each camera stream was encoded independently.

The tokenizer performs $8\times$ temporal and $32\times$ spatial downsampling, yielding a compressed representation with latent dimension $L = 64$, and an effective total compression rate of approximately 400 ($\frac{24 \times 448 \times 960 \times 3}{3 \times 14 \times 30 \times 64} \simeq 400$).

The tokenizer's loss function is composed of a combination of image reconstruction, perceptual, and semantic alignment terms: (1) DINO v2 (Large) [14] distillation in the latent space with a weight of 0.1. (2) KL divergence [1] between the latent distribution and a unit Gaussian, with a low weight of $1e{-}6$ to encourage smoothness. (3) Pixel-level losses: L_1 loss (weight 0.2), L_2 loss (weight 2.0), and LPIPS perceptual loss [26] (weight 0.1).

An exponential moving average (EMA) of the tokenizer parameters ϕ was maintained throughout training, with a decay factor of 0.9999 and updates at every training step. The EMA weights were used at inference.

Training was optimized using AdamW with: $2{,}500$ warm-up steps to the base learning rate of $1e{-}4$, and $5{,}000$ cooldown steps to a final learning rate of $1e{-}5$; Adam betas $[0.9, 0.95]$, weight decay 0.1, and gradient clipping at 1.0. After initial training, the tokenizer decoder was fine-tuned for an additional $20{,}000$ steps using a GAN loss (weight 0.1) in combination with the previous reconstruction losses. The discriminator was optimized with a learning rate of $1e{-}5$.

World Model. The latent world model was trained for $460{,}000$ steps with a batch size of 256 on 256 H100 GPUs. Inputs consisted of 48 video frames at native capture frequencies (20, 25, or 30 Hz) and at spatial resolution 448×960.

To encourage generalization, we sampled different training tasks as follows: 70% from-scratch generation, 20% contextual prediction, and 10% spatial inpainting. To regularize the model and enable classifier-free guidance, conditioning variables were randomly dropped. Each individual conditioning variable was independently dropped with 80% probability, and all of them were simultaneously dropped with 10% probability.

Input camera views were randomly dropped with 10% probability to enhance robustness to partial observability. Latent tokens were normalized using a fixed mean of $\mu_x = 0.0$ and standard deviation $\sigma_x = 0.32$, as empirically determined during tokenizer training.

As with the tokenizer, we maintained an EMA of the world model parameters θ with a decay factor of 0.9999 and updates at every training step. EMA weights were used for inference. The optimizer was AdamW with: $2{,}500$ warm-up steps to an initial learning rate of $5e{-}5$ and cosine decay over the full training duration to a final learning rate of $6.5e{-}6$; Adam betas $[0.9, 0.99]$, weight decay 0.1, and gradient clipping 1.0.

5 Inference

The GAIA-2 model supports a range of inference tasks that showcase its flexibility and controllability across video generation scenarios. These tasks are unified by a shared denoising process operating in the latent space, followed by decoding into pixel space via the video tokenizer.

Inference Tasks. We consider four primary inference modes, each illustrating a distinct capability of the model:

- **Generation from scratch** involves sampling pure Gaussian noise and denoising them with guidance from conditioning variables defined in Section 2.2.3. The resulting latents are then decoded to video frames using the video tokenizer decoder, with temporally consistent outputs produced via the rolling window decoding mechanism illustrated in Figure 3.

- **Autoregressive prediction** enables forecasting future latents from a sequence of past context latents. Given an initial context window of $k = 3$ temporal latents, the model predicts the next set of latents, appends it to the context, and repeats the process using a sliding window. This approach allows for long-horizon rollouts while incorporating conditioning signals such as ego motion. An example is provided in Figure 9.

- **Inpainting** allows selective modification of video content. A spatial-temporal mask is applied to the latent input, and the masked regions are regenerated via conditional denoising. Optional guidance from dynamic agent conditioning (e.g., agent locations) can steer the generation within the masked area. An example is shown in Figure 12.

- **Scene editing** is achieved by partially noising latents extracted from real video, followed by denoising with altered conditioning. This enables targeted semantic or stylistic transformations—such as changing the weather, time of day, or road layout—without re-generating the full scene. Figure 6 illustrates this capability.

These modes demonstrate that GAIA-2 can serve as a general-purpose simulator for a wide variety of scene manipulation tasks, whether starting from noise, context, or existing video.

Inference Noise Schedule. For all inference tasks, we adopt the linear-quadratic noise schedule introduced by [20]. This schedule begins with linearly spaced noise levels, which are effective for capturing coarse scene layouts and motion patterns. In later stages, the schedule transitions to quadratically spaced steps that allow more efficient refinement of high-frequency visual details. This hybrid approach improves both generation quality and computational efficiency. In our experiments, we use a fixed number of 50 denoising steps.

Classifier-Free Guidance. Classifier-free guidance (CFG) is not used by default during inference. However, for challenging or out-of-distribution scenarios, such as those involving rare edge cases or unusual agent configurations (e.g., Figure 11), we activate CFG with a guidance scale ranging from 2 to 20, depending on the complexity of the scene.

In scenarios involving dynamic agent conditioning, where the latent tokens associated with agent-specific regions are known a priori, we apply spatially selective CFG. In this case, guidance is applied only to the spatial locations influenced by the conditioning (e.g., 3D bounding boxes), which enhances generation quality in targeted regions without unnecessarily affecting the rest of the scene. This targeted approach enables more precise control over scene elements while preserving global coherence.

6 Results

We evaluate GAIA-2 through qualitative examples and quantitative metrics that highlight its generative fidelity, controllability, and suitability for synthetic data generation in autonomous driving. A comprehensive set of generated video examples is available at https://wayve.ai/thinking/gaia-2.
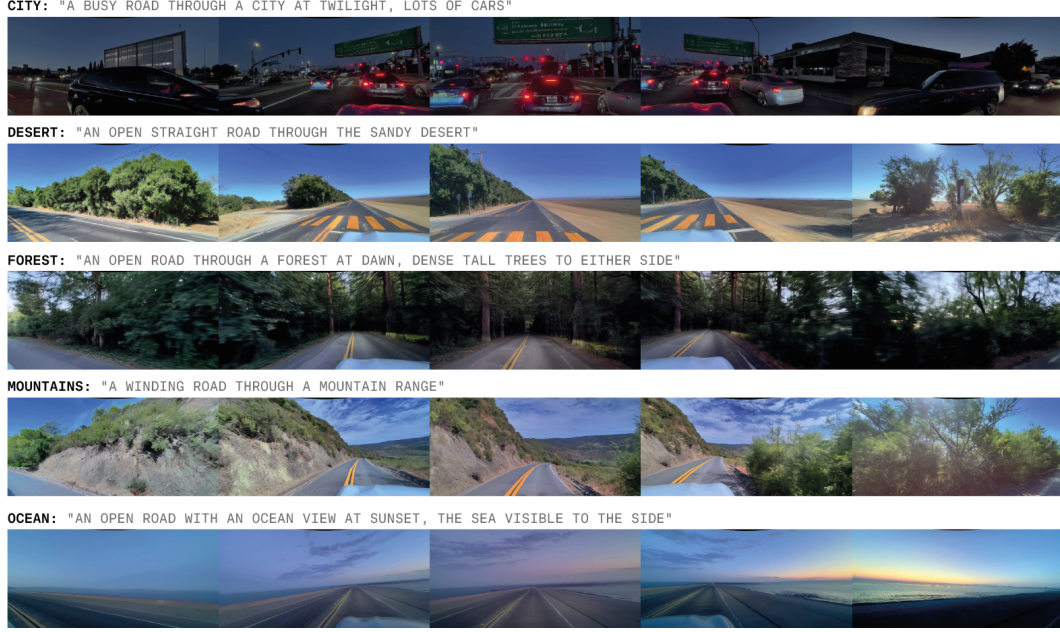
Figure 4: CLIP conditioning. Conditioning GAIA-2 generations with text prompts encoded through the CLIP text encoder provides nuanced control over environmental features.

6.1 Qualitative Examples

Diverse Scenario Generation. GAIA-2 supports the generation of highly diverse driving scenarios, spanning multiple countries, weather conditions, times of day, and road layouts. In addition to structured conditioning via metadata, GAIA-2 can be guided by CLIP embeddings, allowing control over scene semantics not explicitly captured in labels. This includes geographic and environmental context such as urban, mountainous, or coastal scenes (Figure 4).

Furthermore, due to its exposure to varied camera configurations during training, GAIA-2 can simulate driving videos across different vehicle embodiments. By conditioning on camera parameters, it maintains spatial and temporal consistency across viewpoints for rigs mounted on sports cars, SUVs, and large vans (Figure 5).

Augmenting Real-World Data. GAIA-2 enables dataset augmentation through a range of techniques that allow for visual and contextual diversification of real-world sequences.

- **Partial noise and denoise:** Latents derived from real videos are partially noised and then denoised under altered conditioning, such as different weather or lighting. This approach preserves semantics and ego motion while yielding diverse visual outputs (Figure 6). We demonstrate the ability to modify environmental aspects of the scene while leaving core semantic and functional components as per the original. Essentially turning a single real-world example into multiple new scenarios: the same ego trajectory but visually diversified to take place during sunshine, rain, at sunset, at night, and in the snow.

- **Scenario embeddings:** Scenario embeddings derived from our proprietary driving models give GAIA-2 the ability to generate multiple plausible variations from a single real-world example, providing many new examples of diverse yet contextually coherent synthetic data (Figure 7). Because the embedding space is meaningfully organized by the driving model, we are able to reliably generate scenarios that are semantically interpretable given their location in the embedding space, such as accelerating, decelerating, or overtaking other agents. By additionally conditioning on environmental factors, we can synthetically expand our dataset by increasing coverage across various conditions. By conditioning on different camera parameters, we can effectively replicate our corpus across any given vehicle platform.

11

Figure 5: Multi-rig video generation. GAIA-2 supports various vehicle platforms and camera setups, maintaining spatial and temporal consistency. The two examples shown include camera arrangements from a sports car, an SUV, and a large van.

- **Action-based generation:** GAIA-2 can synthesize new scenes purely from ego-vehicle action trajectories. Given the speed and curvature profiles, it generates plausible visual contexts aligned with those dynamics, such as traffic light changes, braking behavior, or turning maneuvers (Figure 8). We show three examples: (1) conditioning on a '*start from stopped*' speed profile, GAIA-2 generates plausible visual observations to fit those actions, in this case, a UK traffic light turning from red and amber to green; (2) conditioning on a '*slow to a stop*' speed trajectory, GAIA-2 generates a scenario where the ego agent is slowing down behind a London taxi; and (3) conditioned on a strong leftward curvature and slow ramping speed, GAIA-2 generates a U-turn at a US intersection.

Generating Safety-Critical Scenarios. GAIA-2 is able to generate rare but safety-critical scenarios that would otherwise occur infrequently in the real world. We consider two classes of safety-critical situations: those caused by the ego agent, and those caused by other agents.

- **Ego-vehicle induced scenarios:** By conditioning on extreme or unsafe ego-vehicle actions (e.g., abrupt steering into oncoming traffic), GAIA-2 generates realistic scenarios critical to testing autonomous system resilience under hazardous conditions (Figure 9 and Figure 10).
- **Other-agent induced scenarios:** Using 3D bounding boxes conditioning, GAIA-2 precisely controls the placement and motion of other agents, creating scenarios involving aggressive driving behaviors, emergency braking, or hazardous crossings, crucial for testing reactive capabilities of autonomous systems (Figure 11).

Inpainting. GAIA-2 supports spatial-temporal inpainting, enabling selective content editing. When provided with masked regions and corresponding agent-level conditioning, GAIA-2 inserts agents seamlessly into existing contexts without disrupting unrelated areas (Figure 12).

These qualitative results highlight GAIA-2's ability to perform fine-grained and controllable video generation, supporting its use in scalable simulation and diverse scenario creation for autonomous systems.

6.2 Metrics

To quantitatively evaluate GAIA-2, we use a suite of metrics that assess visual fidelity, temporal consistency, and conditioning accuracy.

Figure 6: Augmentation through partial noising. By partially noising and denoising video frames, GAIA-2 transforms real video into diverse versions under different environmental settings, such as weather and time of day, while preserving semantic content and ego-actions.

Visual Fidelity. Similarly to [27], we use the Frechet DINO Distance (FDD) and Frechet Inception Distance (FID) [28] to measure the distance between the distributions of generated and real videos. The feature encoder is DINOv2 [14] ViT-L/14, and the input images are bilinearly rescaled to a higher resolution 448×952 compared to 299×299 for InceptionV3 [29] in FID. We observe that FDD is less noisy than FID and saturates much later in training.

Temporal Consistency. To evaluate the temporal consistency of the world model, we use the Frechet Video Motion Distance (FVMD) [30]. This metric compares distributions of explicit key-point motion features for generated and ground-truth videos. We find this metric to be more aligned with human preferences for temporal consistency compared to FVD [31].

Dynamic Agent Conditioning. We evaluate dynamic agent fidelity by comparing the projection of the 3D bounding boxes (our conditional signal) with segmentation masks obtained via OneFormer [32] on the generated video. We then compute class-based Intersection-over-Union (IoU) to measure how well the model adheres to the spatial and categorical specifications of dynamic agents during generation.

We report the metrics and the validation loss in Figure 13. The validation loss is evaluated on the generation from scratch task with $n = 2048$ samples. The other metrics are evaluated on $n = 250$ samples due to inference time constraints, as they are computed on top of generated videos that require going through the entire denoising process. Other metrics are computed on $n = 250$ samples due to inference time constraints. Despite some noise from small sample sizes, all metrics exhibit a downward trend throughout training. Notably, the validation loss correlates most strongly with human preference, making it a valuable proxy for perceptual quality.
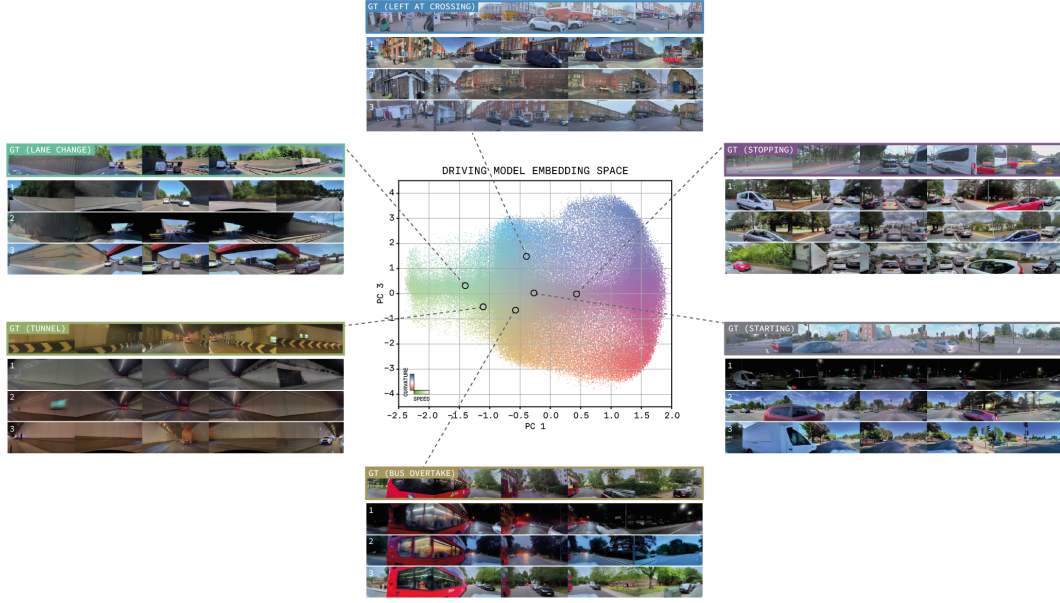
Figure 7: Scenario embedding conditioning. Scenario embeddings from a proprietary driving model enable GAIA-2 to generate diverse yet semantically consistent variations from real-world scenarios. For each scenario type the top row shows the real data the embedding is derived from, and the bottom three rows show synthetic variants of it, guided by the embedding as conditioning. We also optionally apply additional conditioning such as country, weather, time of day, or camera parameters to control diversity.

Figure 8: Action-based generation. GAIA-2 synthesizes diverse scenes from specified speed and curvature profiles. Each scenario is contextually plausible, despite the original video inputs being dropped out.

14

Figure 9: Ego-induced safety-critical scenarios. By applying unsafe action conditioning, GAIA-2 can generate hazardous situations, such as steering into oncoming traffic. We provide real video frames as context and GAIA-2 rolls out from these starting frames given some new action conditioning. In this example, we control the ego vehicle steering by conditioning on rightward curvature. We drop out speed so that GAIA-2 can generate multiple diverse outcomes. Note how the ego vehicle slows down while veering sharply into oncoming traffic. The oncoming vehicle swerves to avoid the ego vehicle.

Figure 10: Extreme generalization behavior. When conditioned on high speed and strong curvature, GAIA-2 extrapolates off-road trajectories, capturing rare behaviors like driving into fields or forests.
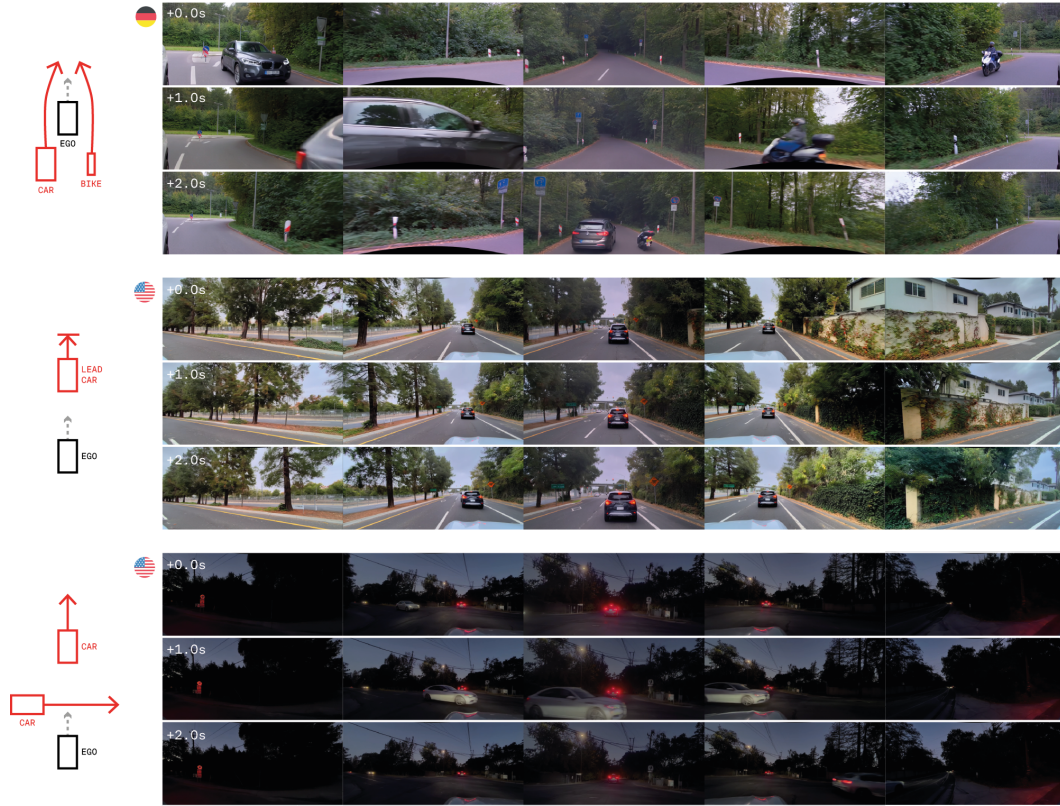
Figure 11: Other-agent induced safety-critical scenarios. 3D bounding box conditioning enables fine control over dynamic agents, allowing simulation of risky behaviors such as sudden braking, overtaking, or speeding through intersections.
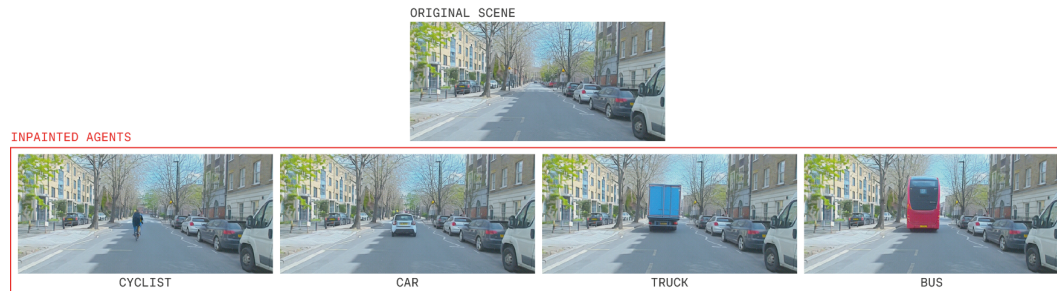

Figure 12: Agent inpainting. GAIA-2 can insert dynamic agents into masked regions of a video, guided by 3D bounding box conditioning. Background consistency and semantic realism are preserved.
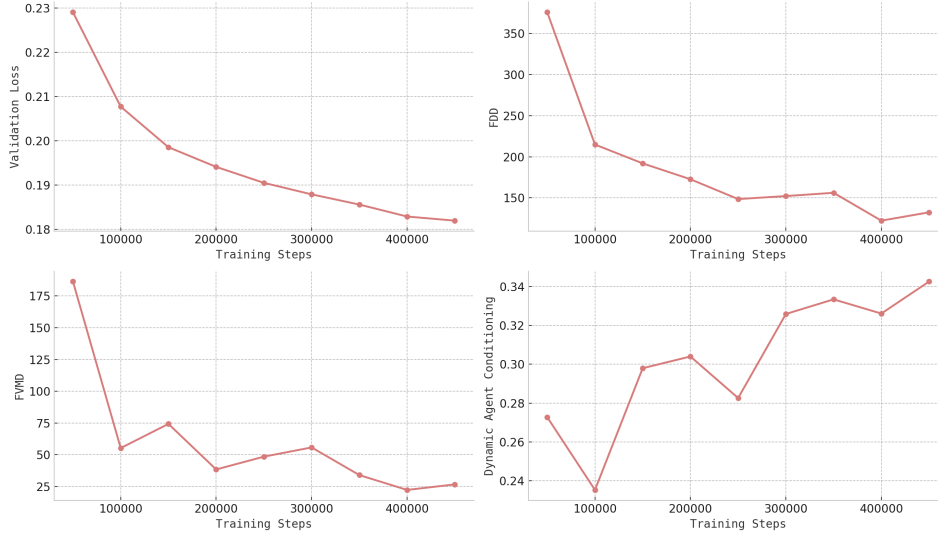
16

Figure 13: Validation loss (evaluated on $n =$2048 samples) and other metrics (evaluated on $n =$250 samples). We found that the validation loss correlates well with human perceptual preferences.

7 Related Work

Video Generative Models. Recent advancements in video generative models have mainly been driven by models that leverage latent representations derived through dimensionality reduction. These representations can be either continuous [1, 33] or quantized [3, 4], and the success of video generation heavily depends on how efficiently these representations capture spatiotemporal information. One of the key challenges lies in ensuring that the mappings to and from latent space preserve both visual fidelity and semantic coherence. While scaling to large datasets improves generalization [2], challenges remain in preserving both visual realism and temporal consistency.

Autoregressive and masked generative models using quantized tokens have demonstrated strong modeling of temporal dependencies and motion dynamics [34, 35, 36, 37, 38, 39, 40, 41, 42]. However, their sequential generation processes lead to slow inference and error accumulation, which can limit their scalability for long sequences or complex multi-agent scenes. In contrast, latent diffusion models offer a more parallelizable alternative, using iterative denoising to generate full video sequences. These models, including Stable Video Diffusion [43], DIAMOND [44], Sora [45], MovieGen [20], Gen-3 [46], and Cosmos [2], have achieved impressive quality on text-to-video and image-to-video tasks, especially in open-domain settings.

However, most of these works prioritize visual aesthetics and domain-agnostic generation, offering limited support for structured control over scene elements or agent behavior. This makes them less suitable for applications (such as autonomous driving) that require precise, semantic, and multi-modal scene control across space and time. In this context, GAIA-2 aims to bridge the gap between high-fidelity generation and domain-specific controllability required for safety-critical systems.

Generative World Models in Autonomous Driving. Unlike general-purpose video generation, generative world models for autonomous driving must satisfy stricter requirements: multi-agent interaction, ego-motion control, environmental diversity, and multi-camera coherence. Early work in this space includes GAIA-1 [5], which introduced text and ego action conditioning in a discrete world model with a video diffusion decoder to improve temporal consistency. CommaVQ [47] similarly leveraged a causal transformer on discrete tokens for ego-motion control. Although these methods took important steps toward controllability, they often targeted single-camera settings and partial controllability via action or text conditioning, limiting their applicability to comprehensive driving simulations.

Subsequent approaches adopted latent diffusion for higher fidelity generation and introduced more flexible control. DriveDreamer [6] adopts a latent diffusion model conditioned on 3D bounding boxes,

HD maps, and ego actions, and further incorporates an action decoder to predict future ego actions. Drive-WM [7] extends this approach to a multi-camera setting (up to six cameras surrounding the ego vehicle) and introduces controllability over ego actions, dynamic agents, and environmental conditions, such as weather and lighting. UniMLVG [48] similarly enables multi-view generation conditioned on text, camera parameters, 3D bounding boxes, and HD maps, while MaskGWM [49] extends UniMLVG to support longer video generations. Vista [8] employs latent diffusion for generating high-resolution, long-duration videos, and Delphi [50] uses a data-driven approach to guide video generation toward failure cases in driving models, thereby improving the synthetic data's training utility. More recently, GEM [51] further generalizes scene control to include ego motion, object dynamics, and even human poses, suggesting broader applicability across domains such as human motion synthesis and drone navigation. A recent line of work has explored 4D geometry-aware generation. DriveDreamer4D [9] uses video generative models to synthesize videos along novel trajectories and combines real and generated footage to optimize a 4D Gaussian Splatting model for closed-loop simulation. Similarly, DreamDrive [52] leverages generative priors to construct high-quality 4D scenes from in-the-wild driving data.

Despite these advancements, most of these solutions only address subsets of the requirements for realistic driving simulation. Some are limited to single-camera settings, others lack structured agent-level control, and few integrate multi-view generation and agent-level semantics into a single framework. Additionally, many do not support fine-grained editing tasks such as inpainting or targeted scene modification, essential for data augmentation and scenario variation in simulation loops.

Motivated by these limitations, **GAIA-2** is designed as a unified latent diffusion framework that brings together:

- multi-camera generation (up to five views);
- structured and semantic conditioning over ego action, dynamic agents, and scene-level metadata;
- flexible inference modes, including generation from scratch, inpainting, scene editing, and long-horizon rollouts; and
- support for external latent spaces like CLIP and driving scenario embeddings.

By combining continuous latent representations with large-scale training on a geographically and environmentally diverse dataset, GAIA-2 delivers both visual fidelity and scenario controllability. It fills a critical gap between general-purpose diffusion models and the domain-specific needs of autonomous driving, namely, the ability to simulate realistic, controllable, multi-view driving scenes for robust, scalable training and evaluation.

8 Conclusions and Future Work

We presented GAIA-2, a domain-specialized latent diffusion model for video generation, designed to meet the diverse and stringent requirements of simulation in autonomous driving. GAIA-2 sets a new benchmark in generative world modeling by unifying multi-camera coherence, structured semantic conditioning, and fine-grained control within a scalable architecture. The model supports controllable generation conditioned on ego-vehicle dynamics, environmental features, road layout, and dynamic agents, while maintaining spatial and temporal consistency across up to five camera views.

The architecture combines a space-time factorized latent diffusion world model with a compact and semantically meaningful video tokenizer, enabling high-fidelity generation across diverse driving environments. GAIA-2 supports multiple inference modes—including generation from scratch, autoregressive rollouts, spatial inpainting, and real-scene editing—each empowering systematic exploration of the driving scene space. The model excels at producing typical as well as rare safety-critical scenarios, thus significantly enhancing data diversity, scenario coverage, and validation rigor in autonomous driving systems.

By scaling controllability and scene diversity in simulation, GAIA-2 bridges the gap between real-world data limitations and the increasing demands of training and evaluating robust AI-driving models. In doing so, it serves as a powerful tool for accelerating iteration cycles, improving generalization to out-of-distribution conditions, and stress-testing AI agents across complex, long-tail scenarios.

Future Work. While GAIA-2 represents a significant step forward, several research directions remain open and will guide future iterations of this work: (*i*) Like all generative models, GAIA-2 occasionally produces temporal or semantic inconsistencies, particularly in long-horizon or complex scenarios. Improving the reliability and consistency of video generation through better failure detection, refinement models, or constraint-aware sampling is a key challenge. (*ii*) Although GAIA-2 enables parallelized generation, real-time or near-real-time video synthesis remains computationally intensive. Future work will explore model distillation, efficient transformer variants, and inference-time acceleration techniques to improve deployment feasibility in real-world pipelines. (*iii*) Despite supporting a rich set of conditionings, new efforts will target richer agent behavior modeling, more nuanced environmental conditions, and open-ended natural language control. Further enrichment of the training dataset with rare and safety-critical events—especially those involving human agents, complex infrastructure, or social interactions—will push the boundaries of realistic simulation.

Together, these directions will ensure that GAIA-2 and its successors continue to advance the role of generative models as core infrastructure in the development of safe, robust, and generalizable autonomous systems.

Acknowledgments

References

[1] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[2] N. Agarwal, A. Ali, M. Bala, Y. Balaji, E. Barker, T. Cai, P. Chattopadhyay, Y. Chen, Y. Cui, Y. Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[3] A. van den Oord, O. Vinyals, and K. Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[4] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[5] 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. *Technical Report arXiv:2309.17080*, 2023.

[6] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759, 2024.

[8] 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. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[9] G. Zhao, C. Ni, X. Wang, Z. Zhu, X. Zhang, Y. Wang, G. Huang, X. Chen, B. Wang, Y. Zhang, et al. Drivedreamer4d: World models are effective data machines for 4d driving scene representation. *arXiv preprint arXiv:2410.13571*, 2024.

[10] J. Chen, H. Cai, J. Chen, E. Xie, S. Yang, H. Tang, M. Li, Y. Lu, and S. Han. Deep compression autoencoder for efficient high-resolution diffusion models. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[11] Y. HaCohen, N. Chiprut, B. Brazowski, D. Shalem, D. Moshe, E. Richardson, E. Levin, G. Shiran, N. Zabari, O. Gordon, P. Panet, S. Weissbuch, V. Kulikov, Y. Bitterman, Z. Melumian, and O. Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint*, 2024.

[12] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[14] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

[15] R. Zhang. Making convolutional networks shift-invariant again. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

[16] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

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

[18] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[19] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023.

[20] A. Polyak, A. Zohar, A. Brown, A. Tjandra, A. Sinha, A. Lee, A. Vyas, B. Shi, C.-Y. Ma, C.-Y. Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint*, 2024.

[21] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin, R. Jenatton, L. Beyer, M. Tschannen, A. Arnab, X. Wang, C. Riquelme, M. Minderer, J. Puigcerver, U. Evci, M. Kumar, S. van Steenkiste, G. F. Elsayed, A. Mahendran, F. Yu, A. Oliver, F. Huot, J. Bastings, M. P. Collier, A. Gritsenko, V. Birodkar, C. Vasconcelos, Y. Tay, T. Mensink, A. Kolesnikov, F. Pavetić, D. Tran, T. Kipf, M. Lučić, X. Zhai, D. Keysers, J. Harmsen, and N. Houlsby. Scaling vision transformers to 22 billion parameters. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

[22] J. B. W. Webber. A bi-symmetric log transformation for wide-range data. *Measurement Science and Technology*, 2012.

[23] S. Wang, Y. Liu, T. Wang, Y. Li, and X. Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3621–3631, 2023.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[26] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. 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.

[27] G. Stein, J. C. Cresswell, R. Hosseinzadeh, Y. Sui, B. L. Ross, V. Villecroze, Z. Liu, A. L. Caterini, J. E. T. Taylor, and G. Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. 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.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[30] J. Liu, Y. Qu, Q. Yan, X. Zeng, L. Wang, and R. Liao. Fréchet video motion distance: A metric for evaluating motion consistency in videos. In *Proceedings of the International Conference on Machine Learning, workshop (ICMLw)*, 2024.

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

[32] J. Jain, J. Li, M. Chiu, A. Hassani, N. Orlov, and H. Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[33] NVIDIA. Cosmos tokenizer: A suite of image and video neural tokenizers. https://research.nvidia.com/labs/dir/cosmos-tokenizer, 2024.

[34] W. Yan, Y. Zhang, P. Abbeel, and A. Srinivas. VideoGPT: Video generation using vq-vae and transformers. In *arXiv preprint*, 2021.

[35] G. L. Moing, J. Ponce, and C. Schmid. CCVS: Context-aware controllable video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[36] S. Ge, T. Hayes, H. Yang, X. Yin, G. Pang, D. Jacobs, J.-B. Huang, and D. Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[37] Y. Seo, K. Lee, F. Liu, S. James, and P. Abbeel. HARP: Autoregressive latent video prediction with high-fidelity image generator. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2022.

[38] C. Hawthorne, A. Jaegle, C. Cangea, S. Borgeaud, C. Nash, M. Malinowski, S. Dieleman, O. Vinyals, M. Botvinick, I. Simon, H. Sheahan, N. Zeghidour, J.-B. Alayrac, J. Carreira, and J. Engel. General-purpose, long-context autoregressive modeling with Perceiver AR. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

[39] V. Micheli, E. Alonso, and F. Fleuret. Transformers are sample-efficient world models. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[40] W. Yan, D. Hafner, S. James, and P. Abbeel. Temporally consistent transformers for video generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

[41] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, and D. Erhan. Phenaki: Variable length video generation from open domain textual description. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[42] L. Yu, Y. Cheng, K. Sohn, J. Lezama, H. Zhang, H. Chang, A. G. Hauptmann, M.-H. Yang, Y. Hao, I. Essa, and L. Jiang. MAGVIT: Masked Generative Video Transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[43] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint*, 2023.

[44] E. Alonso, A. Jelley, V. Micheli, A. Kanervisto, A. Storkey, T. Pearce, and F. Fleuret. Diffusion for world modeling: Visual details matter in atari. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[45] OpenAI. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators, 2024.

[46] Runway. Introducing gen-3 alpha: A new frontier for video generation. https://runwayml.com/research/introducing-gen-3-alpha, 2024.

[47] comma.ai. commavq. https://github.com/commaai/commavq, 2023.

[48] R. Chen, Z. Wu, Y. Liu, Y. Guo, J. Ni, H. Xia, and S. Xia. Unimlvg: Unified framework for multi-view long video generation with comprehensive control capabilities for autonomous driving. *arXiv preprint arXiv:2412.04842*, 2024.

[49] J. Ni, Y. Guo, Y. Liu, R. Chen, L. Lu, and Z. Wu. Maskgwm: A generalizable driving world model with video mask reconstruction. *arXiv preprint arXiv:2502.11663*, 2025.

[50] E. Ma, L. Zhou, T. Tang, Z. Zhang, D. Han, J. Jiang, K. Zhan, P. Jia, X. Lang, H. Sun, et al. Unleashing generalization of end-to-end autonomous driving with controllable long video generation. *arXiv preprint arXiv:2406.01349*, 2024.

[51] M. Hassan, S. Stapf, A. Rahimi, P. Rezende, Y. Haghighi, D. Brüggemann, I. Katircioglu, L. Zhang, X. Chen, S. Saha, et al. Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. *arXiv preprint arXiv:2412.11198*, 2024.

[52] J. Mao, B. Li, B. Ivanovic, Y. Chen, Y. Wang, Y. You, C. Xiao, D. Xu, M. Pavone, and Y. Wang. Dreamdrive: Generative 4d scene modeling from street view images. *arXiv preprint arXiv:2501.00601*, 2024.