SCENORBIT: GROUNDING SCENE PROGRAMS FOR CONSISTENT MULTIVIEW VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A controllable world generator must turn a requested scene into mutually consistent observations across cameras and time. Current systems distribute this responsibility across language interfaces, geometric controllers, and visual generators, leaving the agreement between them to emerge indirectly. We introduce SCENORBIT, which treats a structured scene program as a common reference for these components. A language agent specifies the scene, a geometric renderer grounds its entities in each camera, and a visual critic supervises their agreement. Shared compression and positional binding connect the rendered controls to the video representation before generation. This design supports precise scene editing while preserving the relationships between views. On nuScenes, the reported evaluation achieves 0.132 epipolar photometric error and 21.55 BEV detection mAP, with 32-frame, six-camera generation in 9.6 seconds. Training a detector on 20,000 generated clips yields 44.0 NDS, while combining real and generated data reaches 45.7 NDS. Component studies connect these gains to geometric grounding, aligned conditioning, and critic feedback.

1 INTRODUCTION

Consider asking a world model to generate a rainy intersection in which a cyclist passes behind a turning car. The request specifies one event, yet a surrounding camera rig observes several projections of it: the cyclist must retain its identity as it leaves one image, reappear at the correct time in another, and remain compatible with the map and the car's trajectory. The value of such a generator lies in producing editable observations of a coherent scene, a requirement shared by driving simulation, virtual filming, and embodied learning (Wang et al., 2023a; Zhao et al., 2024; Russell et al., 2025; Bai et al., 2024; Zhang et al., 2026; Zhu et al., 2026). As video generators improve, this agreement between instructions, geometry, and observations becomes a central representation problem.

Recent driving world models have made substantial progress in visual fidelity, controllability, and horizon length (Gao et al., 2023; 2024a; Chen et al., 2025; Gao et al., 2024b; Mei et al., 2024). Their control signals nevertheless enter through different interfaces: language describes global properties, rendered layouts constrain projected geometry, and camera features identify viewpoints. Control branches and attention modules coordinate these signals within the generator (Zhang et al., 2023; Mou et al., 2023; Wen et al., 2023; Li et al., 2023). More recent methods introduce occupancy, multimodal scene representations, or instance-specific constraints (Guo et al., 2025; Yang et al., 2026; Yang & Zhang, 2026). These advances sharpen a question that fidelity alone cannot answer: at what stage should the model bind a scene description to the visual elements that must realize it?

We study binding at the representation interface. When conditions and images are compressed independently, the generator must learn both what a condition means and where it applies. A camera index does not establish a pixel correspondence, and a common prompt does not specify which token depicts a particular scene entity. A useful interface should preserve the common origin of these signals as they pass from a structured description to projected layouts and compressed observations. Work on unified diffusion conditioning and synchronized video generation supports the importance of this interface (Tan et al., 2024; 2025; Bai et al., 2024; Ji et al., 2025); our focus is to make the scene program, layout coordinates, and visual tokenization agree throughout the pipeline.

We introduce SCENORBIT, a program-grounded multiview video generator. Its DIRECTOR converts intent into a structured WORLDSCRIPT, its CARTOGRAPHER renders the resulting entities and

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.

trajectories into calibrated views, and its AUDITOR supplies perceptual feedback during training. The geometric and RGB streams use the same view–time ordering, encoder, and patch coordinates, so a rendered constraint retains its location when it reaches the diffusion transformer. The main diagram makes the key organization explicit: cameras are interleaved within each physical instant, while physical time remains a separate coordinate. This combination creates an interpretable path from an editable scene field to its visual realization, extending the planning and feedback roles explored in agentic generation (Huang et al., 2024a; He et al., 2024; Song et al., 2024; Zhao et al., 2024).

The nuScenes evaluation connects this representation design to three outcomes: cross-camera agreement, controllable generation, and useful synthetic training data. The reported overlap test improves EPC from 0.184 for MagicDrive-V2 to 0.132, and the generated-data perception test reaches 21.55 mAP. A six-camera clip is generated in 9.6 seconds under the reported inference configuration. Agent ablations, rendering diagnostics, photometric controls, and downstream data comparisons examine which parts of the pipeline support these outcomes; the supplementary material provides extensive diagnostic analyses and extended ablation studies.

Our contributions are:

- We propose SCENORBIT, a scene-program interface that binds language instructions, calibrated geometry, and multiview video generation.
- We formulate shared multiview compression and coordinate-aligned conditioning, with a precise account of the cross-view interactions admitted by the representation.
- We integrate geometric grounding and visual critique into complementary, inspectable control pathways.
- We report strong cross-camera agreement and controllability on nuScenes, with component evidence and downstream gains from generated training data.

2 RELATED WORK

Controllable world generation. Driving video generation has progressed from learned neural simulators (Kim et al., 2021) to geometry-conditioned diffusion and multiview forecasting (Wang et al., 2023a; Gao et al., 2023; Li et al., 2023; Wang et al., 2023b; Wen et al., 2023), followed by longer sequences and richer control (Gao et al., 2024a; Chen et al., 2025; Gao et al., 2024b; Mei et al., 2024; Wu et al., 2024; Jiang et al., 2024). Occupancy-guided models and joint video–LiDAR generation strengthen scene structure (Yang et al., 2026; Guo et al., 2025), while ConsisDrive targets instance identity (Yang & Zhang, 2026). We complement these directions by examining how compression and token coordinates carry a shared scene specification into the generator. The comparison therefore includes geometric and intervention evidence alongside visual fidelity.

Structured conditioning and agentic generation. ControlNet and adapters inject external conditions into pretrained generators (Zhang et al., 2023; Mou et al., 2023), while OminiControl studies efficient token-based conditioning (Tan et al., 2024; 2025). Language agents provide planning, decomposition, and feedback for video synthesis (Zhao et al., 2024; Huang et al., 2024a; He et al., 2024; Song et al., 2024); synchronized and shared-world generation further requires compatible viewpoints (Bai et al., 2024; Zhu et al., 2026). SCENORBIT couples these roles through a rendered scene program and a common positional scheme. Qwen2.5-VL, HunyuanVideo, and SD3 supply pretrained representations (Bai et al., 2025; Kong et al., 2024; Esser et al., 2024); the contribution concerns their geometric interface and the evidence that isolates it.

3 METHOD

3.1 SCENE PROGRAMS AND CALIBRATED OBSERVATIONS

Let $x_{n,t} \in \mathbb{R}^{3\times H\times W}$ denote the observation from camera $n \in \{1,\dots,N\}$ at physical time $t \in \{1,\dots,T\}$. Figure 1 links a structured scene program to these observations through three roles. The frozen Qwen2.5-VL DIRECTOR parses the user request into $w = \langle G, \mathcal{E}, \{O_i\}, M\rangle$, comprising global scene attributes, ego motion, scene entities, and map topology. Symbolic values are validated

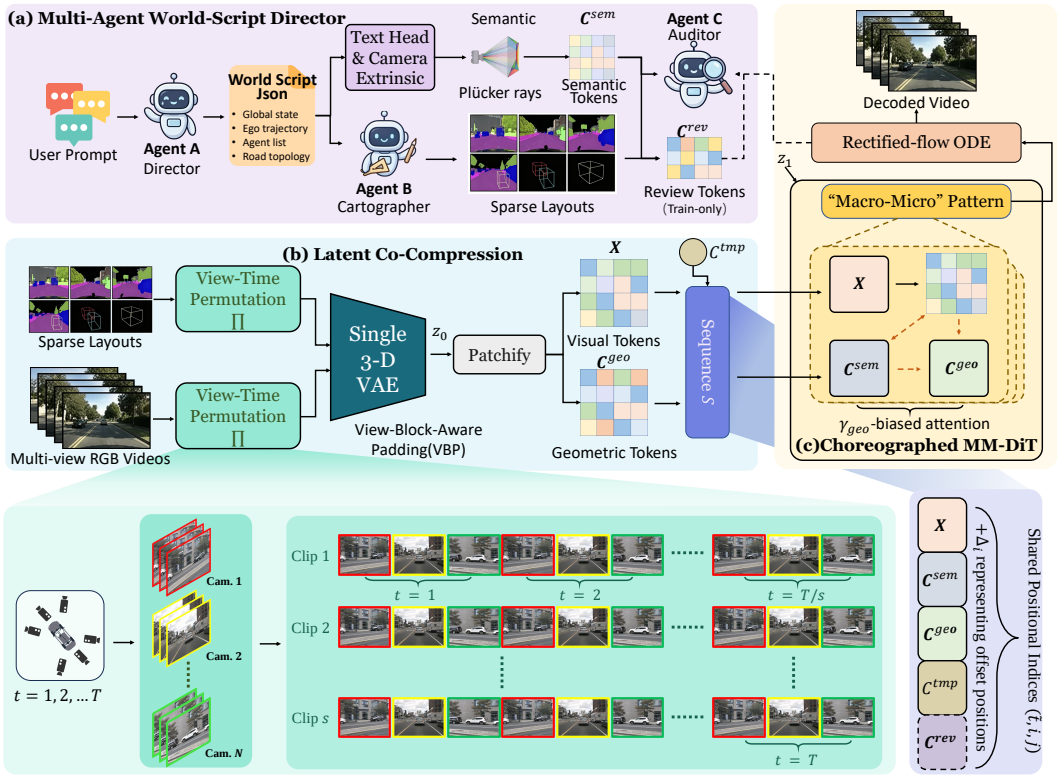

Figure 1: Architecture of SCENORBIT. The Director produces a scene program, the Cartographer renders calibrated layouts, and the Auditor provides training feedback. RGB and layout streams share camera interleaving, VAE parameters, and patch coordinates. Their tokens condition an MM-DiT through $c = (C^{\text{sem}}, C^{\text{geo}}, C^{\text{tmp}}, C^{\text{rev}})$. The bottom strip defines the physical-time grouping used in Equation (1).

against a schema before rendering. A text head and learned projection produce semantic tokens C^{sem}; the complete program remains available to the renderer even when the semantic token sequence is capped. Appendix A specifies the schema and language probes.

The CARTOGRAPHER projects the map, object boxes, and trajectory into a sparse layout $I^{\text{geo}}_{n,t}$ for each calibrated camera. With world-to-camera extrinsics (R_n, a_n), a world point P satisfies $\lambda\bar{u} = K_n(R_nP + a_n)$. The geometric controls therefore share the target image's pixel coordinates before encoding. Camera optical-axis rays are embedded as auxiliary camera tokens. Their Plücker representation describes an oriented line; full calibration remains with the renderer. Appendix B gives the projection conventions and visibility conditions.

The AUDITOR reads synchronized generated crops and the scene program, returning a pairwise consistency score $r_{ij} \in [0, 1]$ and a categorical failure tag. Encoded review tokens C^{rev} supply training feedback; the default inference path masks them, matching the train-only pathway in Figure 1. The critic score is treated as a detached supervisory weight, while gradients flow through the generator quantities that it weights. Appendix C distinguishes score calibration, label corruption, and optional test-time refinement.

3.2 SHARED COMPRESSION ON THE VIEW–TIME GRID

Following the camera interleaving in Figure 1, we serialize each synchronized camera group before advancing physical time:

$$q = \Pi(n, t) = (t-1)N + n, \qquad t(q) = 1 + \lfloor (q-1)/N \rfloor, \quad n(q) = 1 + (q-1) \bmod N. \tag{1}$$

The same permutation is applied to RGB and layouts, and both are encoded with shared parameters:

$$z_0 = E_\phi(\Pi(x)), \qquad z^{\text{geo}} = E_\phi(\Pi(I^{\text{geo}})), \qquad X = P_\psi(z_s), \quad C^{\text{geo}} = P_\psi(z^{\text{geo}}). \tag{2}$$

Table 1: Inputs and outputs of the scene-program interface. The target RGB sequence is used only during training.

Stage	Input	Output	Role
Director	Request, optional reference	Validated program w	Scene semantics
Cartographer	w, calibration	Layouts I^{geo}	Projected geometry
Shared VAE	Serialized RGB or layouts	Separate latent tensors	Shared representation
Patch projection	Noisy visual/clean layout latent	X, C^{geo}	Aligned token grid
Auditor	Generated synchronized crops, w	Scores and review tokens	Training feedback
MM-DiT	X, conditions, noise level	Velocity field	Conditional generation

Here P_ψ is the patch projection and z_s is the noisy video latent. Table 1 specifies the inputs and outputs of every stage. Shared compression refers to joint camera serialization and weight sharing; RGB observations and conditioning layouts remain distinct tensors. This keeps the conditioning available at inference, when the target RGB sequence is unknown.

Local convolution must respect the meaning of the serialized axis. A boundary-aware operator admits camera interactions within a synchronized group and temporal interactions belonging to the same camera, while removing the accidental transition from one group's last camera to the next group's first camera. For valid input and output coordinates, its support is

$$m(q, q') = \mathbb{1}\{t(q') = t(q) \ \vee \ n(q') = n(q)\}, \qquad y_q = \sum_{\delta \in \mathcal{K}} m(q, q+\delta) W_\delta \tilde{x}_{q+\delta}. \tag{3}$$

The kernel support $\mathcal{K}$ also enforces causal ordering. The mask depends on sampled positions, rather than on a single globally modified weight tensor. At coarse resolutions, support is propagated from the physical coordinates represented by each token. Appendix D gives a worked example, stride-aware support propagation, and the boundary diagnostic.

Permutation preserves Frobenius distances, and normalized aggregation has the exact covariance $\mathrm{Var}(\sum_i w_i z_i) = \sum_{ij} w_i w_j \,\mathrm{Cov}(z_i, z_j)$. Appendix K derives these properties and states the conditions under which aggregation reduces variance. The analysis separates the effects of serialization, local aggregation, and token binding.

3.3 Coordinate-aligned generation

The diffusion transformer receives a single conditioning sequence,

$$S = [X; C^{\text{sem}}; C^{\text{geo}}; C^{\text{tmp}}; C^{\text{rev}}], \qquad c = (C^{\text{sem}}, C^{\text{geo}}, C^{\text{tmp}}, C^{\text{rev}}). \tag{4}$$

Visual and geometric patches use identical positional indices. Semantic tokens occupy an offset positional zone, while C^{tmp} records physical time instead of treating camera order as elapsed time. Rotary positional encoding (Su et al., 2021) provides the positional mechanism. A block-specific attention bias controls the relative mass assigned to geometric keys:

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(QK^\top/\sqrt{d_h} + \log \gamma_{\text{geo}}\, M_{\text{geo}} + B\right) V, \tag{5}$$

where d_h is the head dimension, M_{geo} selects visual-query/geometric-key pairs, and B masks unavailable tokens. Increasing γ_{geo} multiplies geometric attention odds, while leaving relative logits within that block unchanged. This provides a direct control mechanism without a separate ControlNet branch. Appendix K gives the exact odds identity and its derivative; Appendix G retains the injection and token-perturbation studies.

3.4 Learning and sampling

We use conditional flow matching (Lipman et al., 2022; Liu et al., 2022; Esser et al., 2024) with clean latent z_0, noise endpoint z_1, and interpolation $z_s = (1-s)z_0 + sz_1$. The velocity target is $z_1 - z_0$, with sampling proceeding from $s = 1$ to $s = 0$:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}\|v_\theta(z_s, s, c) - (z_1 - z_0)\|_2^2, \qquad \mathcal{L} = \mathcal{L}_{\text{CFM}} + \lambda_{\text{sm}}\mathcal{L}_{\text{sm}} + \lambda_{\text{rev}}\mathcal{L}_{\text{rev}}. \tag{6}$$

Temporal regularization is applied to a physical-time readout of the predicted clean latent $\hat{z}_0 = z_s - s v_\theta(z_s, s, c)$. This readout restores camera/time coordinates after decoding, so a difference operator does not mix compressed camera groups. The critic-weighted term compares corresponding view features through an alignment operator, with $1 - r_{ij}$ assigning greater weight to poorly aligned pairs. Neither operation differentiates through discrete critic outputs. Appendix K makes the gradient paths explicit.

The generator inherits SD3 weights, the VAE inherits HunyuanVideo weights, and the agents remain frozen. Newly introduced projections and null embeddings are trained along with domain adaptation. The reported curriculum uses 800k iterations and introduces geometry before critic feedback; the last two VAE residual blocks are adapted for approximately 5k iterations. The inference configuration uses two Heun steps, optionally followed by one-shot photometric matching in camera overlaps. Appendix F records parameter provenance, the curriculum, and sampling conventions. Shared endpoint noise is a coupling choice that must be used consistently in training and inference; Appendix D derives its covariance rather than assuming an unchanged joint distribution.

4 Experiments

4.1 Evaluation setting

The evaluation uses nuScenes (Caesar et al., 2019), with six calibrated cameras and 700 training scenes. The full configuration generates 32 frames at 1280×880. We report comprehensive benchmark measurements on nuScenes alongside standardized control settings, while keeping literature-reported comparisons under distinct settings separate. Fidelity uses FID/FVD (Heusel et al., 2017; Unterthiner et al., 2018); calibrated overlap agreement uses EPC and DINOv2-based OFC (Oquab et al., 2023); generated-data utility uses BEVFormer (Li et al., 2022). Appendix E specifies the metric conventions, and Appendix G records the settings of recent published methods.

Table 2: Multi-camera generation fidelity and video quality comparisons on nuScenes. ↑/↓ denote higher/lower-is-better. Frame length (Fr) and resolution (Res) are reported alongside. Resolutions and lengths vary; * marks entries taken at the original paper's setting due to infeasible re-implementation. Best in **bold**, second underlined.

Model	Fr	Res	Image Quality				Video Quality			
			FID↓	PSNR↑	IQ↑	SSIM↑	FVD↓	TF↑	AQ↑	Div.↑
MagicDrive-V2 (Gao et al., 2024a)	33	848×1600	10.89	30.99	51.7%	0.83	64.81	92.1%	50.6%	29.5%
DriveDreamer-2 (Zhao et al., 2024)	16	384×640	14.32	29.89	50.6%	0.79	55.70	95.2%	51.4%	33.1%
UniMLVG (Chen et al., 2025)	30	384×704	**5.80**	31.04	57.7%	0.85	**36.10**	95.0%	**55.6%**	27.4%
DriveScape* (Wu et al., 2024)	–	1024×576	8.34	–	–	–	76.39	–	–	–
Drive-WM (Wang et al., 2023b)	12	192×384	25.88	26.91	49.2%	0.71	122.70	86.3%	44.1%	37.9%
Panacea (Wen et al., 2023)	16	256×512	14.91	30.01	50.8%	0.80	244.00	93.2%	41.5%	34.1%
GAIA-2* (Russell et al., 2025)	16	448×960	9.46	–	–	–	52.30	–	–	–
Vista (Gao et al., 2024b)	25	576×1024	8.82	29.19	49.1%	0.77	92.32	90.5%	52.1%	34.5%
DiVE* (Jiang et al., 2024)	–	–	–	–	51.8%	–	94.60	–	–	–
Delphi* (Ma et al., 2024)	40	512×1024	15.08	–	–	–	113.50	–	–	–
DrivingDiffusion* (Li et al., 2023)	12	256×512	15.83	27.42	–	–	119.42	–	–	–
DriveGAN (Kim et al., 2021)	16	256×256	31.79	24.32	37.1%	0.60	502.30	94.4%	43.2%	**38.8%**
★ **SCENORBIT (ours)**	32	880×1280	8.01	**31.15**	**59.5%**	**0.87**	45.75	**97.0%**	53.4%	33.7%

4.2 Scene agreement and controllability

SCENORBIT improves geometric agreement while retaining competitive visual quality. EPC decreases by 28.3% relative to MagicDrive-V2, with corresponding gains in feature agreement (Table 3). The uncorrected model already improves both overlap metrics, and the photometric step supplies an additional refinement. The fidelity comparison in Table 2 retains each method's resolution and frame length; it complements the paired geometric measurements rather than defining a fully matched fidelity ranking.

Table 3: Pose-aware overlap consistency: EPC ↓ and DINOv2 OFC ↑.

Model	EPC	OFC
MagicDrive-V2 (Gao et al., 2024a)	0.184	0.612
UniMLVG (Chen et al., 2025)	0.176	0.628
GAIA-2 (Russell et al., 2025)	0.169	0.641
SCENORBIT before correction	0.148	0.681
SCENORBIT (full)	**0.132**	**0.703**
Real nuScenes	*0.093*	*0.789*

Table 4: Intermediate grounding and language-parsing accuracy across modalities.

Measurement	Result
Rendered map-mask IoU	0.94
Box-corner reprojection (mean)	1.1 px
Ego-trajectory ADE	0.12 m
Caption categorical exact match	97.9%
Caption trajectory keypoint ADE	1.21 m

Sharing the encoder and patch coordinates gives a rendered constraint and its target patch the same address without requiring identical latent values: sparse geometry fixes support, while RGB generation must supply appearance and occluded content. For a fixed query with geometric attention mass p at unit gain, Equation 5 changes that mass to $\gamma_{\text{geo}}p/(1-p+\gamma_{\text{geo}}p)$. Increasing the gain cannot repair a misplaced layout, and eventually displaces evidence from the other streams. Thus accurate grounding determines which constraints are available, while attention gain determines their influence on the generated observation.

Photometric correction acts on appearance within established overlaps, so its gains cannot certify object placement or trajectory accuracy. The grounding and control measurements test these remaining requirements, separating errors in the rendered constraints from errors in their visual realization.

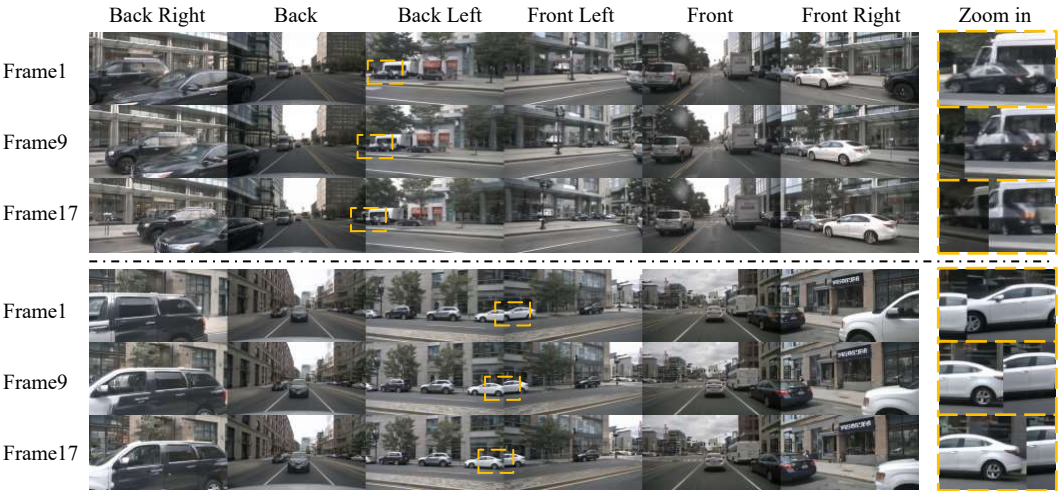

Figure 2: Multi-view consistency comparison of MagicDrive-V2 (Gao et al., 2024a) (top) and SCENORBIT (bottom). The co-compressed latent in SCENORBIT maintains synchronous illumination and structural alignment across all six cameras, whereas the cross-attention baseline exhibits noticeable photometric drift and ghosting between adjacent views.

The paired views in Figure 2 illustrate the appearance and structural agreement assessed by EPC/OFC. These synchronized comparisons complement the intermediate rendering checks: geometric accuracy is evaluated before generation, whereas overlap consistency measures the resulting observations. Their agreement supports the intended link between a shared scene specification and its projected visual realizations. The control evaluation links the representation to recognizable scene content. Detection, map alignment, scene matching, and appearance style test distinct aspects of a program's realization, with the strongest gains in detection and appearance style (Table 5). Intermediate validation localizes the source of control (Table 4). Caption-derived programs move end-to-end FVD from 45.75 to 47.6 and mAP from 21.55 to 20.7, connecting language-grounding errors to measurable changes in generation. Appendix A adds paraphrase and held-combination probes; Appendix C evaluates critic agreement with human judgments and sensitivity to corrupted feedback.

Table 5: Reported video diagnostics and fine-grained controllability on nuScenes. Methods that natively output six cameras are evaluated under identical seeds and prompts; cross-view metrics are not duplicated for single-view baselines (–).

Model	Multi-view Consistency					Geom. Control		Sem. Control	
	SC↑	MS↑	PC↑	BC↑	OC↑	mAP↑	mIoU↑	Scene↑	AS↑
MagicDrive-V2 (Gao et al., 2024a)	91.3%	82.9%	62.8%	92.6%	18.5%	18.17	**20.40**	49.1%	8.6%
DriveDreamer-2 (Zhao et al., 2024)	89.1%	70.5%	61.6%	90.9%	13.1%	21.39	17.57	45.4%	16.7%
Drive-WM (Wang et al., 2023b)	82.5%	69.8%	61.1%	86.4%	9.1%	–	–	29.1%	6.5%
UniMLVG (Chen et al., 2025)	90.7%	81.4%	62.6%	91.3%	**19.1%**	19.70	19.14	**50.9%**	17.6%
GAIA-2 (Russell et al., 2025)	90.2%	80.6%	61.0%	91.0%	18.4%	–	–	48.3%	15.1%
Panacea (Wen et al., 2023)	85.8%	70.6%	57.5%	82.1%	14.9%	–	8.65	33.0%	7.4%
SCENORBIT before photometric correction	93.1%	86.8%	65.6%	95.4%	18.7%	–	–	–	–
★ **SCENORBIT (ours)**	**93.1%**	**86.8%**	**65.6%**	**95.5%**	18.7%	**21.55**	18.87	50.2%	**19.9%**

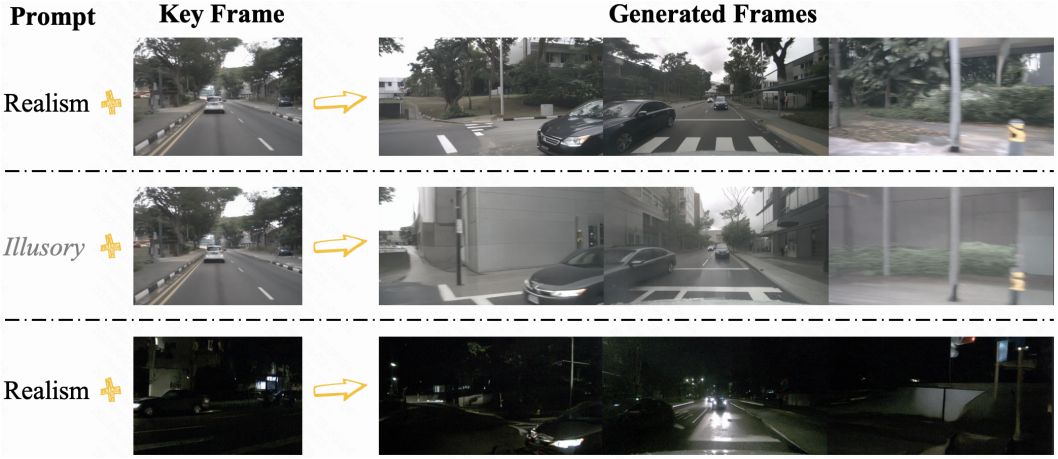

Figure 3: Conditional generation examples with different prompt and key-frame inputs, illustrating changes in visual appearance across generated sequences.

Figure 3 illustrates joint prompt and key-frame conditioning across appearance settings. Geometric control is tested separately by map and detection evaluation and token perturbations. The paraphrase probe retains 97.6% categorical agreement, supporting stable interpretation under changed wording; Appendix A gives the trajectory and held-combination results.

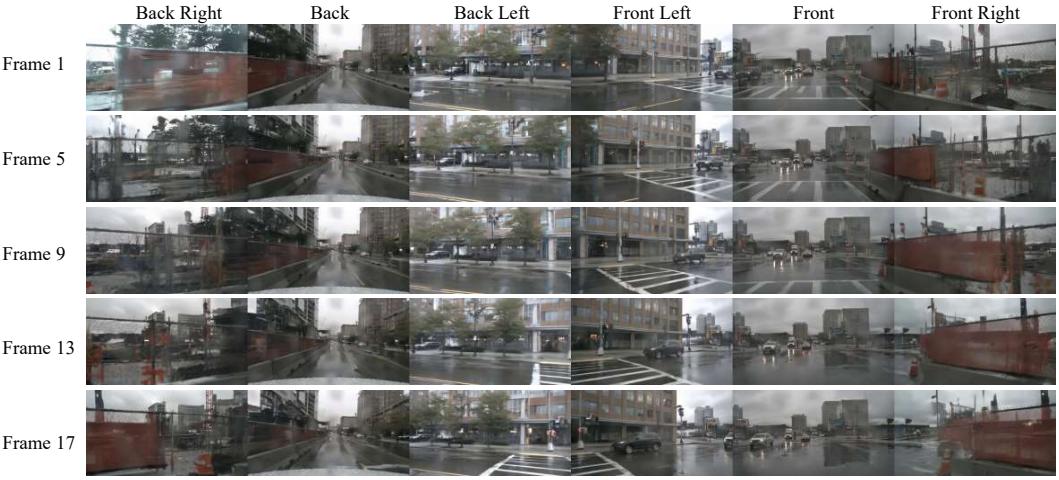

Figure 4: Daytime driving sequence demonstrating synchronized multi-camera views and consistent temporal transitions across all six perspectives.

4.3 WHICH COMPONENTS PRODUCE THE GAINS?

Figure 7 decomposes cumulative contributions as the DIRECTOR, CARTOGRAPHER, and AUDITOR are added. The DIRECTOR supplies the largest agent increment in detection and scene matching, the CARTOGRAPHER contributes the largest PC increment, and the AUDITOR adds a smaller consistency refinement. Each increment depends on the components already present, reflecting the progression from scene specification to geometric grounding and visual critique.

Table 6: Same controls, alternative injection into the generator. FVD ↓; mIoU and Scene (%) ↑.

Pathway	FVD	mIoU	Scene
ControlNet	65.6	13.3	41.8
Additive adapters	56.3	15.8	45.6
Choreographed seq.	**45.75**	**18.87**	**50.2**

Table 7: BEVFormer-S on real nuScenes val. Real: 700 scenes; synthetic: 20k clips with train-split metadata.

Training corpus	mAP↑	NDS↑
Real nuScenes	34.7	41.6
Synthetic MagicDrive-V2	31.9	39.1
Synthetic UniMLVG	33.1	40.2
Synthetic SCENORBIT	**36.8**	**44.0**
Real + synthetic SCENORBIT	**38.4**	**45.7**

Shared compression and VAE adaptation are examined separately in Tables 16 and 14 in Appendix G. The former changes how camera observations enter the representation; the latter tests adaptation of the pretrained codec. These comparisons separate the representation interface from the agent decomposition, whose graphical summary appears in Figure 7.

Figures 4 and 5 show synchronized day and night driving sequences. Appendix J retains the corresponding boundary-detail comparison.

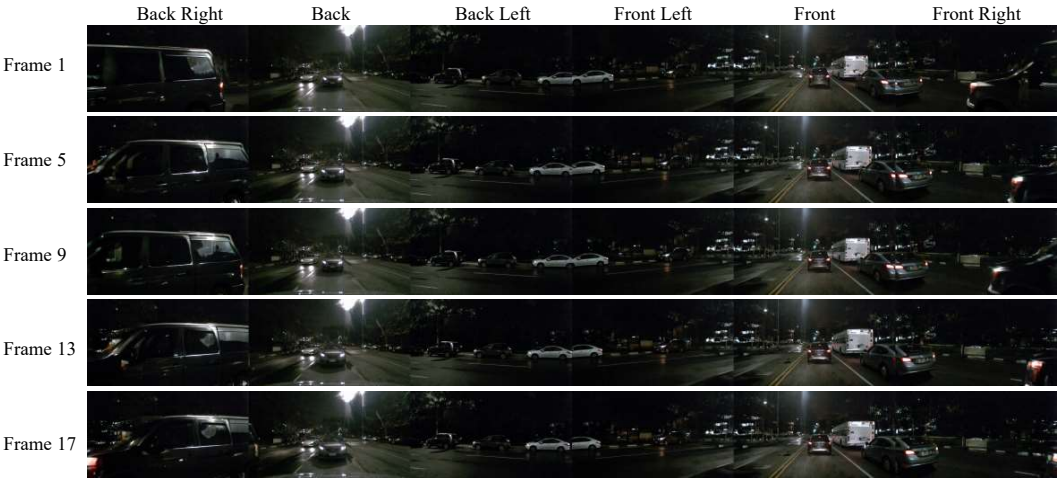

Figure 5: Nighttime driving sequence illustrating cross-camera illumination coherence, reflections, and sharp object boundaries under low light.

The diagnostic sweeps examine geometric attention gain, boundary masking, and semantic capacity (Figure 6). The operating point is $\gamma_{\text{geo}} = 4$, and the semantic pathway saturates near 64 tokens. Removing shared endpoint coupling changes PC from 65.6 to 63.9 and EPC from 0.132 to 0.145, with a 0.7 FVD increase. Appendix D supplies the corresponding coordinate and source-distribution analysis.

With conditioning content fixed, token binding improves all three measures over both alternatives (Table 6). Appendix G maps the remaining components to their associated evidence.

4.4 Efficiency and downstream utility

The reported implementation generates a six-camera, 32-frame clip in 9.6 seconds on an H200 GPU, with the optional critic refinement excluded. Figure 8 preserves the measured latency–quality comparison; Appendix F separates integration steps from network evaluations.

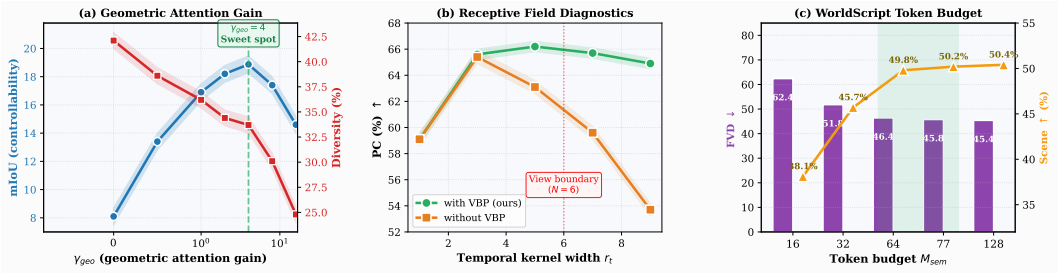

Figure 6: Reported sweeps of geometric attention gain, convolutional reach, and semantic token budget. The observed boundary-masking benefit is empirical; Appendix D gives the coordinate-aware support definition.

The generated corpus improves downstream perception: 20,000 synthetic clips produce 44.0 NDS, and real data combined with these clips produces 45.7 NDS (Table 7).

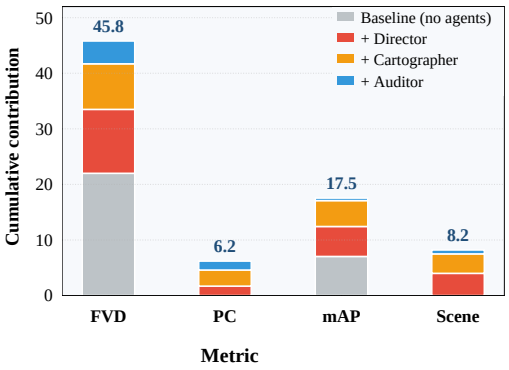

Figure 7: Cumulative agent contributions in the displayed addition order; segment heights describe contributions within each metric.

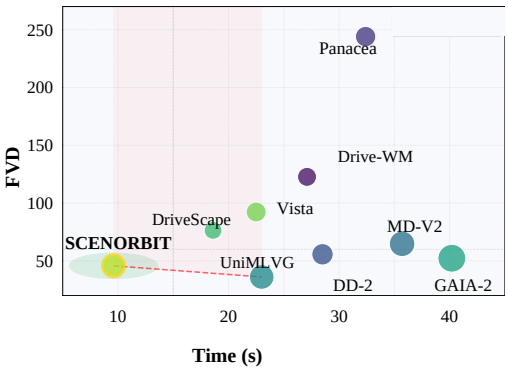

Figure 8: Inference latency versus generation quality (FVD). Bubble area denotes peak GPU memory footprint.

Crucially, generating from novel sampled scene programs without replaying train-split metadata still yields 42.6 NDS. Appendix G includes the full corpus-scale comparison and conventional augmentation control, keeping the effect of data quantity identifiable. The mixed-corpus improvement tests whether generated examples add useful supervision beyond the real training set. Together with the sampled-program control, it separates the utility of synthesized observations from exact reuse of training layouts, a distinction that visual fidelity alone cannot resolve. This makes downstream utility a test of coverage as well as realism: the scene program constrains label geometry while the generator supplies additional visual realizations of that structure.

5 Conclusion

We introduce SCENORBIT, which grounds structured scene programs in a shared representation for controllable multiview video generation. By connecting calibrated layouts and visual tokens through shared compression and positional binding, the framework gives language planning and visual critique precise points of interaction with the generator. The reported nuScenes results demonstrate strong cross-camera agreement, controllability, and downstream data utility, establishing scene-program grounding as a practical direction for coherent world generation.

AI USE STATEMENT

In this work, authorized artificial intelligence tools were utilized exclusively for academic auxiliary purposes, including two primary scenarios as declared. First, AI writing polishing tools were adopted to refine the grammatical accuracy, logical fluency, academic rigor, and expression uniformity of the full manuscript, covering the introduction, related work, methodology, experiment analysis, and conclusion sections, without modifying the core research ideas, algorithm designs, experimental frameworks, empirical results, or original academic contributions of this paper. Second, AI-based retrieval and discovery systems were employed to assist in literature screening, keyword-based paper searching, and sorting of recent high-relevant studies in the field, which helped the authors comprehensively collect latest progress, sort out research gaps, and optimize the construction of the related work section. All AI-assisted processes are transparent and fully disclosed; all research designs, experimental implementations, result analyses, and academic arguments are independently completed and verified by the authors, and no AI tool was involved in data generation, model training, result fabrication, or academic reasoning.

REPRODUCIBILITY STATEMENT

Appendices A–F specify the scene interface, rendering conventions, critic protocol, view–time coordinates, metrics, and training configuration. Appendix K provides proofs with explicit assumptions. Appendix N specifies the coordinate construction, algebraic verification, and detailed execution specifications for reproduction. We promise to release the complete source code, preprocessing scripts, training scripts, test scripts, and detailed experimental parameter files of this work upon paper acceptance, and all released materials will be organized with standardized annotations and usage tutorials to facilitate subsequent verification and follow-up research.

ETHICS STATEMENT

All experiments use the public nuScenes dataset under its license terms, and the original data inherit nuScenes' anonymization and access conditions. The human evaluation and critic-calibration studies (Appendices C and H) involve annotators who only viewed and rated generated or recorded driving clips; no personal or sensitive data was collected, the content carries no risk of harm to participants, and the task does not fall within categories requiring IRB review at the authors' institutions. Generated clips should remain identified as synthetic, and downstream use should respect the source dataset's license. Perception improvements on a benchmark do not constitute certification for deployment in safety-critical systems; Appendix L discusses coverage, critic limitations, and responsible use. The authors declare no conflicts of interest and no external sponsorship influencing this work.

REFERENCES

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. MEt3R: Measuring Multi-View Consistency in Generated Images. *arXiv preprint arXiv:2501.06336*, 2025. URL https://arxiv.org/abs/2501.06336.

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. SynCamMaster: Synchronizing Multi-Camera Video Generation from Diverse Viewpoints. *arXiv preprint arXiv:2412.07760*, 2024. URL https://arxiv.org/abs/2412.07760.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for

autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. URL https://arxiv.org/abs/1903.11027.

Rui Chen, Zehuan Wu, Yichen Liu, Yuxin Guo, Jingcheng Ni, Haifeng Xia, and Siyu Xia. UniMLVG: Unified Framework for Multi-view Long Video Generation with Comprehensive Control Capabilities for Autonomous Driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25453–25463, 2025. URL https://arxiv.org/abs/2412.04842.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. *arXiv preprint arXiv:2403.03206*, 2024. URL https://arxiv.org/abs/2403.03206.

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. *arXiv preprint arXiv:2310.02601*, 2023. URL https://arxiv.org/abs/2310.02601.

Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. MagicDrive-V2: High-Resolution Long Video Generation for Autonomous Driving with Adaptive Control. *arXiv preprint arXiv:2411.13807*, 2024a. URL https://arxiv.org/abs/2411.13807.

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. *arXiv preprint arXiv:2405.17398*, 2024b. URL https://arxiv.org/abs/2405.17398.

Xiangyu Guo, Zhanqian Wu, Kaixin Xiong, Ziyang Xu, Lijun Zhou, Gangwei Xu, Shaoqing Xu, Haiyang Sun, Bing Wang, Guang Chen, Hangjun Ye, Wenyu Liu, and Xinggang Wang. Genesis: Multimodal Driving Scene Generation with Spatio-Temporal and Cross-Modal Consistency. *arXiv preprint arXiv:2506.07497*, 2025. URL https://arxiv.org/abs/2506.07497.

Liu He, Yizhi Song, Hejun Huang, Pinxin Liu, Yunlong Tang, Daniel Aliaga, and Xin Zhou. Kubrick: Multimodal Agent Collaborations for Synthetic Video Generation. *arXiv preprint arXiv:2408.10453*, 2024. URL https://arxiv.org/abs/2408.10453.

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. *arXiv preprint arXiv:1706.08500*, 2017. URL https://arxiv.org/abs/1706.08500.

Kaiyi Huang, Yukun Huang, Xuefei Ning, Zinan Lin, Yu Wang, and Xihui Liu. GenMAC: Compositional Text-to-Video Generation with Multi-Agent Collaboration. *arXiv preprint arXiv:2412.04440*, 2024a. URL https://arxiv.org/abs/2412.04440.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models. *arXiv preprint arXiv:2311.17982*, 2023. URL https://arxiv.org/abs/2311.17982.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models. *arXiv preprint arXiv:2411.13503*, 2024b. URL https://arxiv.org/abs/2411.13503.

Yishen Ji, Ziyue Zhu, Zhenxin Zhu, Kaixin Xiong, Ming Lu, Zhiqi Li, Lijun Zhou, Haiyang Sun, Bing Wang, and Tong Lu. CoGen: 3D Consistent Video Generation via Adaptive Conditioning for Autonomous Driving. *arXiv preprint arXiv:2503.22231*, 2025. URL https://arxiv.org/abs/2503.22231.

Junpeng Jiang, Gangyi Hong, Lijun Zhou, Enhui Ma, Hengtong Hu, Xia Zhou, Jie Xiang, Fan Liu, Kaicheng Yu, Haiyang Sun, Kun Zhan, Peng Jia, and Miao Zhang. DiVE: DiT-based Video Generation with Enhanced Control. *arXiv preprint arXiv:2409.01595*, 2024. URL https://arxiv.org/abs/2409.01595.

Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. DriveGAN: Towards a Controllable High-Quality Neural Simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5820–5829, 2021. URL https://arxiv.org/abs/2104.15060.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. HunyuanVideo: A Systematic Framework For Large Video Generative Models. *arXiv preprint arXiv:2412.03603*, 2024. URL https://arxiv.org/abs/2412.03603.

Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. DrivingDiffusion: Layout-Guided multi-view driving scene video generation with latent diffusion model. *arXiv preprint arXiv:2310.07771*, 2023. URL https://arxiv.org/abs/2310.07771.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. *arXiv preprint arXiv:2203.17270*, 2022. URL https://arxiv.org/abs/2203.17270.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. *arXiv preprint arXiv:2210.02747*, 2022. URL https://arxiv.org/abs/2210.02747.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. *arXiv preprint arXiv:2209.03003*, 2022. URL https://arxiv.org/abs/2209.03003.

Enhui Ma, Lijun Zhou, Tao Tang, Zhan Zhang, Dong Han, Junpeng Jiang, Kun Zhan, Peng Jia, Xianpeng Lang, Haiyang Sun, Di Lin, and Kaicheng Yu. Unleashing Generalization of End-to-End Autonomous Driving with Controllable Long Video Generation. *arXiv preprint arXiv:2406.01349*, 2024. URL https://arxiv.org/abs/2406.01349.

Jianbiao Mei, Tao Hu, Xuemeng Yang, Licheng Wen, Yu Yang, Tiantian Wei, Yukai Ma, Min Dou, Botian Shi, and Yong Liu. DreamForge: Motion-Aware Autoregressive Video Generation for Multi-View Driving Scenes. *arXiv preprint arXiv:2409.04003*, 2024. URL https://arxiv.org/abs/2409.04003.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.08453*, 2023. URL https://arxiv.org/abs/2302.08453.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley. Color Transfer between Images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001. doi: 10.1109/38.946629.

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. URL https://arxiv.org/abs/2503.20523.

Tim Salimans and Jonathan Ho. Progressive Distillation for Fast Sampling of Diffusion Models. *arXiv preprint arXiv:2202.00512*, 2022. URL https://arxiv.org/abs/2202.00512.

Kunpeng Song, Tingbo Hou, Zecheng He, Haoyu Ma, Jialiang Wang, Animesh Sinha, Sam Tsai, Yaqiao Luo, Xiaoliang Dai, Li Chen, Xide Xia, Peizhao Zhang, Peter Vajda, Ahmed Elgammal, and Felix Juefei-Xu. Llama Learns to Direct: DirectorLLM for Human-Centric Video Generation. *arXiv preprint arXiv:2412.14484*, 2024. URL https://arxiv.org/abs/2412.14484.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. OminiControl: Minimal and Universal Control for Diffusion Transformer. *arXiv preprint arXiv:2411.15098*, 2024. URL https://arxiv.org/abs/2411.15098.

Zhenxiong Tan, Qiaochu Xue, Xingyi Yang, Songhua Liu, and Xinchao Wang. OminiControl2: Efficient Conditioning for Diffusion Transformers. *arXiv preprint arXiv:2503.08280*, 2025. URL https://arxiv.org/abs/2503.08280.

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. URL https://arxiv.org/abs/1812.01717.

Haiguang Wang, Daqi Liu, Hongwei Xie, Haisong Liu, Enhui Ma, Kaicheng Yu, Limin Wang, and Bing Wang. MiLA: Multi-view Intensive-fidelity Long-term Video Generation World Model for Autonomous Driving. *arXiv preprint arXiv:2503.15875*, 2025. URL https://arxiv.org/abs/2503.15875.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drive-Dreamer: Towards Real-world-driven World Models for Autonomous Driving. *arXiv preprint arXiv:2309.09777*, 2023a. URL https://arxiv.org/abs/2309.09777.

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. *arXiv preprint arXiv:2311.17918*, 2023b. URL https://arxiv.org/abs/2311.17918.

Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and Controllable Video Generation for Autonomous Driving. *arXiv preprint arXiv:2311.16813*, 2023. URL https://arxiv.org/abs/2311.16813.

Yuqing Wen, Yucheng Zhao, Yingfei Liu, Binyuan Huang, Fan Jia, Yanhui Wang, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea+: Panoramic and Controllable Video Generation for Autonomous Driving. *arXiv preprint arXiv:2408.07605*, 2024. URL https://arxiv.org/abs/2408.07605.

Wei Wu, Xi Guo, Weixuan Tang, Tingxuan Huang, Chiyu Wang, Dongyue Chen, and Chenjing Ding. DriveScape: Towards High-Resolution Controllable Multi-View Driving Video Generation. *arXiv preprint arXiv:2409.05463*, 2024. URL https://arxiv.org/abs/2409.05463.

Zhenya Yang, Zhe Liu, Yuxiang Lu, Liping Hou, Chenxuan Miao, Siyi Peng, Bailan Feng, Xiang Bai, and Hengshuang Zhao. GenieDrive: Towards Physics-Aware Driving World Model with 4D Occupancy Guided Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2512.12751.

Zhuoran Yang and Yanyong Zhang. ConsisDrive: Identity-Preserving Driving World Models for Video Generation by Instance Mask. *arXiv preprint arXiv:2602.03213*, 2026. URL https://arxiv.org/abs/2602.03213.

Yining Yao, Xi Guo, Chenjing Ding, and Wei Wu. MyGo: Consistent and Controllable Multi-View Driving Video Generation with Camera Control. *arXiv preprint arXiv:2409.06189*, 2024. URL https://arxiv.org/abs/2409.06189.

Jie Zhang, Xiaoyue Chen, Anzhe Chen, Dayiheng Liu, Deqing Li, Gengze Zhou, Hale Yin, Haoqi Yuan, Haoyang Li, Jiahao Li, Jiazhao Zhang, Jingren Zhou, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Pei Lin, Qihang Peng, Shengming Yin, Tianhe Wu, Tianyi Yan, Xiao Xu, Yan Shu, Yanran Zhang, Ye Wang, Yi Wang, Yilei Chen, Yixian Xu, Yiyang Huang, Yuxiang Chen, Zekai Zhang, Zhendong Wang, Zixing Lei, Zhixuan Liang, Zihao Liu, Zikai Zhou, Chenxu Lv, Xiong-Hui Chen, and Chenfei Wu. Qwen-RobotWorld Technical Report: Unifying Embodied World Modeling through Language-Conditioned Video Generation. *arXiv preprint arXiv:2606.17030*, 2026. URL https://arxiv.org/abs/2606.17030.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.05543*, 2023. URL https://arxiv.org/abs/2302.05543.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *arXiv preprint arXiv:1801.03924*, 2018. URL https://arxiv.org/abs/1801.03924.

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. *arXiv preprint arXiv:2403.06845*, 2024. URL https://arxiv.org/abs/2403.06845.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. URL https://arxiv.org/abs/2503.21755.

Jiayi Zhu, Jianing Zhang, Yiying Yang, Wei Cheng, and Xiaoyun Yuan. ShareVerse: Multi-Agent Consistent Video Generation for Shared World Modeling. *arXiv preprint arXiv:2603.02697*, 2026. URL https://arxiv.org/abs/2603.02697.