Prompt2Drive: Language-Directed 2D Video Edits for ADAS

Anonymous CVPR submission

Paper ID *****

Author Information

Sagar Chandrashekhar Bellad (sbellad),
Badhrinarayanan K V (badhrik)
Arun Reddy Anugu (aanugu)

Abstract

We study a simple, pragmatic pipeline for editing existing 2D driving videos *from natural language. Users can (i)* ***add pedestrians****, (ii)* ***add traffic objects*** *(e.g., lights with red/yellow/green states; common signs), and (iii)* ***request weather changes*** *(reserved for a companion section). Given a short prompt ("add a pedestrian at ∼20 m" / "insert an overhead 3-aspect traffic light, start red"), subagents infer placement and transforms, diffusion models synthesize the content, and a compositor inserts it with depth-aware ordering. We do* not *claim perfect geometry, shadows, or occlusions; the goal is fast, consistent edits good enough for* simulation-for-perception *data. We keep the design 2D-first to reduce cost and to complement richer 3D systems like ChatSim [15].*

1. Problem Statement

Given a dashcam-like video and a prompt, produce an edited video where inserted content is (a) roughly metric-consistent, (b) temporally stable, and (c) composited with plausible ordering. We focus on three abilities:

1. **Add pedestrians** at user-specified approximate distances/locations.
2. **Add traffic objects**, e.g., lights (with color/state) and common signs.
3. **Change weather** (interface stub left here for companion work).

We explicitly *do not* guarantee perfect lighting, shadows, or geometry; our target is useful, controllable edits for ADAS data augmentation.

2. Related Work

- **ChatSim [15]: Natural Language Editing for Photo-Realistic Driving Simulation.** This work addresses key limitations in editable scene simulation by introducing **ChatSim**, a system that allows for photo-realistic 3D scene editing via **natural language commands**. ChatSim's contributions are highly relevant to our project:
  1. It uses a **Large Language Model (LLM) agent collaboration** framework to achieve high user interaction efficiency and command flexibility.
  2. It employs a novel **multi-camera Neural Radiance Field (NeRF)** method to ensure photo-realistic and visually consistent rendering across all sensor views.
  3. It integrates external digital assets seamlessly using a **multi-camera lighting estimation method** to maintain scene-consistent realism.

  Validated on the Waymo Open Dataset, ChatSim sets a new standard for high-fidelity, user-friendly simulation data generation, which is foundational to our approach.

3. Proposed Method

We propose an instruction-guided diffusion framework that performs controllable and realistic editing of autonomous-driving scenes given natural-language prompts such as *"add more pedestrians"*, *"remove traffic lights"*, or *"change weather to fog"*. Unlike previous text-to-image diffusion approaches, we leverages the **Segment Anything Model (SAM)** [8] to extract rich geometric and semantic priors that keep the generated edits physically consistent.

Our pipeline (Fig. 2) comprises four stages: (1) multimodal scene encoding via SAM and auxiliary estimators, (2) instruction parsing and layout planning, (3) multi-control diffusion editing, and (4) label propagation and synthetic-dataset generation.

3.1. Multimodal Scene Encoding via SAM

Given an image x, we derive conditioning cues that describe scene structure and semantics. **SAM** [8] provides dense instance masks that delineate all visible entities without class supervision. We fuse these masks with

- depth maps from ZoeDepth [1],
- semantic maps from Mask2Former [3], and
- edge maps from Canny/HED filters,

forming a composite conditioning tensor. Multiple ControlNets [16] consume these signals to ensure edits remain faithful to existing geometry and object boundaries.

3.2. Instruction Parsing and Layout Planning

A lightweight parser, inspired by InstructPix2Pix [2] and T2I-Adapter [4], converts the textual prompt p into a structured *EditPlan*:

```
{"operation":"insert","class":"pedestrian",
 "count":3,"region":"sidewalk"}
```

Using SAM and semantic maps, the **Layout Planner** (cf. LayoutDiffusion [9]) samples physically plausible object locations and scales consistent with local depth and perspective, guaranteeing realism in insertions and removals.

3.3. Multi-Control Diffusion Editing

Editing is executed with a latent-diffusion backbone akin to Stable Diffusion [12], augmented with Depth, Semantic, SAM, and Edge ControlNets. We apply null-text inversion [10] to embed x into latent space, followed by Prompt-to-Prompt attention control [6] to confine edits to SAM-derived masks. Weather transformations use a physics-aware residual branch based on the Koschmieder scattering model [14], modulated by depth and lighting cues for realistic rain/fog/snow synthesis. Training employs LoRA adapters [7] and a region-aware perceptual loss that discourages off-target changes.

3.4. Label Propagation and Synthetic-Dataset Generation

After generating the edited image x', labels are automatically updated for downstream tasks. Unchanged objects retain annotations through IoU matching of SAM masks, while inserted or removed entities obtain bounding boxes from the Layout Planner and are refined with a teacher detector such as Mask R-CNN [5]. This self-consistent procedure, similar to SynthDet [11], yields aligned image–label pairs. By varying prompts (*“increase pedestrian density”*, *“nighttime fog”*), out method produces diverse synthetic data that improve perception robustness to weather and density shifts.

3.5. Generative Modeling for Weather Simulation

Current generative models often lack the domain-specific fidelity needed for autonomous driving, failing on multi-agent interactions, fine-grained control, and multi-camera consistency. Our project will develop a novel latent diffusion world model—similar to the approach demonstrated by GAIA-2 [13]to unify these critical capabilities. The model will enable controllable, spatiotemporally consistent video generation conditioned on structured inputs (e.g., dynamics and road semantics). This integration will facilitate the scalable synthesis of both common and rare driving scenarios, significantly advancing the utility of world models as a core development tool.

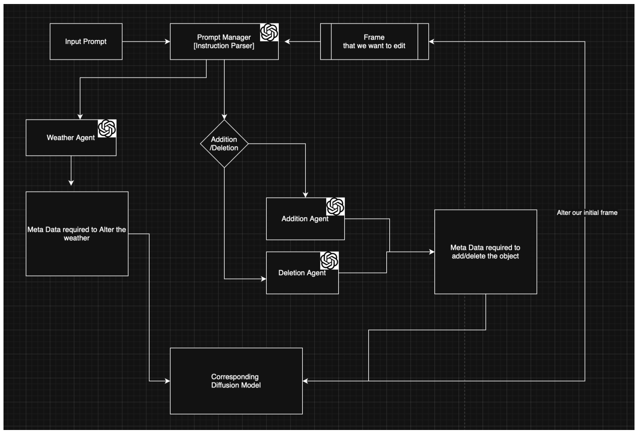

Figure 1. Overall flow.

4. Future Work

(i) Fill in the weather module (physics + diffusion) in this stub. (ii) Scale data generation using segmentation-driven bootstrapping to train multiple diffusion back-ends, covering the common object set in autonomous driving so most prompts can be handled out-of-the-box. (iii) Explore improved temporal coherence without requiring full 3D reconstruction.

References

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Zoedepth: Zero-shot transfer by combining relative and metric depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[2] Tim Brooks and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[3] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Mask2former: Unified architecture for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[4] Changqian Gao, Yixuan Huang, Fangyun Wang, Yunpeng Bai, Yongchao Xu, Xiaoguang Hu, and Shuicheng Lu. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Gir-

shick. Mask r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017. 2

[6] Amir Hertz, Ron Mokady, Jonathan B. Tenenbaum, Ariel Shamir, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[7] Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1

[9] Zhen Li, Zexin Zhang, Ziyu Guo, Yandong Zhang, and Ziwei Liu. Layoutdiffusion: Controllable layout-to-image generation. *arXiv preprint arXiv:2303.17185*, 2023. 2

[10] Ron Mokady, Amir Hertz, Jonathan B. Tenenbaum, Ariel Shamir, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[11] Sergey Nikolenko, Xiaolong Peng, Jonathan Tremblay, Thang To, and Stan Birchfield. Synthdet: Synthetic data generation for object detection. *arXiv preprint arXiv:2006.12073*, 2020. 2

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[13] 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, 2025. 2

[14] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Sir^2: A synthetic-to-real benchmark for image restoration under adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[15] Yuxi Wei, Zi Wang, Yifan Lu, Chenxin Xu, Changxing Liu, Hao Zhao, Siheng Chen, and Yanfeng Wang. Editable scene simulation for autonomous driving via collaborative llm-agents, 2024. 1

[16] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2