DrivePTS: A Progressive Learning Framework with Textual and Structural Enhancement for Driving Scene Generation

Supplementary Material

6. Appendix

6.1. Details on Multi-View Hierarchical Description Generation

In previous works on driving scene generation, text-based conditional inputs often lack the multi-view and fine-grained descriptions necessary for high-fidelity and contextually consistent reconstruction. To overcome this limitation, we leverage the advanced multimodal model Qwen2.5-VL-72B to generate comprehensive descriptions across six distinct semantic aspects for each view. We fine-tune VLM with LoRA [10] on driving-domain datasets (DriveLM [32] and nuscenes-OmniDrive [36]) to reduce hallucinations and enhance scene understanding. To further align generated captions with factual scene attributes, we apply direct preference optimization (DPO) [30] as a reinforcement correction mechanism, guiding the model toward human-consistent outputs. This facilitates more precise control over the generation process and improves realism and diversity in the synthesized scenes. Furthermore, to accommodate the constraints of the CLIP text encoder within the Stable Diffusion UNet framework, the total token count for each scene description is constrained to a maximum of 77 tokens. The specific prompt employed for the VLM is presented in Listing 1.

As shown in Fig. 7, there are some example scenes paired with their multi-view hierarchical descriptions. Beyond the fundamental temporal information, the generated texts effectively convey weather conditions, including overcast, rainy, clear sky, etc. Additionally, it categorizes various road types, such as straight roads, cross-junctions, right-turn lanes, roundabouts, split roads, and scenarios with no road. Furthermore, the descriptions identify the types of objects and infrastructure present in the scene and articulate their spatial relationships.

6.2. More Qualitative Results

Supplementary examples are provided to further demonstrate the superior capabilities of our proposed DrivePTS. The multi-view hierarchical descriptions significantly outperform the short captions from existing datasets regarding scene detail recovery, as illustrated in Fig. 8 and 9. These comprehensive descriptions enable a more nuanced and accurate reconstruction of complex driving scenarios. Additionally, Fig. 10, 11, and 12 demonstrate DrivePTS's ability to align with modified maps through successful road removal and addition across diverse environments. These results highlight the framework's adaptability to dynamic road configurations and demonstrate its potential for generating various road types to validate navigation tasks.

6.3. Generalization to Video Generation

We further evaluate the generalization capability of the proposed strategy and components on the video generation task. Following the design of MagicDrive, we extend DrivePTS with spatio-temporal attention to model temporal dependencies. For quantitative evaluation, we adopt the Fréchet Video Distance (FVD) [35] metric to measure temporal coherence and overall video quality. The results are summarized below.

Table 5. Comparison of temporal consistency for different video generation methods on the NuScenes dataset. Lower FVD scores indicate better temporal coherence.

Metric	MagicDrive	DriveDreamer	Panacea	Drive-WM	Ours (SD-2.1)	**Ours (SD-3.5)**
FVD↓	221	353	139	122	128	**110**

As shown in Tab. 5, our temporal variant still achieves lower FVD scores than most existing baselines. It is worth noting that Drive-WM[38] benefits from additional conditional inputs, including driving actions and control signals, enabling a future-aware world model with stronger temporal coherence. Without introducing video-specific designs, our framework still shows strong adaptability, confirming that the proposed strategy and components can be effectively extended to temporal generation tasks. Furthermore, when equipped with the more advanced SD3.5 [8] backbone following the DiT paradigm, DrivePTS achieves the best FVD score, benefiting from DiT-based spatio-temporal modeling capacity. These results highlight the generalizability and extensibility of DrivePTS.

6.4. Limitation and Future Work

Despite overall improvements, the fidelity of generated lane markings and traffic sign details remains suboptimal, indicating the need for more precise spatial and semantic control. Future work will explore incorporating more advanced fine-grained constraints to further refine local structures and contextual consistency. These enhancements are expected to facilitate the synthesis of more realistic and coherent driving scenes while supporting fine-grained perception tasks such as lane detection, traffic sign recognition, and signal understanding.

```
You are an image captioner specialized in autonomous driving scenes.
You will be provided with multiple images taken by the ego vehicle's
       360-degree surround view cameras, with viewpoints at the "
       CAM_FRONT_LEFT", "CAM_FRONT", "CAM_FRONT_RIGHT", "CAM_BACK_RIGHT
       ", "CAM_BACK", "CAM_BACK_LEFT" of the ego vehicle.

Given multiple images from different camera viewpoints, you must output
        a valid JSON object with exactly the following structure:
{
        "CAM_FRONT_LEFT": {
                "time": <time of day, choose from e.g. "daytime", "
                       night", "evening", "dawn", "dusk">,
                "weather": <concise weather condition, choose from e.g.
                       "clear", "sunny", "overcast", "cloudy", "foggy
                       ", "rainy", "drizzle", "snowy", "hazy", "stormy
                       ">,
                "road type": <brief description of the road surface,
                       choose from "straight road", "split road", "left
                       -turn lane", "right-turn lane", "cross-junction
                       ", "T-junction", "Y-junction", "roundabout", "
                       merging road", "no road">,
                "surroundings": <brief description of the scene type, e
                       .g. "urban intersection", "residential street",
                       "multi-lane highway", "construction zone", "
                       urban park area">,
                "static_and_dynamic_objects": <comma-separated list of
                       visible and relevant static and dynamic entities
                       , e.g. "cars, trucks, buses, bicycles,
                       pedestrians, traffic lights, traffic signs,
                       trees, buildings, cones, barriers">,
                "spatial relationships": <a rich and informative
                       sentence that describes the relationship among
                       the scene layout, interactions, or driving-
                       relevant dynamics. Do not simply repeat values
                       from the previous fields. Focus on spatial
                       relationships, motion patterns, visual semantics
                       , or notable features>},
        "CAM_FRONT": {
                // Same structure as above},
        "CAM_FRONT_RIGHT": {
                // Same structure as above},
        "CAM_BACK_RIGHT": {
                // Same structure as above},
        "CAM_BACK": {
                // Same structure as above},
        "CAM_BACK_LEFT": {
                // Same structure as above}
}

Constraints:
- Note!! Each camera viewpoint's caption (including field names,
       punctuation, and values) must not exceed 110 tokens.
- Output ONLY the JSON object--no additional text.
- The images may be blurred due to rain or movement. If the text cannot
        be read clearly, please do not guess the content of the text
       blindly. So, for textual content, it is important to answer
       accurately.
- For congestion: If there are other cars in front of ego vehicle at a
       close distance and there is a large flow of traffic in the same
       lane next to it, it can be considered congested.
- Common static objects include traffic signs/signals, road
       infrastructure, road markings, road obstacles/barriers, parking
       facilities, traffic monitoring equipment, roadside facilities,
       roadside greenery, advertising/information boards, and public
       facilities.
- Common dynamic objects include motor vehicles, non-motorized vehicles
       , pedestrians, emergency vehicles, construction equipment/
       vehicles, and public transportation vehicles.
- If a camera viewpoint shows no clear road or driving scene, set "road
        type" to "no road" and adjust other fields accordingly.

Example format:
{
        "CAM_FRONT_LEFT": {
                "time": "daytime",
                "weather": "sunny, clear sky",
                "road type": "left-turn lane",
                "surroundings": "urban street scene",
                "static_and_dynamic_objects": "bus, car, fence, trees,
                       building",
                "spatial relationships": "The image shows a street
                       scene with a green fence, trees, and buildings.
                       There's an orange bus on the road, and part of a
                        car is visible in the foreground."},
        "CAM_FRONT": {
                "time": "daytime",
                "weather": "sunny, clear sky",
                "road type": "straight road",
                "surroundings": "urban intersection",
                "static_and_dynamic_objects": "traffic lights, cars,
                       crosswalk, buildings",
                "spatial relationships": "Forward view shows an
                       intersection with traffic lights overhead,
                       vehicles waiting in lanes, and crosswalk
                       markings visible on the road surface."}
        // ... continue for all other camera viewpoints
}
```

Listing 1. Prompt template for VLM-based scene description generation.

Cross roundabout, construction vehicle, ped walking along the road side

"daytime",
"cloudy",
"straight road",
"urban commercial area",
"cars, buildings, shipping containers, pedestrians, road signs",
"A car drives past a row of shipping containers used as storefronts, with a pedestrian walking nearby and a restaurant named 'Yankee Lobster' in the background."

"daytime",
"cloudy",
"roundabout",
"urban industrial area",
"signs, cars, buildings, trees",
"A roundabout with a yield sign and construction warning. Cars navigate around while industrial buildings and trees surround the area under a cloudy sky."

"daytime",
"cloudy",
"straight road",
"construction zone",
"fence, concrete barriers, traffic signs",
"A construction zone with a red mesh fence supported by concrete barriers. Traffic signs indicate road directions under a cloudy sky."

"daytime",
"cloudy",
"no road",
"urban street",
"buildings, fence, barrier",
"A view of a straight urban road with a red mesh fence and concrete barriers, under an overcast sky with buildings in the background."

"daytime",
"cloudy",
"straight road",
"urban construction zone",
"excavator, truck, buildings, barriers",
"A construction site with an excavator and a truck is visible on the left side of a straight urban road lined with tall buildings under an overcast sky."

"daytime",
"cloudy",
"straight road",
"urban street scene",
"cars, pedestrians, buildings, tent structure",
"Cars drive along a straight road past a large tent-like structure with people walking nearby and buildings in the background under an overcast sky."

Nature, bus stop, parked cars, bus exits from the intersection

"daytime",
"overcast",
"no road",
"urban park area",
"trees, stairs, railing, signboard",
"The scene depicts a covered staircase leading up to a walkway surrounded by greenery and urban structures, with a signboard indicating 'Portsdown Road'."

"daytime",
"overcast",
"straight road",
"urban park area",
"trees, bus stop, cars, road markings",
"A quiet urban park road with a bus stop on the left, cars parked on the right, and lush green trees lining both sides of the road."

"daytime",
"overcast",
"straight road",
"suburban area",
"small building, trees, fence, car",
"A small white building with a red roof sits beside a road, surrounded by trees and a fence, with a parked car nearby under an overcast sky."

"daytime",
"overcast",
"straight road",
"urban park area",
"trees, grass, pole, electrical box",
"A quiet urban park scene with lush green trees and grass beside a road lined by a striped curb. A utility pole and electrical box are visible near the edge."

"daytime",
"overcast",
"right-turn lane",
"urban park area",
"trees, buildings, road markings",
"A right-turn lane curves gently through a park-like setting with lush trees and distant high-rise buildings under an overcast sky."

"daytime",
"overcast",
"no road",
"urban park area",
"trees, trash bags, bollard, building structure",
"The scene depicts a park-like area with dense vegetation, a pile of trash bags, a yellow bollard, and a modern building structure in the background."

Cross busy intersection, bendy bus changing lane

"daytime",
"overcast",
"straight road",
"urban street",
"truck, buildings, trees, road markings",
"A truck carrying a shipping container drives along a city street lined with brick buildings and small trees, passing by clear road markings."

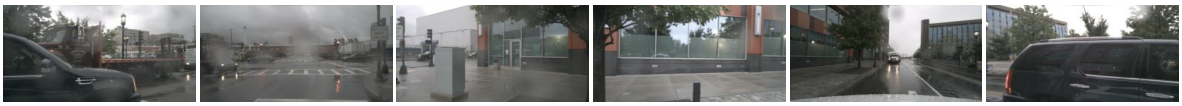
"daytime",
"overcast",
"straight road",
"urban street scene",
"cars, buildings, traffic lights, fences, pedestrian bridge",
"The image captures a city street with cars moving along a straight road under an overcast sky, flanked by tall buildings and a pedestrian bridge overhead."

"daytime",
"overcast",
"cross-junction",
"urban construction site",
"traffic light, warning sign, trucks, buildings, barriers",
"A truck navigates a cross-junction near a construction site with barriers and warning signs. Overcast skies loom over urban buildings in the background."

"daytime",
"overcast",
"cross-junction",
"urban commercial area",
"cars, traffic signs, barricades, buildings",
"Vehicles wait at a cross-junction under overcast skies, with barricades and signs indicating sidewalk closure. Modern buildings frame the scene."

"daytime",
"overcast",
"straight road",
"construction zone",
"cranes, scaffolding, traffic lights, pedestrians",
"The scene captures a wide road leading towards a construction site with cranes and scaffolding under a cloudy sky, with pedestrians crossing nearby."

"daytime",
"overcast",
"cross-junction",
"urban construction zone",
"cars, pedestrian, traffic lights, buildings, scaffolding",
"A pedestrian crosses at a red light while cars wait at the intersection. Construction scaffolding looms in the background, and modern buildings frame the scene."



Rain, parking lot, bicycle rack, buses, jaywalker, parked scooter, arrive at intersection

"daytime",
"rainy",
"straight road",
"urban street",
"trucks, cars, traffic lights, buildings, trees",
"A rainy urban street scene with a flatbed truck carrying a forklift, surrounded by cars and buildings under overcast skies."

"daytime",
"rainy",
"cross-junction",
"industrial urban area",
"cars, trucks, traffic lights, buildings, smokestacks",
"A wet cross-junction with cars and trucks navigating through. Smokestacks emit steam into the overcast sky, and industrial buildings line the streets."

"daytime",
"rainy",
"no road",
"urban street",
"traffic lights, cars, buildings, trees",
"A wet urban street scene with cars parked along the side, traffic lights, and a building with reflective windows showing reflections of the surroundings."

"daytime",
"rainy",
"no road",
"urban sidewalk",
"trees, building, bollards",
"A quiet urban sidewalk scene with a tree partially framing the view, a modern building with large windows reflecting the surroundings, and a bollard in the foreground."

"daytime",
"rainy",
"straight road",
"urban street",
"trees, buildings, bus, streetlights",
"A rainy urban street with wet pavement reflecting light. A bus drives ahead, flanked by trees and modern buildings under a gray sky."

"daytime",
"rainy",
"straight road",
"urban street",
"SUV, trees, buildings, bicycle",
"A black SUV is parked on a wet urban street with trees and modern buildings in the background; a bicycle is visible further down the road."

Wait at intersection, heavy truck, personal mobility

"daytime",
"clear",
"no road",
"industrial area",
"truck, trailer",
"A truck with a trailer displaying 'Available Liftgate Service' is parked in an industrial setting, with no visible road or other vehicles nearby."

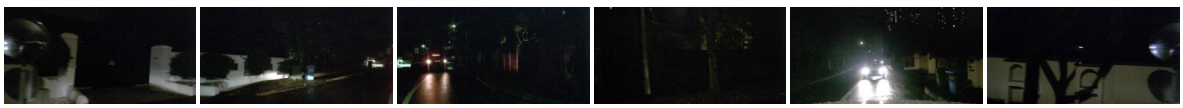
"daytime",
"clear",
"cross-junction",
"urban intersection",
"cars, truck, traffic lights, buildings, cranes",
"A truck and several cars wait at a red light in an urban setting with tall buildings and construction cranes in the background."

"daytime",
"clear",
"split road",
"urban area under construction",
"traffic lights, buildings, cranes, road signs",
"A split road in an urban setting with a large building under construction, cranes in the background, and traffic lights showing red signals."

"daytime",
"clear",
"straight road",
"urban industrial area",
"cars, buildings, fences, streetlights",
"A silver SUV drives past a white car on a straight road near industrial buildings with fences and streetlights under a clear sky."

"daytime",
"clear",
"straight road",
"urban street",
"cars, truck, buildings, streetlights",
"Vehicles are moving along a multi-lane urban road with buildings on the right and a clear view of the road ahead under the sky."

"daytime",
"clear",
"no road",
"urban street",
"truck, reflectors, road",
"A close-up view of a truck's side panel with three diamond-shaped reflectors and red reflective strips along the bottom edge."



Night, residential, car turning

"night",
"clear",
"no road",
"residential area",
"trees, buildings, street lighting",
"Dark nighttime scene with minimal visibility showing residential buildings and vegetation along roadside with sparse illumination."

"night",
"clear",
"straight road",
"residential area",
"trees, wall, trash bins, street light",
"A quiet residential street at night with a white wall, trees, and illuminated trash bins along the side, leading towards a distant street light."

"night",
"clear",
"straight road",
"residential street",
"trees, streetlights, car headlights",
"A dimly lit residential street at night with wet pavement reflecting car headlights and sparse streetlights illuminating the path ahead."

"night",
"dark",
"no road",
"forest area",
"trees, darkness",
"The image captures a dark forest scene with dense trees barely visible under night conditions, creating a mysterious atmosphere."

"night",
"clear",
"straight road",
"rural area",
"cars, road markings, buildings",
"A nighttime rural scene with cars approaching head-on, their headlights illuminating the dark road. Buildings line the right side, barely visible in the darkness."

"night",
"clear",
"no road",
"rural area",
"wall, graffiti, light fixture",
"The image shows an indoor wall with graffiti under dim lighting. A light fixture casts shadows, creating a dark and moody atmosphere."

Figure 7. Comparison of original dataset captions and VLM-driven multi-view hierarchical descriptions. Each group shows: (top) six-view images, (middle) original dataset captions, and (bottom) our multi-view hierarchical descriptions.

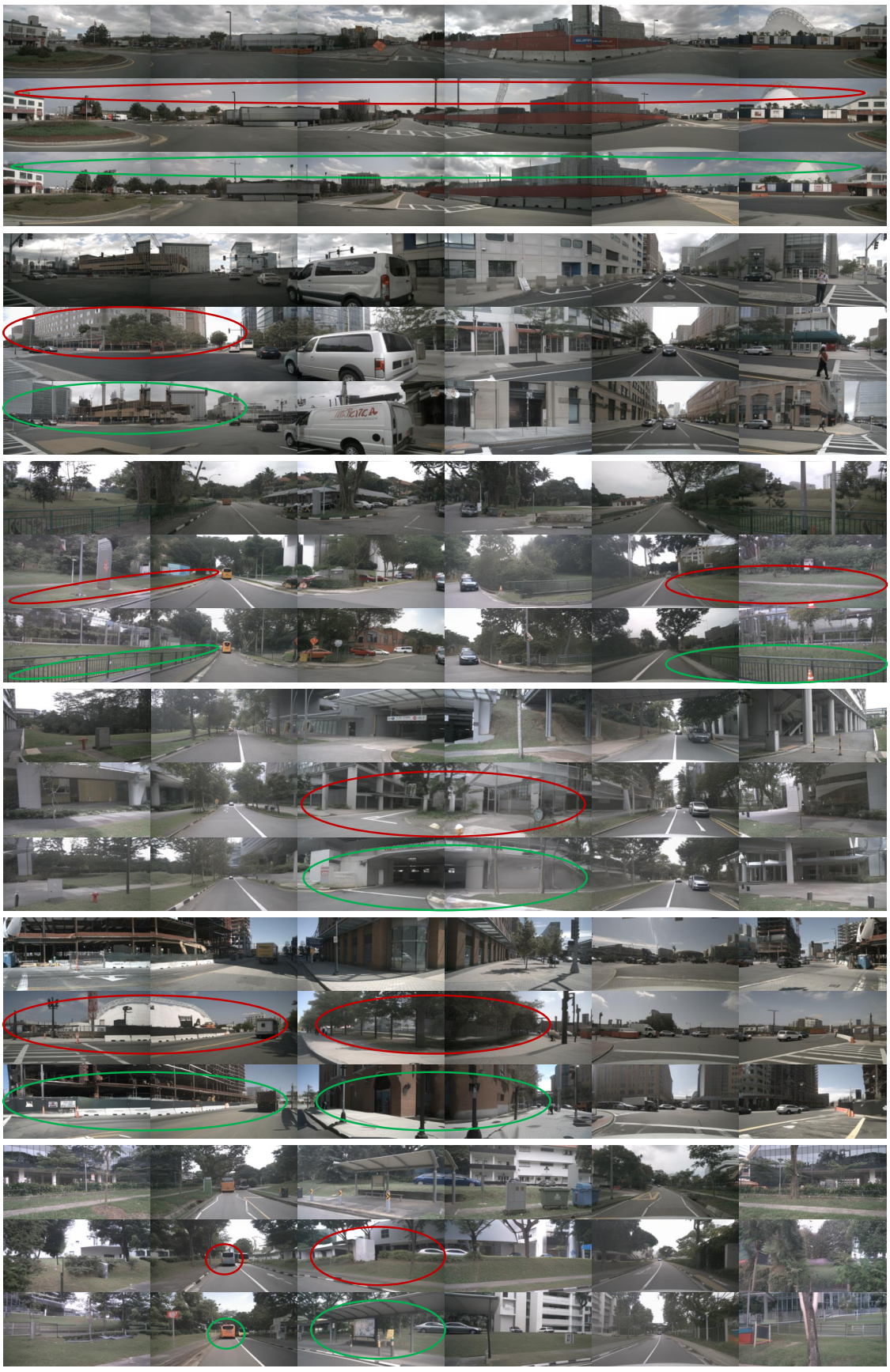


Figure 8. Qualitative comparison of scene reconstruction quality between original dataset captions and our multi-view hierarchical descriptions. For each example: original image (top), reconstruction from original captions (middle), and our results (bottom). Red regions indicate missing details with original captions, while green regions highlight successful recovery through our fine-grained descriptions.

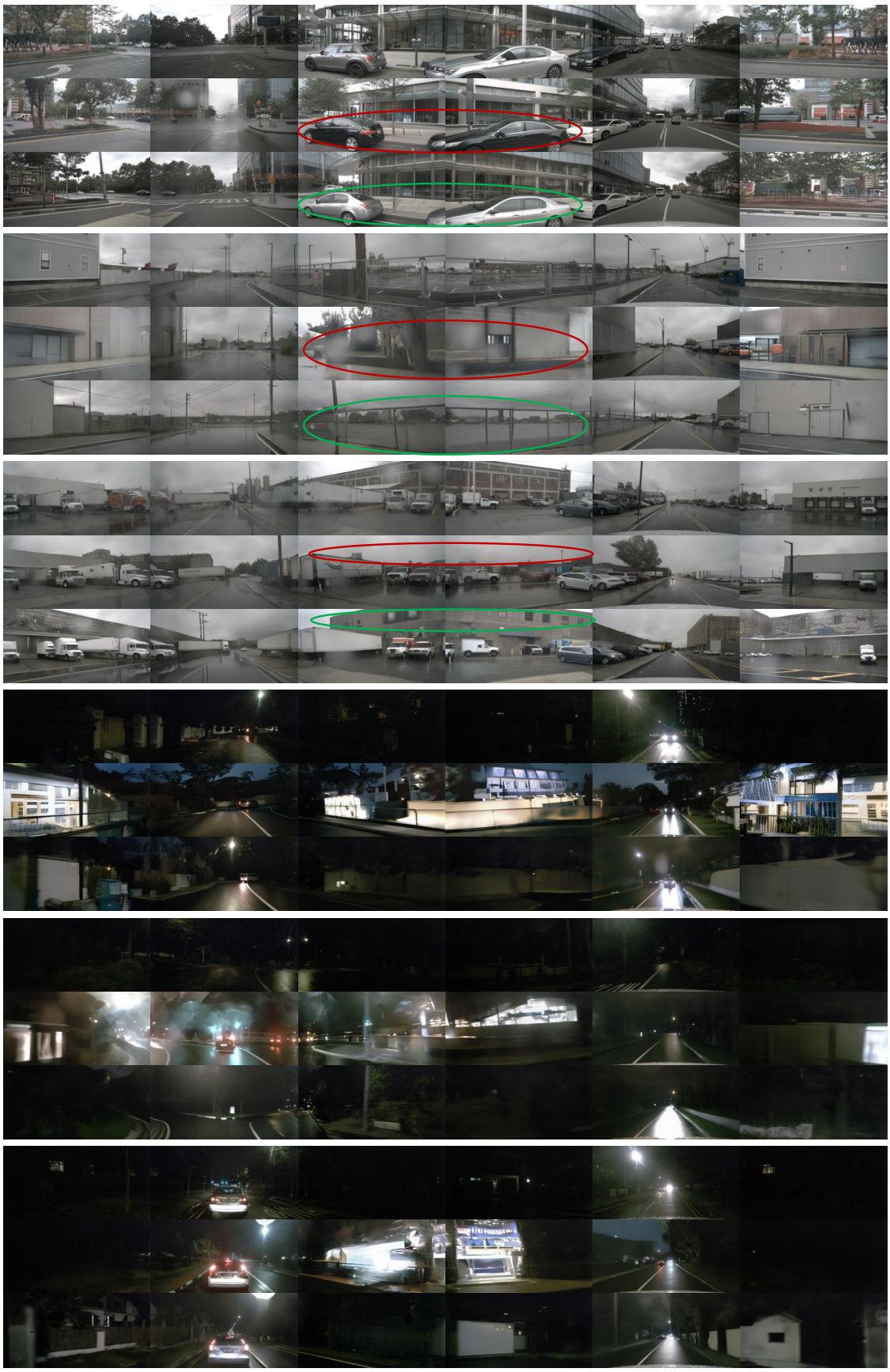


Figure 9. Qualitative comparison of scene reconstruction quality between original dataset captions and our multi-view hierarchical descriptions. Notably, in night scene generation, original captions tend to produce hallucinated illuminated areas due to insufficient contextual information, while our multi-view hierarchical descriptions faithfully reconstruct authentic nighttime atmospheres.

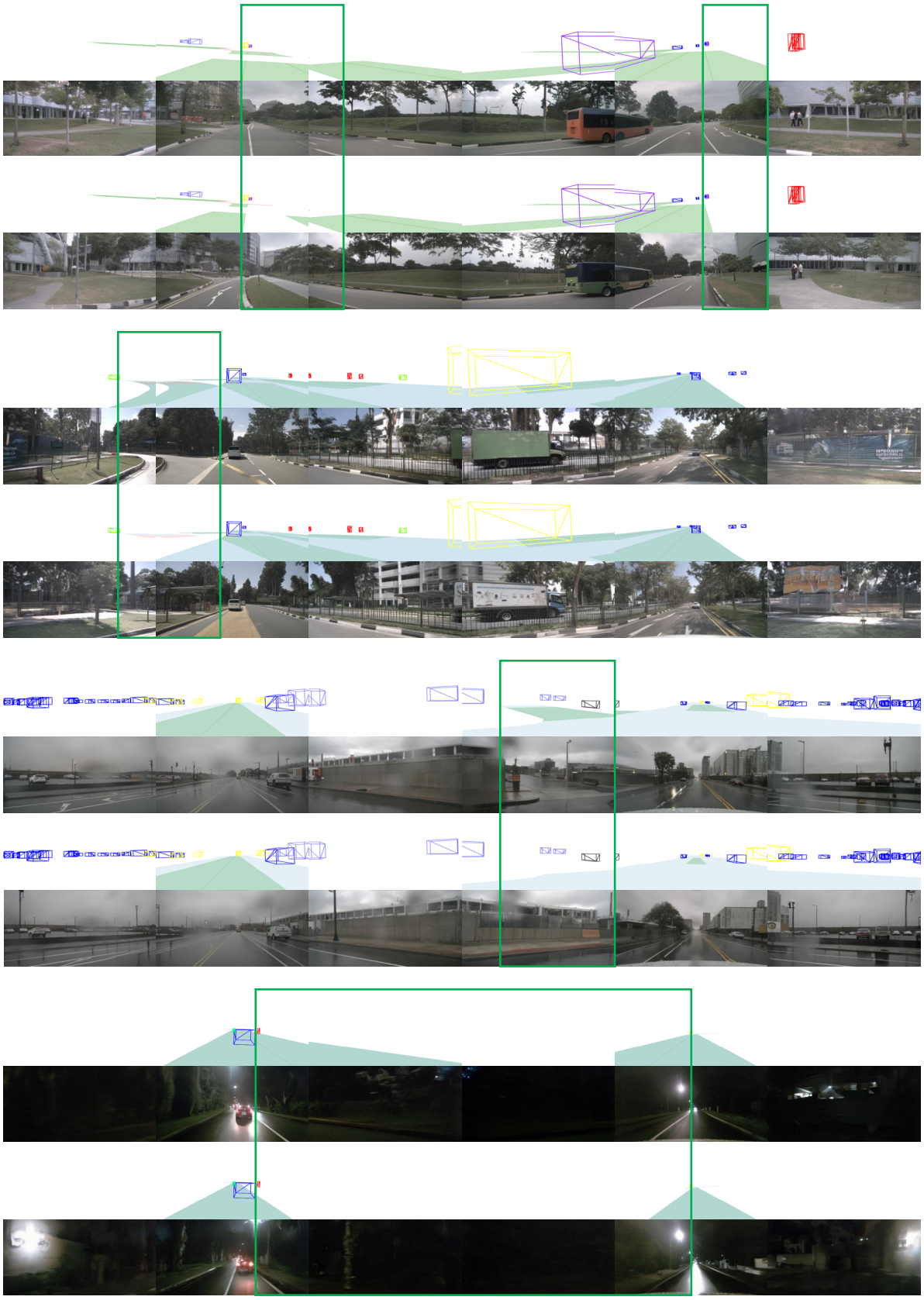


Figure 10. Examples of targeted road removal in driving scene generation using our DrivePTS framework. Each group shows: (1) original geometric conditions, (2) original scene generation, (3) modified geometric conditions after road removal, and (4) updated scene generation. Green boxes indicate areas corresponding to the geometric modifications.

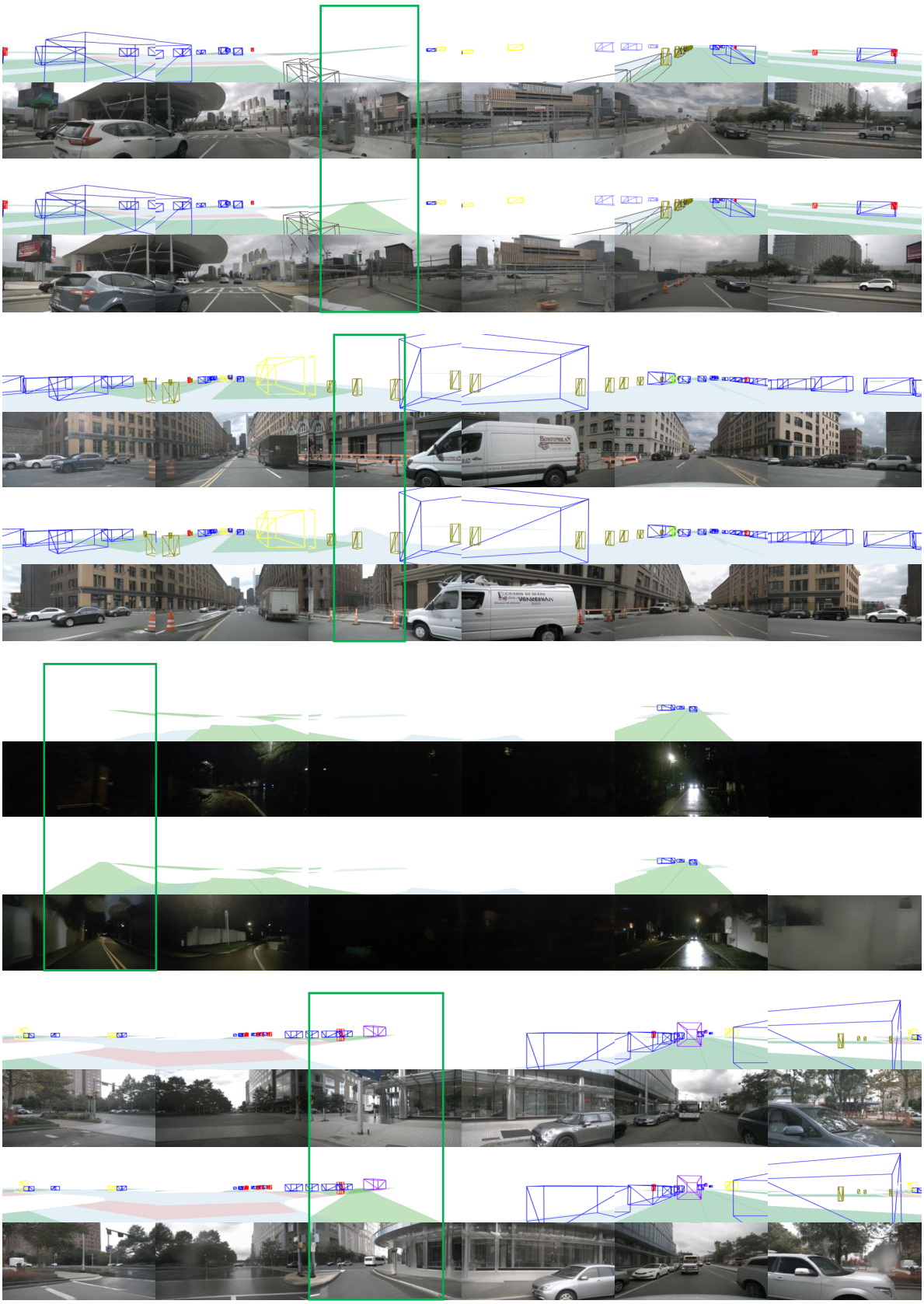


Figure 11. Examples of targeted road addition in driving scene generation using our DrivePTS framework. The visualization follows the same four-row format as above, where the third row shows geometric conditions with added roads.

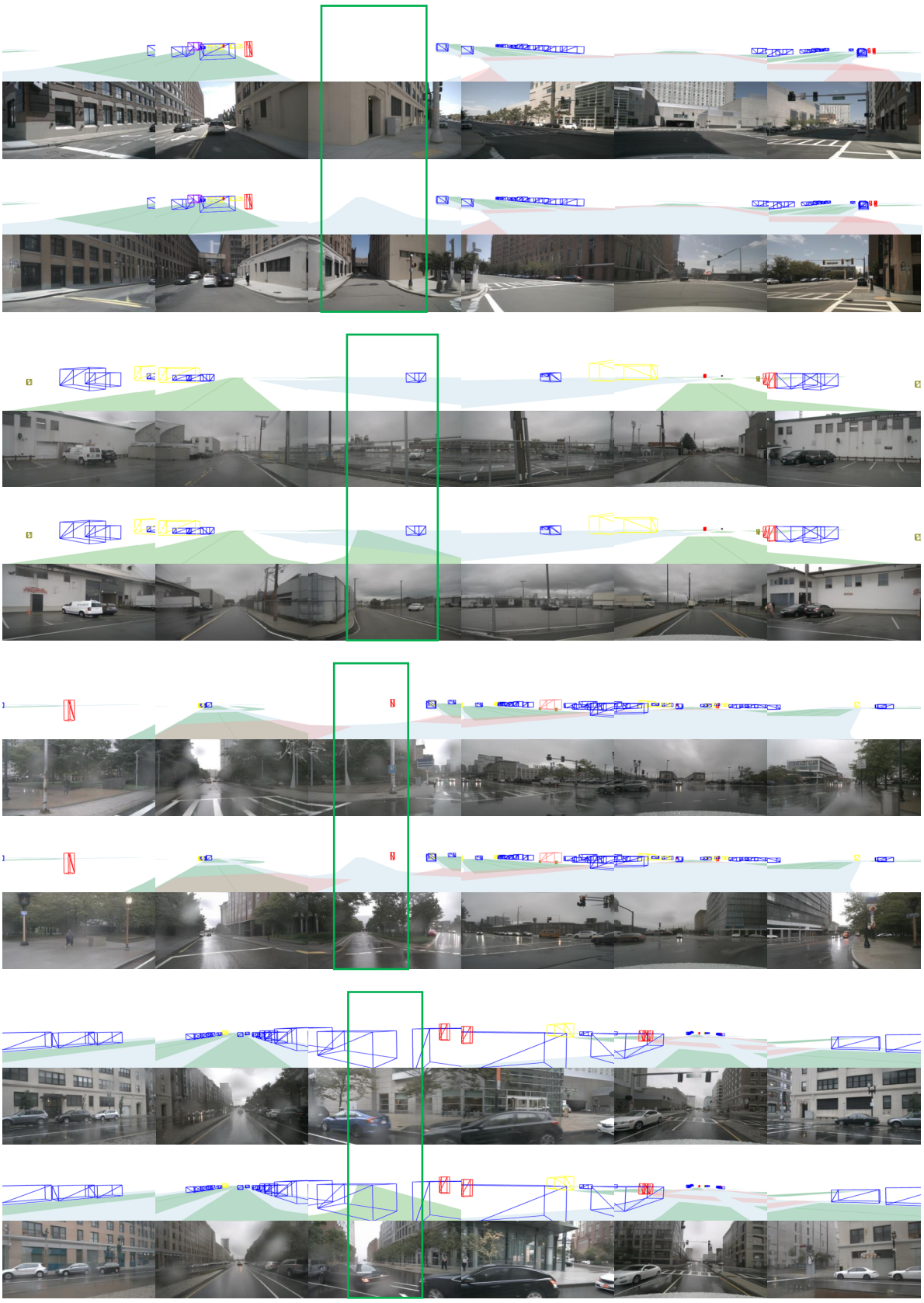


Figure 12. Additional examples of targeted road addition using our DrivePTS framework. The visualization format follows the same structure as above.