Supplementary Material for SIGMA-Lane

1 Experimental Details

1.1 Implementation Details

We adopt OMR's encoder-decoder architecture [4], using a ResNet18 [3] backbone and an eigenlane detection head with $M = 6$ basis vectors. Feature maps are extracted at resolution H=96, W=160 with C=128 channels. For temporal aggregation, we instantiate Mamba2 [2] with state dimension d_{state}=128, convolutional kernel d_{conv}=4, expansion factor expand=2, and a temporal window of T=8 frames. The Scale-Pyramid low-resolution branch uses a spatial downsampling factor s=2, with learnable fusion weight α_{low} initialized to 0.1. Obstacle masks are pseudo-labels from a frozen SegFormer-B5 [10], following OMR for fair comparison. The segmenter remains frozen throughout training. We keep the pretrained encoder-decoder fixed and train only the temporal aggregation modules with the following loss:

$$\mathcal{L} = \ell_{\text{cls}}(P, \bar{P}) + \ell_{\text{reg}}(C, \bar{C}), \tag{1}$$

where P denotes the lane probability map, C the eigenlane coefficient map, and $\bar{\cdot}$ their ground-truth counterparts. ℓ_{cls} is the focal loss [6] and ℓ_{reg} is the LIoU loss [12]. We use KINS [9] synthetic occlusion for data augmentation, the AdamW optimizer [7] with an initial learning rate of 10^{-4}, and a batch size of 4 for 20 epochs on a single NVIDIA RTX 4090 GPU. Input images are resized to 384×640.

1.2 Datasets

We evaluate on two complementary video lane benchmarks that emphasize different aspects of temporal modeling.

VIL-100 [11] provides 100 short clips annotating up to 6 lanes per frame. A major challenge is persistent large-vehicle occlusion, where trucks and buses obscure lane markings for multiple consecutive frames. This setting provides a direct test of state contamination in recurrent temporal models.

OpenLane-V [5] is a large-scale benchmark derived from OpenLane [1], spanning ∼90K frames across 590 videos with up to 4 annotated lanes following CULane conventions [8]. Beyond occlusion, it includes harsh weather, low illumination, and unmarked intersections, which test whether the temporal model remains stable when visual cues are weak or absent.

1.3 Evaluation Metrics

Per-frame quality. Following the CULane protocol [8], each lane is modeled as a 30-pixel wide strip. A prediction is deemed a true positive if its IoU with the ground truth exceeds a threshold τ. We evaluate at τ=0.5 and the stricter τ=0.8, reporting F1@0.5 and F1@0.8 respectively, along with Precision, Recall, mIoU, and pixel-level Accuracy:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2}$$

mIoU averages the IoU scores of correctly detected lanes. Accuracy measures overall pixel-level classification:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{3}$$

Temporal consistency. We adopt the video metrics from [5]. For each ground-truth lane present in two consecutive frames, the detection outcome falls into one of three categories: *Stable* (detected in both), *Flickering* (detected in one only), or *Missing* (missed in both). Let N denote the total number of such adjacent-frame pairs. The flickering rate and missing rate are:

$$R_F = \frac{N_F}{N}, \quad R_M = \frac{N_M}{N}. \tag{4}$$

2 Controlled Contamination

To test whether corrupted observations persist through state propagation, we inject controlled perturbations into the video memory during inference. We corrupt the stored current-frame feature F_t along occlusion boundaries before it is consumed by the temporal aggregation module of subsequent frames. This creates a controlled "contaminated memory" setting without changing the training pipeline.

Protocol. We evaluate four cases under the same checkpoint and validation pipeline: clean and perturbed SIGMA-Lane, and clean and perturbed ungated control. The ungated control disables the input and output gating paths at inference time while leaving the rest of the network unchanged, isolating the role of gated state writing. The perturbation is applied only within the same short interval for all perturbed runs, while evaluation still covers the full validation set. We extract the occlusion boundary band by dilating and eroding the obstacle mask ($M_t > 0.5$) with a 5×5 kernel and taking their difference. We then inject additive noise with amplitude $\alpha = 0.3$ into F_t within this boundary band for frames 200–202 in the evaluation stream. This boundary band is where the backbone's receptive field mixes occluded and non-occluded pixels, producing ambiguous features for the gating module.

Table 1: Controlled memory contamination analysis on OpenLane-V validation.

Setting	F1@0.5(↑)	mIoU(↑)	R_M@0.5(↓)
SIGMA-Lane (clean)	0.838	0.761	0.162
SIGMA-Lane (perturbed)	0.836	0.752	0.182
Ungated control (clean)	0.820	0.746	0.236
Ungated control (perturbed)	0.811	0.726	0.367

Observed behavior. Table 1 reports both the full SIGMA-Lane model and its ungated control. For the full model, structured perturbation causes a modest degradation: F1@0.5 drops from 0.838 to 0.836 ($\Delta = -0.002$), mIoU from 0.761 to 0.752 ($\Delta = -0.009$), and the persistent missing rate R_M@0.5 increases from 0.162 to 0.182 ($\Delta = +0.020$). In the ungated control, the same perturbation produces larger shifts: F1@0.5 drops from 0.820 to 0.811 ($\Delta = -0.009$), mIoU from 0.746 to 0.726 ($\Delta = -0.020$), and R_M@0.5 rises from 0.236 to 0.367 ($\Delta = +0.131$). Across all three metrics, the ungated variant degrades more than the gated model under identical perturbation, with degradation magnitudes 4.5× larger for F1@0.5, 2.2× for mIoU, and 6.6× for R_M@0.5. This trend is consistent with the write-path analysis in Sec. 3.2 of the main paper: the mask-modulated write coefficient limits how much contamination accumulates in the propagated state. The largest contrast appears in R_M, the temporal stability metric.

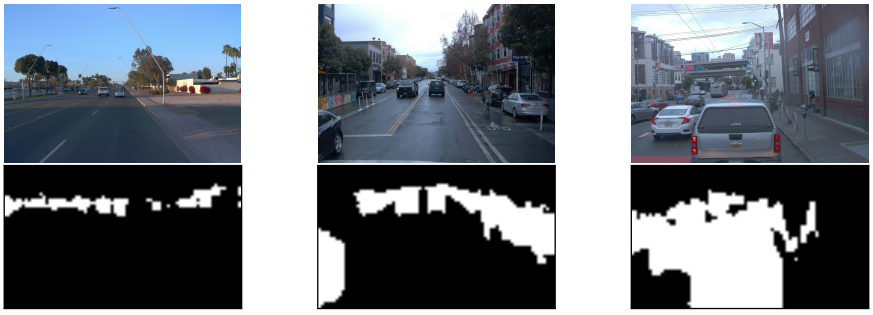
3 Occlusion-Stratified Evaluation

To test whether SIGMA-Lane has its largest gains under heavy occlusion, we stratify frames on the OpenLane-V validation set by occlusion severity using obstacle-mask coverage. All numbers are obtained from the same trained checkpoint used in the main paper without any retraining.

Binning design. For each frame we compute the occlusion ratio ρ, *i.e.* the fraction of spatial positions whose obstacle-mask probability exceeds 0.5:

$$\rho = \frac{\|\mathbf{1}[M_t > 0.5]\|_1}{H \times W}. \tag{5}$$

Frames are then assigned to one of three bins: *Light* ($\rho < 10\%$), *Moderate* ($10\% \leq \rho < 30\%$), and *Heavy* ($\rho \geq 30\%$). The thresholds are chosen so that each bin captures a qualitatively distinct driving condition: Light frames contain little or no foreground vehicles, Moderate frames correspond to typical urban traffic where one or two vehicles partially overlap with the lane region, and Heavy frames represent scenarios where large vehicles substantially occlude the forward view (see Fig. 1 for illustrative examples). For fairness, SIGMA-Lane and OMR [4] are compared under the same frame partition induced by SIGMA-Lane's per-frame ρ values.



(a) Light ($\rho < 10\%$) (b) Moderate ($10\% \leq \rho < 30\%$) (c) Heavy ($\rho \geq 30\%$)

Fig. 1: Examples of the occlusion levels defined by obstacle-mask coverage ρ. The top row shows the original images for light, moderate, and heavy occlusion. The bottom row shows the corresponding obstacle masks.

Motivation. This stratification follows the write-path analysis in Sec. 3.2 of the main paper: as occlusion coverage ρ increases, more spatial positions may carry corrupted features, and a longer or wider occlusion segment can inject more contamination into the recurrent state. If dual-gating suppresses contaminated writes as intended, its benefit should be most visible in frames where occlusion is moderate or heavy.

Results. Table 2 reports mIoU and F1@0.5 on the full OpenLane-V validation split. In the Light bin, OMR and SIGMA-Lane perform comparably, with OMR slightly higher on both metrics. In the Moderate and Heavy bins, SIGMA-Lane is consistently better than OMR. This shift indicates that dual-gating is most useful when state contamination is more likely to affect the recurrent memory.

Table 2: Occlusion-stratified analysis on OpenLane-V validation. Frames are grouped by obstacle mask coverage ρ computed per frame.

		SIGMA-Lane		OMR	
Bin	#Frames	mIoU	F1@0.5	mIoU	F1@0.5
Light ($\rho < 10\%$)	6,507	0.838	0.934	0.844	0.941
Moderate ($10\% \leq \rho < 30\%$)	13,307	0.741	0.811	0.713	0.805
Heavy ($\rho \geq 30\%$)	3,207	0.686	0.756	0.658	0.751
Total	23,021	0.761	0.838	0.742	0.836

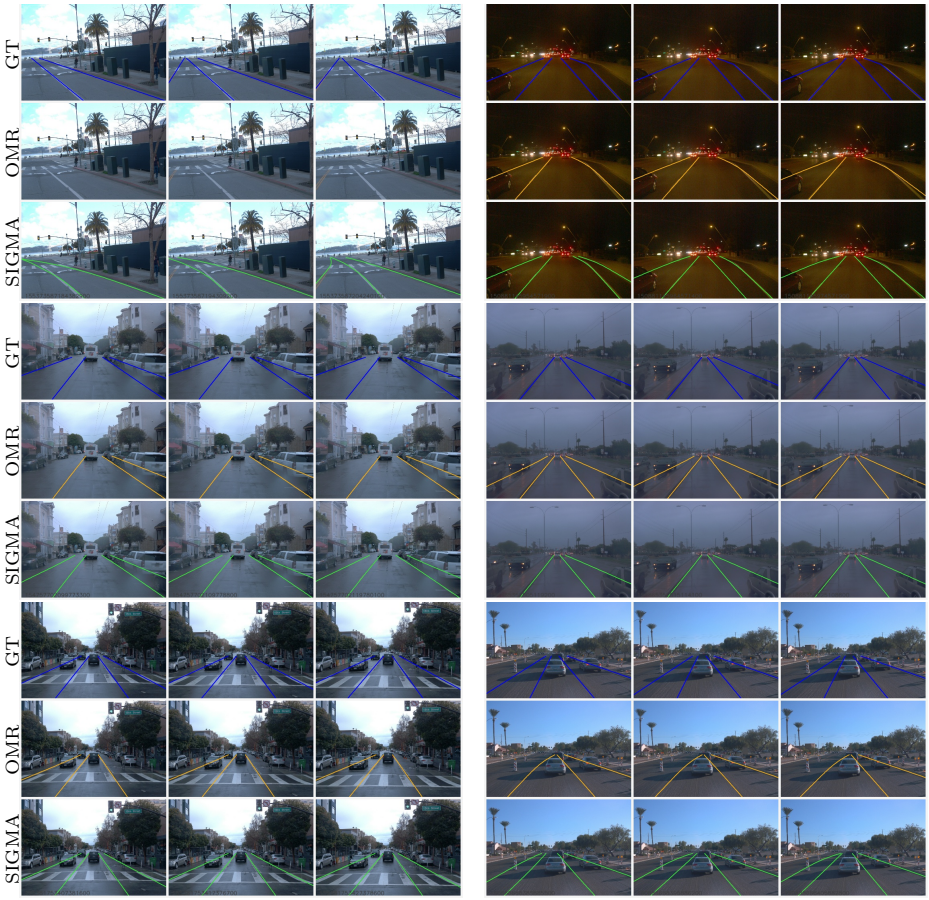


Fig. 2: Representative temporal consistency examples on the matched OpenLane-V clip set. Each block shows three adjacent frames, with ground truth, OMR, and SIGMA-Lane predictions arranged from top to bottom.

4 Temporal Consistency

To visualize temporal consistency under dynamic traffic, we select representative clips from the matched OpenLane-V evaluation set, where SIGMA-Lane and OMR [4] are compared under the same scenarios.

As shown in Fig. 2, we present six cases using three-frame filmstrips to show short-term recovery under transient occlusions, abrupt boundary reveals, and appearance changes. Lane predictions are displayed to focus the comparison on geometric continuity. Across these selected examples, SIGMA-Lane better preserves smooth lane trajectories and partially occluded structures.

In contrast, OMR shows partially missed detections with shortened extents, complete failures on some lanes, and occasional extra false positives under these

conditions. The qualitative cases mirror the lower missing rates reported in the main paper: SIGMA-Lane tends to preserve lane topology when visible lane evidence changes abruptly.

References

1. Chen, L., Sima, C., Li, Y., Zheng, Z., Xu, J., Geng, X., Li, H., He, C., Shi, J., Qiao, Y., Yan, J.: PersFormer: 3d lane detection via perspective transformer and the OpenLane benchmark. In: ECCV. pp. 550–567 (2022)
2. Dao, T., Gu, A.: Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In: ICML. pp. 10041–10071 (2024)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
4. Jin, D., Kim, C.S.: OMR: Occlusion-aware memory-based refinement for video lane detection. In: ECCV. pp. 129–145 (2024)
5. Jin, D., Kim, D., Kim, C.S.: Recursive video lane detection. In: ICCV. pp. 8473–8482 (2023)
6. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017)
7. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
8. Pan, X., Zhan, X., Shi, J., Luo, P., Wang, X., Tang, X.: Spatial as deep: Spatial CNN for traffic scene understanding. In: AAAI. pp. 7276–7283 (2018)
9. Qi, L., Jiang, L., Liu, S., Shen, X., Jia, J.: Amodal instance segmentation with KINS dataset. In: CVPR. pp. 3014–3023 (2019)
10. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and efficient design for semantic segmentation with transformers. In: NeurIPS. pp. 12077–12090 (2021)
11. Zhang, Y., Zhu, L., Feng, W., Fu, H., Wang, M., Li, Q., Li, C., Wang, S.: VIL-100: A new dataset and a baseline model for video instance lane detection. In: ICCV. pp. 15681–15690 (2021)
12. Zheng, T., Huang, Y., Liu, Y., Tang, W., Yang, Z., Cai, D., He, X.: CLRNet: Cross layer refinement network for lane detection. In: CVPR. pp. 898–907 (2022)