Multimodal Coarse-to-Local Transformer for End-to-End Autonomous Driving (Student Abstract)

Yeryeong Cho[1], **Joongheon Kim**[1]
[1]Korea University
{joyena0909, joongheon}@korea.ac.kr

Abstract

End-to-end (E2E) autonomous driving must maintain global consistency while preserving local precision. However, existing E2E approaches rarely achieve both goals simultaneously. Therefore, we propose a multimodal coarse-to-local transformer (MC2L-Transformer), which is composed of a hierarchical transformer architecture. Multimodal inputs are fused into a shared embedding, and global waypoints are produced. Local refinement is then utilized to capture fine interactions around the vehicle. Furthermore, a temporal encoder summarizes recent context, and navigation target and velocity are embedded to guide route- and speed-aware decoding. We evaluate in CARLA, and the results show lower collision and off-route rates even under sudden events. These results indicate that combining a coarse-to-local hierarchical transformer with a lightweight temporal context provides a practical step toward reliable E2E autonomous driving.

Code — https://github.com/CYeryeong/Multimodal-Coarse-to-Local-Transformer-for-Reliable-End-to-End-Autonomous-Driving.git

Introduction

End-to-end (E2E) autonomous driving connects multi-sensor inputs directly to control signals. This design simplifies the pipeline and can reduce error propagation. However, autonomous driving in the real world requires both global scene consistency and precise local interaction around the ego vehicle. Recent papers explore hierarchical transformers in perception or in planning, produced in a single stage (Chitta et al. 2023). Therefore, it entangles global intent with local refinement and weakens closed-loop behavior. These limitations make it hard to balance collision avoidance with off-route prevention, and smooth steering with fast reactions. Therefore, we introduce a hierarchical transformer–based E2E driving model that separates global and local objectives. Multimodal data is fused by an encoder into a shared embedding for trajectory generation. A global stage first outputs coarse waypoints in parallel. A local stage then applies residual refinement to capture near-field interactions around the ego vehicle. Along the temporal axis, a short K-frame memory summarizes recent context and supplies keys and values to both stages. Therefore, the temporal module and the trajectory generator are optimized independently. This yields stable training and efficient computation while clarifying the roles of global and local components. We evaluate our algorithm in CARLA, which is the most similar simulator to the real-world environment. It demonstrates reduced collisions and off-route errors and shows practical real-time potential.

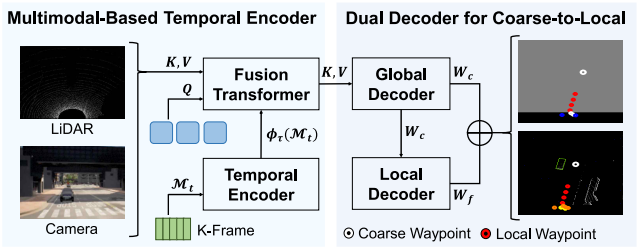


Figure 1: Overview of the proposed MC2L-Transformer.

Multimodal Coars-to-Local Transformer

Our model couples multimodal sensor fusion with a hierarchical transformer for trajectory generation in E2E autonomous driving. Camera and LiDAR data are fused into a shared embedding, a short temporal module supplies recent context, and two decoders generate global-to-local waypoints in a single forward pass, which can avoid heavy autoregression while preserving near-field detail.

Multimodal-Based Temporal Encoder. A fusion Transformer encoder ϕ_{fuse} yields the shared embedding $z_t = \phi_{\text{fuse}}(I_t, L_t)$, and I_t, L_t are camera and LiDAR bird's-eye view (BEV) features, respectively. In practice, this stage aligns features across views and rates, so the downstream trajectory head does not need to relearn cross-sensor consistency. K-frame memory $\mathcal{M}_t = \{z_{t-K+1}, \ldots, z_t\}$ is then summarized by a lightweight temporal encoder ϕ_τ to produce $c_t = \phi_\tau(\mathcal{M}_t)$. It provides short-horizon dynamics. The navigation target and speed of the vehicle supply task conditions. $e_g = \phi_g(g_t)$ encodes the local route goal, and $e_v = \phi_v(v_t)$ encodes desired pace and braking margin. These vectors are concatenated with the current and temporal features to form the keys/values $K_t = [z_t; c_t; e_g; e_v]$ with $V_t = K_t$ for both decoding stages, ensuring that attention

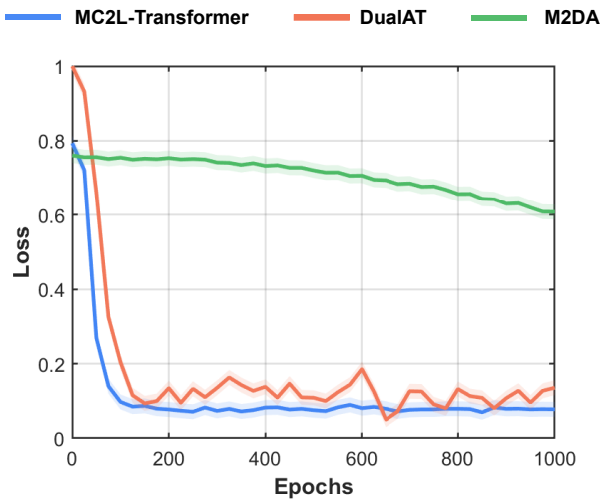


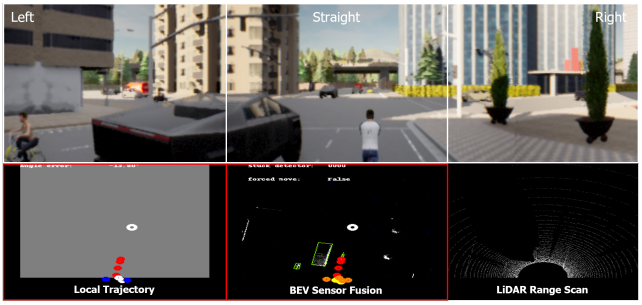
Figure 2: Final L_2 loss between the original and reconstructed image, varying attack method and batch size.

can directly "query" route and speed alongside scene context. To stabilize optimization and avoid long-horizon backpropagation, $\tilde{c}_t = \mathrm{sg}(\phi_\tau(\mathcal{M}_t))$, and K_t is built with $\tilde{c}_t$. This decouples the temporal summarizer from the trajectory generator, improves batch efficiency.

Dual Decoder for Coarse-to-Local. Two sets of learned queries drive parallel T-step prediction. The *global* decoder attends to K_t, V_t and outputs coarse waypoints $W_c \in \mathbb{R}^{T\times 2}$. It represents route-consistent positions over the next horizon without committing to fine collision-avoidance maneuvers. This non-autoregressive output provides a stable scaffold that captures map alignment, heading, and curvature trends at a modest compute cost. The *local* decoder then refines W_c by attending to an augmented context that includes an encoding of W_c; it predicts a residual ΔW and produces the final trajectory $W_f = W_c + \Delta W$. Conditioning on W_c focuses attention on regions near the proposed path to enable the model to make precise lateral adjustments and speed modulation for resolving lane changes, cut-ins, and pedestrian crossings. Therefore, training minimizes the waypoint loss $\mathcal{L}_{\text{wp}} = \frac{1}{T}\sum_{t=1}^{T} \|W_f(t) - W^\star(t)\|_1$ with an optional auxiliary loss on the coarse stage, $\mathcal{L} = \mathcal{L}_{\text{wp}} + \lambda_c \mathcal{L}_{\text{coarse}} + \lambda_s \mathcal{R}_{\text{smooth}}$. $\mathcal{L}_{\text{coarse}}$ mirrors $\mathcal{L}_{\text{wp}}$ on W_c and $\mathcal{R}_{\text{smooth}}$ penalizes curvature spikes to encourage steerable, low-jerk trajectories.

Performance Evaluation

We evaluate the proposed MC2L-Transformer in the CARLA simulator in comparison to DualAT (Chen et al. 2024) and M2DA (Xu et al. 2024), using the public datasets (Chitta et al. 2023). The navigation target and ego speed are embedded into the attention keys and values, and a K-frame temporal memory provides recent context to both decoding stages. To probe robustness to environmental changes, we include day and night scenarios and visualize BEV diagnostics with coarse and refined waypoints, as shown in Fig. 3.



(a) Recognition and trajectory planning during daytime.

(b) Recognition and trajectory planning during night time.

Figure 3: Coarse waypoints (white) and refined path (red) demonstrate day and night consistency and illumination robustness of the coarse-to-local model.

Training Loss and Inference Results. Fig. 2 shows the total loss trajectory during training. DualAT exhibits longer plateaus, and M2DA shows late epoch oscillations. Compared with DualAT and M2DA, our model converges faster and to a lower validation loss. The stop-gradient temporal path further reduces variance and yields a smoother loss curve and earlier. At test time, the model first predicts coarse waypoints with a global decoder and then applies residual refinement to obtain the final trajectory. Fig. 3 overlays coarse (white) and refined (red) waypoints together with detections. It demonstrates that the refined path remains route-consistent while handling near-field interactions under both day and night scenes. Consequently, a non-autoregressive global predictor with residual local refinement lowers error accumulation, and decoupled optimization of the temporal memory improves convergence stability.

Concluding Remarks

This paper proposes MC2L-Transformer, a hierarchical transformer that fuses multimodal inputs, predicts coarse waypoints, and applies residual local refinement with a lightweight temporal module and task-aware conditioning. It yields lower loss compared to other benchmarks and remains consistent across day and night. Furthermore, this paper suggests a practical step toward reliable E2E driving with a realistic simulation in CARLA. In particular, MC2L-Transformer demonstrates that interpretable intermediate structure and efficient temporal reasoning. This paper indicates that scalable and hierarchical planners can be integrated into real-world autonomous driving, while preserving robustness to environmental variation.

Acknowledgments

Funds. This research was supported by the MSIT (Ministry of Science and ICT), Korea by IITP (Institute for Information & Communications Technology Planning & Evaluation) (2022-0-00907, Development of AI Bots Collaboration Platform and Self-organizing AI).

Corresponding Author. The corresponding author of this paper is Joongheon Kim and his postal address is as follows: Engineering Building #214, 145 Anam-ro, Seoul 02841, Korea (Phone: 82-2-3290-3223, E-mail: joongheon@korea.ac.kr).

References

Chen, Z.; Yu, Z.; Li, J.; You, L.; and Tan, X. 2024. DualAT: Dual Attention Transformer for End-to-End Autonomous Driving. In *Proc. IEEE ICRA*. Yokohama, Japan.

Chitta, K.; et al. 2023. TransFuser: Imitation With Transformer-Based Sensor Fusion for Autonomous Driving. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11): 12878–12895.

Xu, D.; Li, H.; Wang, Q.; Song, Z.; Chen, L.; and Deng, H. 2024. M2DA: Multi-Modal Fusion Transformer Incorporating Driver Attention for Autonomous Driving. arXiv:2403.12552.