Mitigating Attention Collapse via Mean-Deviation Constrained Optimization

Jiyun Kim[1] and Kyunghwan Choi[2*]

[1] AI Graduate School, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea
jiyun6606@gm.gist.ac.kr
[2] Cho Chun Shik Graduate School of Mobility, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
kh.choi@kaist.ac.kr
https://kaist-mic-lab.github.io

Abstract. Attention mechanisms are widely used in deep learning to compute contextual representations, but they are prone to collapse when attention weights concentrate excessively on a few tokens, potentially degrading model performance. We propose an Mean-Deviation Constrained Attention (MDCA), an optimization-based attention mechanism that constrains the mean-deviation of attention weights to mitigate attention collapse. The constraint is formulated as an inequality condition and is efficiently handled using the Augmented Lagrangian Method (ALM), enabling explicit control over attention concentration. Unlike heuristic approaches such as dropout or temperature scaling, our method introduces a principled regularization framework grounded in constrained optimization. We evaluate the proposed method on two tasks: (i) selective attention for handwriting classification using the Badge-MNIST dataset, in comparison with standard baselines including vanilla attention, entropy regularization, and temperature scaling; and (ii) imitation learning on the nuPlan dataset, compared with a representative state-of-the-art planner. On Badge-MNIST, our method improves attention selectivity and accuracy across seeds. On nuPlan, it yields safety driving in reactive closed loop and openloop evaluation while maintaining modest.

Keywords: Attention · Attention Collapse · Constrained Optimization · Augmented Lagrangian Method.

1 Introduction

Attention mechanisms [18] are now central mechanisms in learning systems across vision [9], language [8] and multimodal domains [17]. In learning systems, attention assigns the importance of input vectors which called embedded tokens for a given task. By allocating importance over key elements, models are

[*] This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00554087)

enable to create powerful representations. However, attention distributions are often not robust, frequently collapsing onto a few tokens. It leads to poor performance such as distribution shift, shortcuts [10] and occlusion. This effect is exacerbated when the model has only a few queries to attend.

There have been several attempts to mitigate such collapse through heuristic dropout on attention maps [15], softmax temperature scaling [11], [19] entropy bonuses [14], KL-to-uniform [16], L2 [4] and penalty on dispersion [13], [3]. However, these approaches require manual tuning of hyperparameters, and lack a clear bound on peaking often dulling warranted focus or failing to stop collapse.

An optimization-based attention mechanism, Mean-Deviation Constrained Attention (MDCA), is introduced to mitigate attention collapse by constraining the mean-deviation of attention weights via the Augmented Lagrangian Method (ALM). ALM is a classical optimization technique for efficiently handling inequality constraints [2], [13]. Rather than continually flattening attention scores with heuristics, a bound (inequality constraint) is imposed on the mean-deviation of the attention weights and enforced through ALM, yielding explicit control over attention concentration. This principled regularization framework is optimization-centric and does not rely on heuristic tuning.

Empirical evaluation is conducted on two tasks. On **Badge-MNIST**, a comprehensive ablation suite is performed: NONE (vanilla attention), ENT (entropy regularization), TEMP (softmax temperature scaling), KL (KL-to-uniform / label smoothing), PENONLY (penalty without multiplier updates), and the proposed method (Mean-Deviation Constrained Attention). The module is further adapted to imitation learning for autonomous driving within the **nuPlan** framework [5], a large-scale dataset and closed-loop planning benchmark. Across both domains, improvements are modest yet consistent, supporting the objective of enhancing stability and robustness without sacrificing warranted selectivity.

Our contributions are as follows:

- A mean-deviation constrained attention (MDCA) is introduced, implemented via the ALM, to explicitly suppress pathological attention peaking without relying on heuristic smoothing.
- Constraint enforcement is achieved through a dual update process that activates only when necessary, with no additional computation during inference.
- The proposed module is model-agnostic and compatible with any attention-based architecture.
- Our experimental results on Badge-MNIST and nuPlan show consistent improvements over baselines and enhanced robustness under challenging conditions. Across Badge-MNIST, our proposed method demonstrates reliable, consistent improvements and enhanced robustness. On nuPlan, compared to a state-of-the-art planner (planTF), our method improves safety and comfort performance in reactive closed-loop scenarios.

2 Related Works

Attention is a weighting operator over token pairs, widely used in Transformer models [18]. In scaled dot-product attention, queries Q, keys K, and values V yield the score matrix S,

$$S_{ij} = \frac{q_i^\top k_j}{\sqrt{d}}, \tag{1}$$

which is normalized row-wise (over j) to attention weights a,

$$a_{ij} = \frac{\exp(S_{ij})}{\sum_m \exp(S_{im})}, \tag{2}$$

and produces the head output

$$o_i = \sum_j a_{ij} v_j. \tag{3}$$

Here d denotes the key/query dimension. Temperature scaling replaces softmax(s) with softmax(S/T) for $T > 0$. Multi-head attention applies this operation to linearly projected (Q, K, V) across multiple heads and concatenates the outputs [18]. Several attention masks are used to fuse feature maps or predictions from different branches [6]. Common interaction patterns include self-attention [18], [12], within a sequence, and cross-attention [1], [18], across sequences (e.g., encoder–decoder). A known failure mode is *attention collapse*, where the normalized weights concentrate on a few positions and suppress signals from the rest, reducing robustness under noise or missing tokens.

To mitigate collapse, prior work smooths or calibrates the attention distribution, or penalizes dispersion of the normalized weights. However, each method has notable limitations. Temperature scaling [11], [19] controls global sharpness but requires layer and dataset-dependent tuning, making it difficult to generalize. Entropy bonuses [14] and KL-to-uniform [16] encourage spread, yet may over-smooth and suppress meaningful peaks, reducing selectivity. Variance-like penalties [4] nudge weights toward uniformity but act as soft preferences and can be sensitive to sequence length. Quadratic penalty schemes [13], [3] apply a margin penalty without dual feedback, which can either under-enforce the bound or stiffen optimization, harming convergence. Dropout [15] randomly masks tokens or channels to decorrelate activations. It provides only indirect control over the normalized attention and may remove informative inputs, leading to train–test mismatch or loss of critical information.

State Dropout Encoder (SDE) is a novel component of PlanTF [7], a state-of-the-art planner on nuPlan [5]. The term *ego state* denotes the ego vehicle's kinematic variables at time t and represented by $x_t = [p_x,\ p_y,\ \psi,\ v,\ a,\ s]$. During training, SDE is applied to a subset of the tokens so that non-geometric channels (e.g., ψ, v, a, s) are randomly dropped while position and heading tokens are retained. By occasionally removing individual ego-state channels, SDE discourages shortcut reliance on a few tokens and encourages the planner to exploit other inputs (surrounding-agent histories and map features). However, dropout occurs only during training, which can induce a train–test mismatch and occasional re-collapse at inference. The mask is indiscriminate and may drop critical channels, and there is no explicit target dispersion level which makes behavior sensitive.

3 Main Results

3.1 Problem Formulation as a Constrained Optimization Problem

The main contribution is a constrained optimization-based attention mechanism to mitigate attention collapse by bounding the mean-deviation of attention weights. The constraint is formulated as an inequality and efficiently handled using the ALM.

Let θ be the trainable parameters and $L_0(\theta)$ the task loss on a mini-batch of size B. For an attention layer with H heads and C attended channels, let $\alpha_h^{(b)}(\theta) \in \Delta^{C-1}$ be the normalized attention of head h for batch index b (i.e., $\alpha_{h,j}^{(b)} \geq 0$ and $\sum_{j=1}^{C} \alpha_{h,j}^{(b)} = 1$). Heads are averaged to obtain a single distribution:

$$\bar{\alpha}^{(b)}(\theta) = \frac{1}{H} \sum_{h=1}^{H} \alpha_h^{(b)}(\theta) \in \Delta^{C-1}.$$

Define the dispersion from the uniform vector $u = \frac{1}{C}$ as follows

$$\begin{aligned} D\left(\bar{\alpha}^{(b)}(\theta)\right) &= \frac{1}{C} \left\| \bar{\alpha}^{(b)}(\theta) - u \right\|_1 \\ &= \frac{1}{C} \sum_{j=1}^{C} \left| \bar{\alpha}_j^{(b)}(\theta) - \frac{1}{C} \right| . \end{aligned} \tag{4}$$

The batch-averaged deviation is

$$\bar{D}(\theta) = \frac{1}{B} \sum_{b=1}^{B} D\left(\bar{\alpha}^{(b)}(\theta)\right) . \tag{5}$$

Since $\bar{\alpha}^{(b)}(\theta) \in \Delta^{C-1}$, a valid bound is $0 \leq \bar{D}(\theta) \leq \frac{2(C-1)}{C^2}$. Let $m \in \left[0, \frac{2(C-1)}{C^2}\right]$. Consider the constrained problem:

$$\min_{\theta} L_0(\theta) \quad \text{s.t.} \quad \bar{D}(\theta) \leq m. \tag{6}$$

Constraint enforcement employs the augmented Lagrangian with multiplier $\lambda \geq 0$ and penalty weight $\rho > 0$:

$$\mathcal{L}_{\text{aug}}(\theta, \lambda) = L_0(\theta) + \lambda \left[\bar{D}(\theta) - m\right]_+ + \frac{\rho}{2} \left[\bar{D}(\theta) - m\right]_+^2, \tag{7}$$

where $[x]_+ = \max(x, 0)$.

3.2 ALM-based Update Rule

The parameter update rule is derived using the ALM to solve the constrained problem. The augmented objective is

$$\mathcal{L}_{\mathrm{aug}}(\theta, \lambda) = L_0(\theta) + \lambda\,[g(\theta)]_+ + \tfrac{\rho}{2}\,[g(\theta)]_+^2, \tag{8}$$

with penalty weight $\rho > 0$, and the multiplier is updated per step as

$$\lambda \leftarrow \lambda + \rho\,[g(\theta)]_+, \qquad \lambda_0 = 0. \tag{9}$$

When $D(\theta) \leq m$, $[g(\theta)]_+ = 0$ and the constraint term is inactive. Otherwise the penalty and multiplier increase pressure toward feasibility. The quantity $g(\theta)$ is evaluated on the same mini-batch as $L_0(\theta)$, gradients are backpropagated through $g(\theta)$ to the attention weights, and the multiplier is updated once per optimization step.

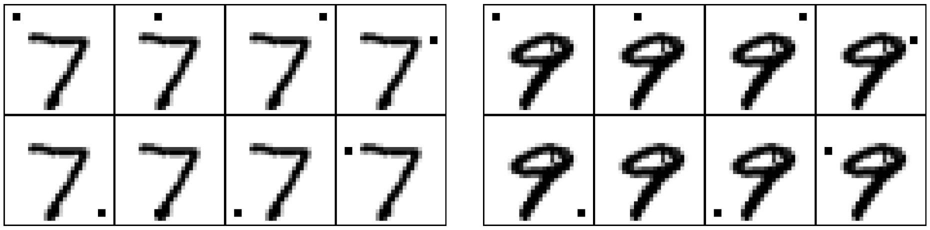
4 Experiments

4.1 Experimental Setup

The proposed method was evaluated on two tasks: (i) selective attention for handwriting classification using the Badge-MNIST dataset, against baselines including NONE, ENT, TEMP, KL, and PENONLY, and (ii) imitation learning on the nuPlan dataset, compared with a representative state-of-the-art planner which implements the dropout mechanism on attention. In the Badge-MNIST task, training was performed for 20 epochs with a total batch size of 256. The Adam optimizer was used with a learning rate of 3×10^{-4} and a weight decay of 1×10^{-4}. Evaluation was conducted for 300 epochs. All comparison methods were trained under identical settings for a fair comparison. In the nuPlan task, training was performed on NVIDIA RTX 4090 GPUs for 20 epochs with a total batch size of 32, using Adam with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-4}.

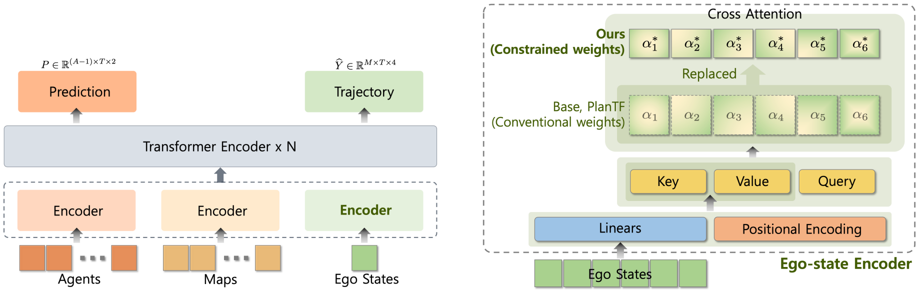
Badge-MNIST is a synthetic dataset based on MNIST. In the Badge-MNIST task, each MNIST digit is divided into a 3×2 grid of patches. A small badge is added to a specific corner of the image based on the digit label, requiring the model to focus on digit-specific evidence. The model applies attention pooling over the six patch tokens to classify the digit. Figure 1 shows example images from the dataset. All comparison methods are trained with the same backbone and optimization schedule.

NuPlan is a large-scale closed-loop imitation learning based planning benchmark for autonomous driving [5]. The dataset consists of 1,282 hours of driving data which collected across four different cities Las Vegas, Pittsburgh, Boston and Singapore. It provides the real-world driving dataset from human drivers and



(a) (b)

Fig. 1: Example dataset of Badge-MNIST : (a) Digit "7" with badge at 8 places. (b) Digit "9" with badge at 8 places.



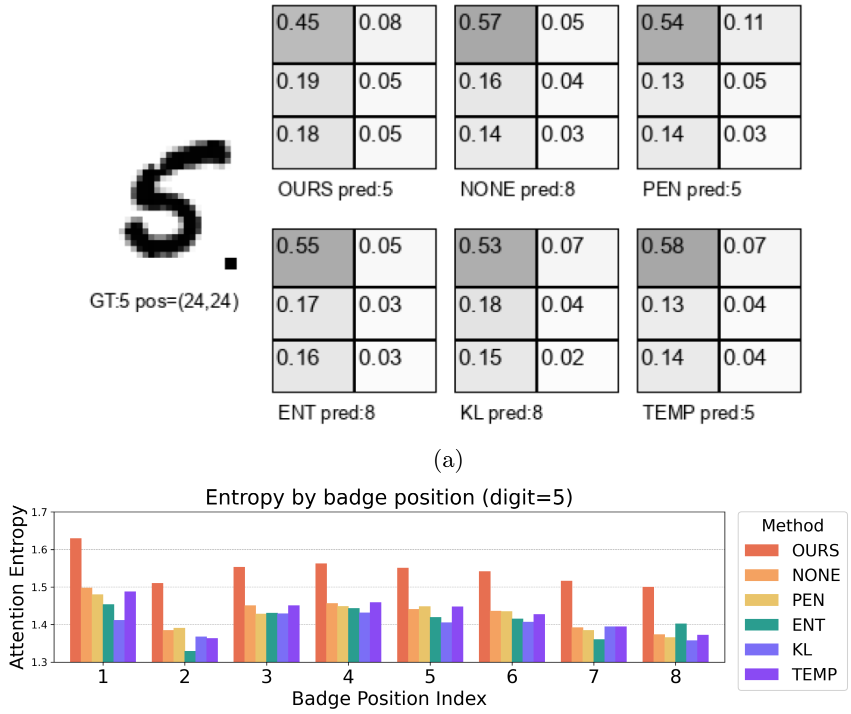
(a) (b)

Fig. 2: Planner architecture of Base (vanilla), PlanTF (with SDE), and Ours (with MDCA): (a) Architecture overview and (b) ego-state encoder (Ours).

various scenarios which annotated tags for specific scenarios. Also, it includes the framework and metrics for training and evaluating the planner in open-loop score (OLS), non-reactive closed-loop score (NR-CLS), and reactive closed-loop score (R-CLS).

The planner adopts a Transformer-based encoder–decoder architecture which shown in Figure 2a inspired by PlanTF [7], which reports state-of-the-art results on nuPlan. The base model follows the PlanTF design but omits the SDE. In PlanTF, SDE tokenizes individual ego-state variables, embeds them, and aggregates them with a learnable query via cross-attention while randomly dropping state tokens (except position and heading) during training to reduce shortcut reliance.

Our planner inputs comprise the ego-state sequence, surrounding-agent histories, and map features. Each input modality is encoded by its own encoder with self attention. Fusion is performed in a joint Transformer block with self-attention over the concatenated modality tokens. The decoder outputs a T-step and M-mode distribution over future ego poses. During training, the ALM regularizer is applied to the head-averaged normalized attention weights of the ego-state encoder cross-attention, as shown in Figure 2b. All other components,



(a)

(b)

Fig. 3: The example of Badge-MNIST experiment, Digit "5" with badge at bottom-right corner. (a) Attention maps of comparison methods. (b) Entropy of attention maps of comparison methods.

including data augmentation and inference, remain identical across comparison methods to ensure a fair comparison.

Here are the model variants used in the experiments:

- **Base (Vanilla).** Ego 6-D state is summarized by a small MLP, with no SDE or ALM. Agent/map encoders and the shared Transformer are identical to the other variants.
- **PlanTF (with SDE).** Same backbone, but the ego state uses a single-query state-attention encoder with SDE during training. Position and heading tokens are retained.
- **Ours (with MDCA).** Same backbone with ego-state encoder cross-attention (ego queries attending to agent and map keys/values). During training, the normalized attention weights of this cross-attention are regularized per head by an ALM dispersion constraint and aggregated by the mean across heads.

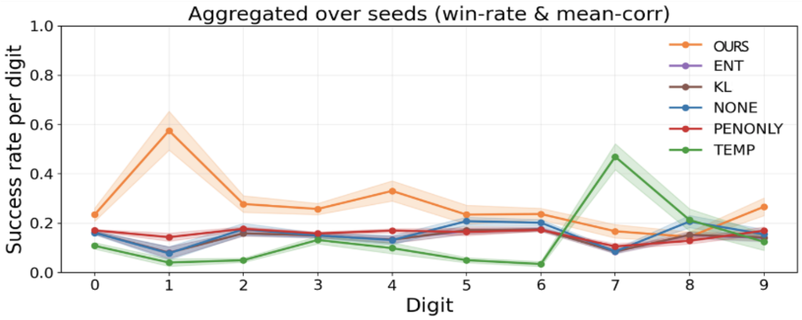


Fig. 4: Badge-MNIST per-digit results aggregated over 25 seeds. Curves show mean win rate per digit, shaded regions indicate variability across seeds (±95% CI).

4.2 Experimental Results

Results on Badge-MNIST Figure 3 illustrates a Badge-MNIST example with a 3×2 attention grid. As shown in Figure 3a, compared to baselines, the proposed model allocates more attention to digit-relevant regions and reduces focus on the badge. Attention is a probability distribution over six patches, and controls overall concentration to avoid attention collapse onto a few patches rather than suppressing a specific cue. Consistent with this, Figure 3b shows that Ours yields higher attention entropy across most badge positions, indicating mitigation of badge-centered collapse and a broader allocation over relevant patches. In contrast, the baselines (NONE, ENT, KL, PEN, TEMP) often show lower entropy at certain positions, consistent with peakier, badge-focused attention.

Per-digit success rates over 25 seeds are shown in Figure 4, where lines indicate the mean across seeds and shaded bands represent 95% confidence intervals. Our proposed method, mean-deviation constrained attention (MDCA), leads on most digits—especially 1, 4, and 9—and keeps variability moderate, indicating gains in both accuracy and stability while preventing attention collapse. Heuristic baselines (NONE, ENT, KL, PENONLY) stay low, reflecting either diffuse attention or collapse onto the badge; TEMP is highly digit-dependent, dipping on most digits but spiking on 7 due to a single global temperature.

Overall, these results show that the MDCA effectively prevents attention collapse onto the badge, maintains selective focus on digit-relevant regions, and consistently improves per-digit accuracy and stability across random seeds. Unlike heuristic regularizations or scaling, ALM activates only when attention concentration exceeds the specified bound, allowing the model to retain sharp focus when appropriate and intervening only when necessary.

Results on nuPlan Evaluation follows the nuPlan benchmark. Results are reported on two official test splits, `test14-random` and `test14-hard`, each comprising 261 scenarios across 14 scenario types. `test14-random` consists of ran-

Method	**Test14-random**			**Test14-hard**		
	OLS	NR-CLS	R-CLS	OLS	NR-CLS	R-CLS
Base	86.64	80.01	74.48	82.48	65.30	53.11
planTF	86.27	85.23	79.36	83.34	70.03	59.83
Ours	87.67	84.91	78.31	86.31	69.48	64.64

Table 1: Comparison of planner performance on nuPlan test14-random and test14-hard splits.

Method	Collisions	TTC	Drivable	Comfort	Progress	Speed
Base	88.11	81.50	92.64	88.23	72.79	98.02
planTF	85.84	80.88	92.64	93.01	**84.55**	97.01
Ours	**90.63**	**85.49**	**94.02**	**98.16**	84.28	**98.22**

Table 2: R-CLS sub-metric comparison on nuPlan test14-hard split. Higher is better.

domly sampled scenarios (fixed after selection), whereas `test14-hard` is a curated set of challenging scenarios. Reported challenges are OLS (open-loop score against logged trajectories), NR-CLS (closed loop with nonreactive log-replay agents), and R-CLS (closed loop with reactive agents). Higher is better for all metrics.

Results are summarized in Table 1 for both `test14-random` and `test14-hard` splits. On `test14-random`, our method achieves the highest OLS (87.67), outperforming Base (+1.03) and planTF (+1.40). In closed-loop evaluation, NR-CLS (84.91) and R-CLS (78.31) are comparable to planTF (85.23 and 79.36), but clearly surpass Base (+4.90 NR-CLS, +3.83 R-CLS), demonstrating strong open-loop fidelity with minimal loss in closed-loop performance. On `test14-hard`, Ours attains the best OLS (86.31) and the best R-CLS (64.64), with NR-CLS (69.48) near planTF (70.03). The R-CLS gain over planTF is +4.81, and over Base is +11.53, indicating a significant improvement in reactive closed-loop performance on challenging scenarios.

To further analyze the planner's behavior, Table 2 reports reactive closed-loop sub-metrics on `test14-hard`. Our method improves safety and comfort compared to baselines. The Collisions, TTC, and Drivable scores are highest, indicating safer and more rule-compliant driving. Comfort is also substantially improved, while efficiency metrics (Speed, Progress) are maintained. This matches the overall R-CLS gains, showing safer and smoother planning without loss of efficiency.

For a more detailed evaluation of open-loop planning, Table 3 presents sub-metrics on `test14-hard`. Average Displacement Error (ADE) and Final Displacement Error (FDE) assess position accuracy, while Average Heading Error (AHE) and Final Heading Error (FHE) measure heading accuracy. Miss Rate

Method	ADE	FDE	AHE	FHE	MR
Base	79.65	61.34	93.91	91.05	95.95
planTF	81.93	64.31	93.43	90.64	96.32
Ours	**84.99**	**67.81**	**94.11**	**91.55**	**97.06**

Table 3: OLS sub-metric comparison on nuPlan test14-hard split. Higher is better.

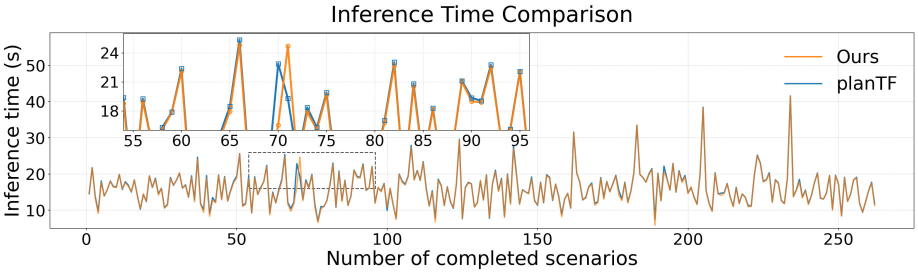


Fig. 5: Inference time (seconds) per scenario comparison on nuPlan.

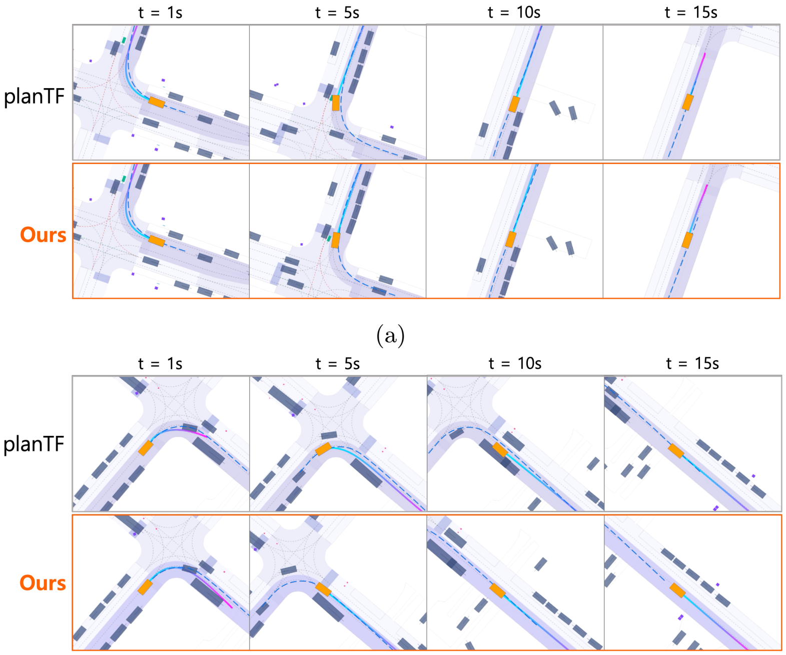
(MR) indicates the proportion of missed predictions, with higher values representing better performance. The proposed method achieves the highest scores across all five metrics, with notable improvements in displacement-based measures (ADE/FDE), indicating stronger trajectory fidelity to logged data. Heading and miss-rate scores also consistently outperform both baselines.

To evaluate model overhead, training and inference runtimes were measured and compared with planTF. Training time per epoch remained stable with low variance. MDCA incurs a small additional cost of approximately 2 minutes per epoch due to the ALM computations, a reasonable trade-off given the observed gains in performance and robustness. Figure 5 shows per-scenario inference times on the `test14-random` split. Inference runtimes for MDCA and planTF are nearly identical. When per-scenario differences occur, they are within 2 seconds, which indicates that MDCA adds negligible inference overhead.

Typical failure–recovery contrasts are illustrated in Figure 6. In (a), planTF turns right and collides with a motorcycle waiting at the signal ($t = 5$s), whereas Ours avoids the obstacle. In (b), planTF becomes confused during the right turn ($t = 5$s) and, after completing the turn ($t = 10$s), clips a parked car, while Ours negotiates the corner without incident and maintains a safe distance. These cases mirror the gains in R-CLS and Comfort, highlighting improved stability and safety in complex scenes.

5 Conclusion

This work presents a constrained optimization-based approach, MDCA to mitigating attention collapse by imposing a mean-deviation constraint on attention



(a)

(b)

Fig. 6: The scenario examples from nuPlan test14-hard split. (a) The planTF collides with a motorcycle waiting at the signal while our policy safely avoids the obstacle. (b) The planTF clips a parked car during a right turn while our method negotiates the corner without incident.

weights. The constraint is formulated as an inequality and efficiently solved using the ALM, providing a principled alternative to heuristic regularization techniques. Experiments on the Badge-MNIST dataset demonstrate consistent improvements in attention selectivity and effective suppression of attention collapse across digits and random seeds. In the nuPlan experiments, the proposed method leads to more balanced and cautious planning behavior, improving safety and comfort in reactive scenarios. Overall, the proposed method offers a generalizable and effective framework for improving the reliability of attention mechanisms by directly addressing collapse in attention distributions.

In large multi-head architectures, a single global λ often blunts useful heads and misses local collapse, yielding oversmoothing. Our future works will explore head-wise, conditional λ that activates only when collapse signals (low entropy, high sharpness, large dominance gap) are detected, while keeping $\lambda \approx 0$ for well-behaved heads. This targets collapse precisely and improves both selectivity and stability without unnecessary smoothing.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2014)
2. Bertsekas, D.P.: Nonlinear programming. vol. 48, pp. 334–334. Taylor & Francis (1997)
3. Bertsekas, D.P.: Constrained optimization and Lagrange multiplier methods. Academic press (2014)
4. Bishop, C.M., Nasrabadi, N.M.: Pattern recognition and machine learning, vol. 4. Springer (2006)
5. Caesar, H., Kabzan, J., Tan, K.S., Fong, W.K., Wolff, E., Lang, A., Fletcher, L., Beijbom, O., Omari, S.: nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles (2021)
6. Cheng, J., Dong, L., Lapata, M.: Long short-term memory-networks for machine reading (2016)
7. Cheng, J., Chen, Y., Mei, X., Yang, B., Li, B., Liu, M.: Rethinking imitation-based planners for autonomous driving. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 14123–14130. IEEE (2024)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale (2020)
10. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
11. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
12. Lin, Z., Feng, M., Santos, C.N.d., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A structured self-attentive sentence embedding (2017)
13. Nocedal, J., Wright, S.J.: Numerical optimization. Springer (2006)
14. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., Hinton, G.: Regularizing neural networks by penalizing confident output distributions (2017)
15. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. vol. 15, pp. 1929–1958. JMLR. org (2014)
16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
17. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the conference. Association for computational linguistics. Meeting. vol. 2019, p. 6558 (2019)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. vol. 30 (2017)
19. Wang, P.H., Hsieh, S.I., Chang, S.C., Chen, Y.T., Pan, J.Y., Wei, W., Juan, D.C.: Contextual temperature for language modeling (2020)