Thank you all for your thoughtful comments; we address your concerns below.

[R3] Clarify relation between TLT, Occam's razor, and interpretability We discussed this a bit in our paper by drawing connections to Feldman's work (L36), but we agree that the relation between the three topics should be expanded upon. Since TLT measures the total stepwise change in attention weights, decreasing TLT can be seen as an instance of the minimum description length (MDL) principle. The MDL principle formalizes Occam's razor and is a reasonable relaxation of Kolmogorov complexity. Previous work in cognitive neuroscience starting from [Hochberg and McAlister. A quantitative approach, to figural "goodness". *Journal of Experimental Psychology*, 46(5):361, 1953.] to its modern follow-up studies empirically demonstrate that such quantitative complexities are inversely proportional to interpretability. We will add the discussion of such relevant studies to section 1.

[R2] "In addition, MAC is only tested on the syntactic CLEVER dataset."
[R3] "the paper does not contain any experiments on any of the visual question answering (VQA) real datasets"

Model(step)	MAC(2)	DAFT MAC(2)	MAC(4)	DAFT MAC(4)	MAC(8)	DAFT MAC(8)
Accuracy	52.4 ± 0.3	52.4 ± 0.7	52.6 ± 0.1	52.6 ± 0.4	52.7 ± 0.5	52.8 ± 0.5
TLT	0.43 ± 0.09	0.19 ± 0.14	0.94 ± 0.23	0.34 ± 0.08	2.02 ± 0.23	0.48 ± 0.05

To show the benefits of DAFT on real-world datasets, we additionally trained and evaluated DAFT MAC on the GQA dataset [Hudson and Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *CVPR*, 2019.]. Due to insufficient time, we trained the models with the 1M subset of the GQA dataset rather than the 14M full training set. We found that DAFT MAC preserves all the benefits: no harm on performance, lower TLT, and consistent and chunked attention maps. Note that that GQA requires less steps than CLEVR, and also that TLT values for GQA are much lower than that for CLEVR (over 4 for MAC(8)). This evidence supports our claim that TLT measures effective reasoning path length. We will add these results and accompanying visualizations to appendix.

[R3] "The manuscript does not talk about the additional cost of running the ODE solvers including run-time analysis and comparisons."

Model (solver)	MAC	DAFT MAC (euler)	DAFT MAC (rk4)	DAFT MAC (dopri5; used in training)
Time (ms)	153.7 ± 3.8 (1x)	167.9 ± 1.7 (1.09x)	189.7 ± 1.9 (1.23x)	365.5 ± 12.5 (2.37x)
Accuracy	98.6 ± 0.2	98.7 ± 0.2	98.9 ± 0.2	98.9 ± 0.2
TLT	2.06 ± 0.15	1.76 ± 0.07	1.62 ± 0.06	1.62 ± 0.06

For more detailed run-time analysis, we measured the time, accuracy, and TLT obtained by using various ODE solvers *during evaluation* of five different 4-step DAFT MAC. We used two fixed-step solvers (Euler method and Runge-Kutta 4th order method with 3/8 rule) and one adaptive-step solver (Dormand-Prince method) that we used during training. We found that during evaluation, rk4 solves all the dynamics generated from CLEVR dataset. Even the simplest euler method results in lower accuracy and higher TLT compared to vanila MAC. We will add these results to section 5.2.

[R2] What if instead of neural ODE, a_t simply depends on a_{t-1}? As discussed in the previous question on run-time analysis, we have tried the Euler method, which is equivalent to the residual form $a_t = a_{t-1} + f(a_{t-1})$. While using such a simple dependence improved over vanila MAC, this scheme was outperformed by DAFT MAC with more sophisticated solver both in terms of accuracy and TLT.

[R2] "the approach is novel, but only applicable for MAC-like solution." We believe that we insufficiently emphasized the generality of DAFT. In fact, DAFT can be applied to any attention update via a one-line change: $\mathbf{a}_{t_1} = \mathbf{W}^{1\times d}(\mathbf{W}_{t_1}^{d\times d}\mathbf{q} \odot \mathbf{cw}) \rightarrow \mathbf{a}_{t_1} = \mathbf{a}_{t_0} + \int_{t_0}^{t_1} f(\mathbf{a}_t, t)dt$. DAFT can therefore be used in the types of attention that R2 mentions such as glimpses or Transformer heads. To highlight this fact, we will add short pseudocode detailing the application of DAFT to other attention mechanisms to section 4.

[R2] "...DAFT doesn't improve the performance of MAC. I do tend to believe that low TLT should also correlate with better performance." The main focus of this paper is not improving performance, but rather on being maximally interpretable while harming performance as little as possible. TLT is highly correlated with the interpretability of the model. This claim is backed by many previous works in cognitive science in addition to our empirical results.

[R1] "The paper lacks explanation about ODE Solvers used in the method..." We will add a concise explanation of fixed- and adaptive-step ODE solvers to section 2.3 for self-containedness.

[R1] "Further discussion and comparison to discrete attention mechanism..." We will add connections between our method to pondering techniques such as ACT and dynamic halting (of Universal Transformer) to section 2.3.

[R2] "The qualitative example (Fig. 3) is kind of weird. The attention seems to focus a lot on the edges, and never on green blocks, can you please comment on that?" In the CLEVR dataset, "blocks" refer to cubes. The image in figure 3 contains no green cubes, so the model is correct in not attending to any other object.