7. Appendix

7.1. What Does the Causal Model Learn from the Future?

To further validate our empirical findings and understand exactly *why* future-privileged supervision improves causal prediction, we conduct two deeper diagnostic analyses on EGTEA Gaze+.

Isolating the Impact of the Foundation Model. A critical necessity in our controlled study is ensuring that the observed performance gains are genuinely driven by the future-aware distillation, rather than merely stemming from the powerful frozen DINOv3 encoder. To strictly isolate this, we re-implemented the causal GLC baseline [13] equipped with the *exact same* frozen DINOv3 backbone.

As shown in Table 7, the DINOv3-equipped GLC achieves only 33.1 F1, lagging significantly behind our distilled causal student (45.9 F1), despite requiring over 13× more FLOPs. This drastic performance gap confirms a key finding: deploying a strong foundation model is, by itself, insufficient for solving online causal gaze prediction. The true driver of the performance boost is the successful distillation of future context coupled with an efficient decoder, validating the efficacy of our controlled future-privileged framework.

Method	Backbone	F1 ↑	FLOPs (G) ↓	FPS ↑
GLC (Causal)	DINOv3	33.1	226.9	40.9
ECOGaze	DINOv3	**45.9**	**17.4**	**59.0**

Table 7. Impact of the foundation model on EGTEA Gaze+. Even when equipped with the exact same frozen DINOv3 encoder, the causal GLC baseline significantly underperforms ECOGaze, confirming that our gains stem fundamentally from future-privileged distillation and our efficient decoder design.

Beyond Hand-Tracking: Robustness to Hand Visibility. In egocentric videos, gaze is often strongly correlated with hand movements. This raises an important question regarding what the causal student actually learns from the future: *Does future-privileged supervision merely teach the model a shallow "hand-tracking bias," or does it impart deeper anticipatory intent?*

To answer this, we partitioned the EGTEA Gaze+ test set into "With Hands" and "Without Hands" subsets. As detailed in Table 8, even when hands are entirely absent from the current frame, our causal student maintains a robust F1 score of 44.8 (compared to 39.9 F1 for the GLC baseline) and a high precision of 59.4. This resilience provides a crucial empirical insight: the future context distills broader scene semantics and high-level task progression (e.g., relevant object affordances) into the causal model, enabling robust intent prediction even when explicit hand-object interaction cues are temporarily invisible.

Method	With Hands			Without Hands		
	F1	Prec.	Rec.	F1	Prec.	Rec.
Attention Transition [12]	26.6	18.4	47.5	19.6	15.7	26.0
GLC (Causal) [13]	43.9	34.7	59.6	39.9	31.0	56.2
ECOGaze (Ours)	**45.9**	**36.8**	**60.1**	**44.8**	**59.4**	**35.9**

Table 8. Robustness to hand visibility on EGTEA Gaze+. ECOGaze maintains a substantial performance lead over the baselines even in frames where hands are entirely absent, demonstrating that it learns deeper anticipatory intent beyond shallow hand-tracking biases.