A. Appendix

This Appendix includes:

1. Reference to video with qualitative examples of EgoT2
2. Experimental Setup
3. Additional Results
4. Additional Visualizations

A.1. Video containing qualitative results

We invite the reader to view the video available at https://vision.cs.utexas.edu/projects/egot2/ where we show qualitative examples of (a) how EgoT2 captures inter-frame and inter-task relations, (b) video retrieval results using attention weights of task tokens and (c) how EgoT2-g makes predictions conditioned on the task prompt and given video.

From these examples, we can see that EgoT2 offers good interpretability on task relations, revealing clearly which temporal segments and which subsets of tasks contribute to improving a given task. Moreover, we run EgoT2-s on all AR validation videos and retrieve video segments with top PNR and OSCC weights. The results show that videos with large PNR and OSCC weights actually all involve heavy human-object interactions, which is the focus of these two tasks. Finally, we observe that EgoT2-g successfully performs task translation conditioned on the task of interest and task tokens through encoder-decoder attention weights.

A.2. Experimental Setup

Below we provide detailed descriptions of the 7 tasks we adopt in our study.

- *Point-of-no-return Keyframe Localization (PNR)*: given a short video of a state change, estimate the keyframe that contains the time at which a state change begins.
- *Object State Change Classification* (OSCC): given a video clip, classify whether an object state change has taken place or not.
- *Action Recognition* (AR): classify the action (verb and noun) of the camera wearer from a short egocentric video clip; there are 115 verb categories and 478 noun categories.
- *Long-term Action Anticipation* (LTA): given a video clip, predict the camera wearer's future sequence of actions; the action vocabulary is identical to that used in AR.
- *Looking At Me* (LAM): given an egocentric video in which the faces of social partners have been localized and identified, classify whether each face is looking at the camera wearer.
- *Talking To Me* (TTM): given a video and audio with the same tracked faces, classify whether each face is talking to the camera wearer.
- *Active Speaker Detection* (ASD): given a cropped face video clip and corresponding audio segments, identify whether this person is speaking.

A.2.1 Dataset Details

Note that Ego4D does not have a common training set that provides labels for all tasks. In all of our experiments, the task-specific models are trained on 7 subsets of Ego4D. Table 7 reports the percentage of training videos shared between task pairs. Among the 7 task datasets, the average data overlap between two tasks is 22.5%, and for 57.1% of the task pairings there is strictly disjoint training data. The limited intersections in these 7 task datasets lend support to the generalizability of EgoT2 across multiple datasets.

	PNR	OSCC	AR	LTA	LAM	TTM	ASD
PNR	100	48.2	32.6	32.6	0	0	0
OSCC	48.2	100	67.0	67.0	0	0	0
AR	32.6	67.0	100	100	0	0	0
LTA	32.6	67.0	100	100	0	0	0
LAM	0	0	0	0	100	12.8	12.8
TTM	0	0	0	0	12.8	100	100
ASD	0	0	0	0	12.8	100	100

Table 7. Percentage of training videos shared between task pairs for the 7 tasks used in the experiments. There is a low level of data overlap between individual task pairs.

A.2.2 Implementation Details

Task-Specific Translation. As shown in Table 1 of the main paper, the LTA task-specific backbone requires videos of 16 seconds while the other three human-object interaction tasks operate on videos of 8 seconds. Therefore, when $\mathcal{T}_p$ is LTA, we slide the other task-specific backbones along the 16-seconds time window to obtain auxiliary task features; the stride size is set to 8 seconds. When $\mathcal{T}_p$ is PNR, OSCC or AR, LTA is not a valid auxiliary task since its task-specific model requires video of a longer temporal span than provided in these three datasets. While it is possible to expand the video for the LTA model to be applicable, we aim at avoiding advantages brought by a longer time window for a fair comparison with prior work and thus exclude LTA as the auxiliary task. Nevertheless, to provide a complete evaluation, we consider one such special case when the primary

task is AR and the auxiliary task is LTA (see results marked with * in Table 9). Similarly, for the 3 human-human interaction tasks, LAM dataset provides video instances of 0.2 seconds while the TTM and ASD task-specific model requires videos of a longer time span. Consequently, LAM is not considered as the primary task.

Task-General Translation. For human-object interaction tasks, we follow common practices [26] and treat predicting verbs and nouns as two separate tasks. EgoT2-g is thus jointly optimized on 6 tasks: PNR, OSCC, AR Verb, AR Noun, LTA Verb and LTA Noun. We simplify the LTA task as predicting actions at a single timestamp into the future as opposed to the 20 timestamps considered in the original benchmark since otherwise the decoder would be heavily biased towards the LTA task (see parameter comparison in Table 2 of the main paper). While EgoT2-s predicts future actions at future 20 timestamps and uses edit distance@20 (ED@20) as the metric, we report verb and noun accuracy for LTA in EgoT2-g. For human-human interactions, while LAM is not considered as the primary task for EgoT2-s, EgoT2-g provides the flexibility to incorporate LAM in training as well. In particular, when task prompt is LAM, we feed LAM tokens as input to the task fusion transformer and do not use other task tokens following the time span guidelines discussed above.

Tokenization and Detokenization. We construct a small task-related vocabulary for the sequence decoder in EgoT2-g to work. Namely, it is based on the label spaces of all candidate tasks and maps the original output label to a vocabulary. For PNR, we transform the output keyframe (*i.e.*, an integer from 0-15) to be its character format. For OSCC/LAM/TTM/ASD, we transform the output label to the word 'True' or 'False'. For AR and LTA, we use the verb and noun vocabulary and transform the label to the word. In addition, we include the 7 task prompts (*i.e.*, PNR, OSCC, AR, LTA, LAM, TTM and ASD) in the vocabulary. Consequently, we can transform the original label for all tasks to be a sequence and transform the predicted sentence back to the original label since it is a one-to-one mapping. Note that EgoT2-g is not sensitive to the choice of prompts. For example, the human-human interaction task prompts are [LAM], [TTM], [ASD], but [TaskA], [TaskB], [TaskC] would work too. Any output tokens outside the target task's label space are considered incorrect predictions. We find that EgoT2-g learns to predict words within the target task dictionary after a few epochs.

Hyperparameters and Optimization. Our implementation is based on the official Ego4D codebase.[6] EgoT2-s retains the same training configurations (*e.g.*, batch size, optimizer, total number of training epochs) unless otherwise specified. (1) $\mathcal{T}_p$ is PNR: Transfer (AR) is implemented as a SlowFast backbone pretrained on AR dataset followed by a 1-layer MLP with hidden dimension of 4096 and the PNR prediction head. Similarly, Finetuning and Transfer (OSCC) consists of a I3D ResNet-50 backbone pretrained on PNR and OSCC respectively followed by a 1-layer MLP with hidden dimension of 512 and the PNR prediction head. Late Fusion uses 3 1-layer MLPs to map features generated by each task-specific model (*i.e.*, PNR, OSCC and AR) to be 512-dimensional and concatenates the three task-specific features; the concatenated features are then passed to the PNR prediction head. EgoT2-s consists of 6-layer transformer encoders with hidden dimension of 128. (2) $\mathcal{T}_p$ is OSCC: we follow the same way as in PNR to implement these baselines, and the task fusion transformer in EgoT2-s has 5 layers with hidden dimension set as 128. (3) $\mathcal{T}_p$ is AR: Late Fusion follows the same design as in PNR and OSCC but has hidden dimension equal to 256. EgoT2-s uses a transformer encoder of 3 layers and hidden dimension set as 128. (4) $\mathcal{T}_p$ is LTA: The hidden dimension of Finetuning, Transfer and Late Fusion is set as 2048. EgoT2-s has a 1-layer transformer encoder with 128 dimension. (5) $\mathcal{T}_p$ is TTM: Finetuning and Transfer baselines are implemented as 3-layer MLPs with hidden dimension set as 1024 and 512. Late Fusion uses a 2-layer MLP to take concatenated features as input and passes the processed features to the TTM prediction head. EgoT2-s uses a 1-layer transformer encoder with hidden dimension of 128. (6) $\mathcal{T}_p$ is ASD: The baselines follow the same design as in TTM, and the hidden dimension of Transfer and Late Fusion is set as 6144 and 2048, respectively. EgoT2-s uses a 1-layer transformer encoder with hidden dimension of 256. Learning rate is set as 1e-3.

For EgoT2-g on human-object interaction tasks, we use a batch size of 4×8 distributed over 8 GPUs. The task translator consists of 3 transformer encoder layers and 3 transformer decoder layers with hidden dimension equal to 512. We use AdamW optimizer with learning rate and weight decay set as 1e-4. For human-human interaction tasks, we set the batch size for LAM, TTM and ASD to be 256, 15 and 1800 respectively to balance three dataloaders. The task translator has 1 transformer encoder layer and 1 transformer decoder layer with hidden dimension set as 128. We use Adam optimizer with learning rate of 5e-4 and weight decay of 5e-5. All models are trained for 20 epochs.

A.3. Additional Results

Analysis on Task Relations. From Table 2-3 in the main paper, we observe the superior performance of EgoT2-s. Moreover, Transfer baseline results from these two tables offer insights on task relations. Intuitively, tasks within one benchmark (*e.g.*, AR and LTA) are very related and can help each other, and tasks across benchmarks (*e.g.*, PNR and AR, OSCC and AR) may seem unrelated at first sight. It is interesting to see that our results capture both inter-benchmark

[6] https://github.com/EGO4D.

	$\mathcal{T}_p$ is TTM		$\mathcal{T}_p$ is PNR		$\mathcal{T}_p$ is OSCC	
	# Params $\cdot 10^6$ Trainable (*All*)	mAP (s)	# Params $\cdot 10^6$ Trainable (*All*)	Error (s) ↓	# Params $\cdot 10^6$ Trainable (*All*)	Acc. (%) ↑
TS model [23]	20.2 *(20.2)*	58.91	32.2 *(32.2)*	0.615	32.2 *(32.2)*	68.22
EgoT2-s (Subset of Tasks)	0.7 *(35.3)*	65.89	5.8 *(70.2)*	**0.608**	5.8 *(70.2)*	69.69
EgoT2-s (All Tasks)	0.7 *(51.1)*	**66.54**	6.4 *(132)*	0.610	7.4 *(133)*	**72.69**

Table 8. Results of EgoT2-s when primary task is $\mathcal{T}_p$ is TTM, PNR and OSCC. We compare EgoT2-s that uses a subset of auxiliary tasks with EgoT2-s using all auxiliary tasks. When $\mathcal{T}_p$ is TTM, 'Subset of Tasks' denote TTM and LAM; When $\mathcal{T}_p$ is PNR or OSCC, 'Subset of Tasks' denote PNR and OSCC.

	$\mathcal{T}_p$ is AR			$\mathcal{T}_p$ is LTA		
	# Params $\cdot 10^6$ Trainable (*All*)	Acc. (%) ↑ Verb	Noun	# Params $\cdot 10^6$ Trainable (*All*)	ED@20 ↓ Verb	Noun
TS model [23]	63.3 *(63.3)*	22.18	21.55	180 *(242)*	0.746	0.789
EgoT2-s (Subset of Tasks)	2.4 *(282)*	21.94*	**23.33***	25.0 *(304)*	0.739	0.774
EgoT2-s (All Tasks)	4.3 *(130)*	**23.04**	23.28	41.8 *(348)*	**0.731**	**0.769**

Table 9. Results of EgoT2-s when primary task is $\mathcal{T}_p$ is AR and LTA. 'Subset of Tasks' denote AR and LTA. The results achieved with expanded video length are makred with a *.

and intra-benchmark task relations: (1) when $\mathcal{T}_p$ is PNR, the Transfer of OSCC or AR features yields similar results, achieving the temporal localization error of 0.611 and 0.613 seconds, respectively; (2) when $\mathcal{T}_p$ is OSCC, surprisingly, Transfer (AR) outperforms Transfer (PNR) and a dedicated OSCC model (*i.e.*, Finetuning) by ∼3%; (3) when $\mathcal{T}_p$ is AR or LTA, PNR and OSCC features transfer better to predicting verbs than predicting nouns. We hypothesize that this is because an object state change is dependent on verbs and agnostic to nouns.

We find that the task of action recognition (AR) is very informative in predicting the other 3 tasks; this suggests that similar to common practices in third-person video understanding (*e.g.*, finetuning an action recognition model pretrained on Kinetics to other downstream tasks), the Ego4D AR model can also serve as a good initialization choice for other egocentric video tasks. In addition, from the task definition, PNR and OSCC are more object-centric while AR and LTA focus on human activities. Besides the obvious task relations (*i.e.*, PNR to OSCC, AR to LTA), we uncover connections between tasks belonging to different benchmarks as well. AR task provides information complementary to primary task features and benefits OSCC. PNR and OSCC models convey information that are helpful for classifying verbs in AR and LTA.

For human-human interactions, when the primary task is TTM, the good results achieved by Transfer (LAM) and Transfer (ASD) indicate that both auxiliary tasks provide informative cues for TTM. This also aligns with our intuition that LAM and TTM are very related tasks as people tend to make eye contact when they talk to someone. In addition, when $\mathcal{T}_p$ is ASD, Transfer baseline results indicate that TTM and LAM are detrimental to the ASD task. We conjecture that this may be because the act of someone looking at the camera wearer does not necessarily relate to the fact that this person is the active speaker. In all, we hope our analysis on task relations can facilitate holistic egocentric video understanding.

Varying the Set of Auxiliary Tasks. In Table 2-3 of the main paper we presented results for EgoT2-s (All Tasks), where all tasks within the same cluster of $\mathcal{T}_p$ are adopted as auxiliary tasks. Here we consider the setting where we constrain the auxiliary tasks to be within the same benchmark as $\mathcal{T}_p$. Results of EgoT2-s using a subset of tasks[7] are shown in Table 8-9.

By comparing results of EgoT2-s (Subset of Tasks) and EgoT2-s (All Tasks) in these two tables, we see that there are cases where EgoT2 can effectively leverage synergies between tasks that belong to different benchmarks. For instance, when $\mathcal{T}_p$ is OSCC, since AR features provide beneficial cues, EgoT2-s with all auxiliary tasks outperforms by 3% the EgoT2-s variant that only uses PNR and OSCC features. Conversely, we would expect that the introduction of inter-benchmark auxiliary tasks may cause a detrimental effect when the benchmarks involve dissimilar tasks, for instance, when $\mathcal{T}_p$ is PNR. However, even in such case EgoT2-g (All Tasks) is still on-par with EgoT2-g (Subset of Tasks) and it outperforms all transfer baselines. This suggests that it has strong ability to mitigate negative transfer.

Ablation Study. In Table 4 of the main paper, we provided an ablation study of EgoT2-s when the primary task is TTM to validate our design choices. Here, we conduct another set

[7]We exclude ASD here since there is no other task from the same benchmark as ASD (see Table 1 in the main paper).

	# Params $\cdot 10^6$ Trainable (*All*)	Auxiliary Tasks	Temporal Information	Frozen TS model	mAP (%) ↑
(a)	8.9 *(105)*		✓	✓	69.68
(b)	7.4 *(133)*	✓		✓	71.65
(c)	133 *(133)*	✓	✓		72.22
(d)	7.4 *(133)*	✓	✓	✓	72.69

Table 10. Ablation study of EgoT2-s ($\mathcal{T}_p$ is OSCC).

Acc. (%)	SlowFast	EgoVLP
TS Model	68.22	73.00
EgoT2-s	72.69	75.77

Table 11. Experiments with the TS model being SlowFast and EgoVLP when $\mathcal{T}_p$ is OSCC. By resorting to auxiliary task information, EgoT2-s demonstrates further performance improvements.

of ablation studies for the case when $\mathcal{T}_p$ is OSCC. The results are summarized in Table 10. The results are consistent with those reported in Table 4. The three components (*i.e.*, the introduction of auxiliary tasks, preserving temporal information and freezing TS backbones) work together and contribute to the efficacy of EgoT2-s.

Experiments with a different TS backbone. In the experiments presented in the main paper, we selected as TS backbones, the baseline models of Ego4D in order to facilitate comparison with prior work and to demonstrate the ability of our approach to achieve state-of-the-art results with simple network designs. However, EgoT2-s provides a flexible framework that can incorporate any advanced architecture. Here we demonstrate this flexibility by replacing the I3D ResNet-50 backbone with a video transformer used in EgoVLP [40] for the case when $\mathcal{T}_p$ is OSCC. We report results in Table 11. We find that the improvement brought by auxiliary task information (*i.e.*, AR in this case) is orthogonal to architecture advances and pretraining techniques. EgoT2-s can further improve the EgoVLP model performance by 2.77%.

Comparison of EgoT2-s and EgoT2-g. We provide a side-by-side comparison of our proposed two variants of EgoT2 over the TS model in Figure 7. As discussed in Sec. A.2.2, LTA has two metrics (accuracy for future 1 timestamp and edit distance for future 20 timestamps). Since EgoT2-s is optimized towards long-term predictions and EgoT2-g is trained to make one-step predictions, EgoT2-s does not perform as well as EgoT2-g in terms of LTA verb and noun accuracy, and ED@20 is not computable for EgoT2-g. In general, EgoT2 achieves great performance gains over the TS models across tasks, and EgoT2-s leads to top performance. Moreover, Table 12 compares the number of trainable parameters and multiply-accumulate operations required for EgoT2-s and EgoT2-g. For EgoT2-s, we sum the trainable parameters (computations) of all task translators within

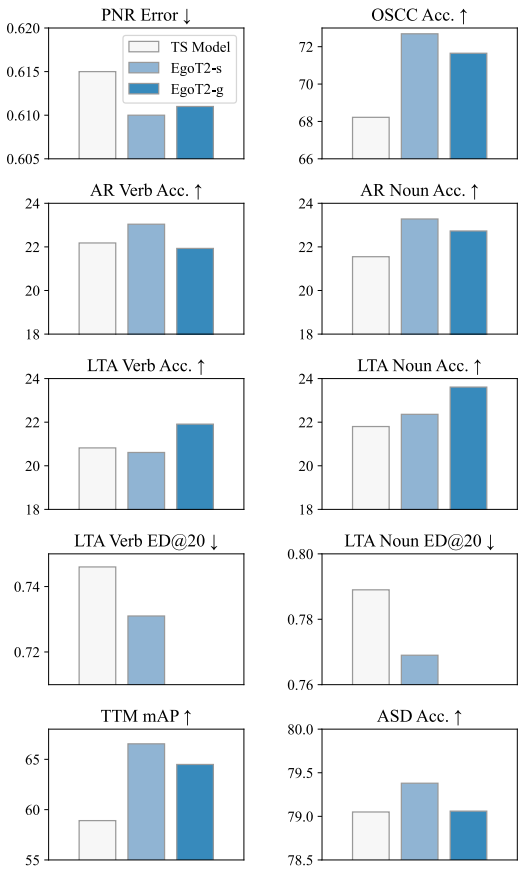

Figure 7. Performance comparison of two variants of EgoT2 with the TS models on 6 tasks. EgoT2 leads to great improvement over the TS model and EgoT2-s achieves top performance.

	Human-Object Tasks # Params	Human-Object Tasks # MACs	Human-Human Tasks # Params	Human-Human Tasks # MACs
Sum of EgoT2-s	2.2	1802.5	59.9	386.6
EgoT2-g	1.4	1803.6	34.5	386.2

Table 12. Efficiency comparison of two variants of EgoT2. We report the number of trainable parameters (in millions) and the multiply-accumulate operations (MACs, in billions) required for one forward pass. Compared with a set of EgoT2-s models developed for each task, EgoT2-g has fewer trainable parameters and similar computational costs.

one cluster. EgoT2-g shares the task translator across tasks within one cluster and hence saves parameters. The computational costs of EgoT2-s and EgoT2-g are similar, as the majority of the computation lies in the task-specific backbones, which are identical in both variants.

EgoT2-g across Task Clusters. In Table 5 of the main paper, we presented separate results of EgoT2-g on the cluster of human-human interaction (HHI) tasks and the cluster of human-object interaction (HOI) tasks due to the sub-

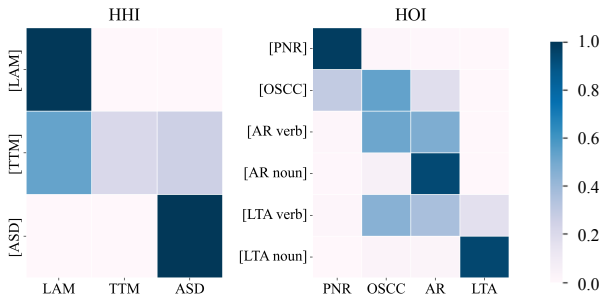

Figure 8. Average encoder-decoder attention weights of EgoT2-g. The heatmaps illustrate how task-specific feature tokens (x axis) contribute to the task of interest (y axis) in task translation.

stantial domain gap between hese two clusters of Ego4D videos. Take Figure 4 as an example, videos from HOI task datasets (upper figure) only capture human-object interactions and do not have other people in the scene. Thus, no HH interactions would be detected in these HOI videos, and as such we would not expect HHI features to contribute to HOI tasks. To verify this hypothesis, we implement a cross-cluster EgoT2-g that attends to two HOI tasks (PNR and OSCC) and one HHI task (LAM) simultaneously and report the results in Table 13. The cross-cluster EgoT2-g yields similar performance with intra-cluster EgoT2-g.

	PNR Error (s) ↓	OSCC Acc. (%) ↑	LAM mAP (%) ↑
EgoT2-g (intra-cluster)	0.612	68.6	77.63
EgoT2-g (cross-cluster)	0.611	68.3	77.56

Table 13. Results of EgoT2-g across 2 task clusters. Due to the domain gap between human-human interaction tasks and human-object interaction tasks, EgoT2-g (cross-cluster) does not lead to further improvement compared with the EgoT2-g variant trained within the same task cluster.

A.4. Additional Visualizations

Finally, Figure 8 shows encoder-decoder attention weights of the last layer transformer produced by EgoT2-g for 3 human-human interaction (HHI) tasks and 6 human-object interaction (HOI) tasks. The attention weights of task-specific tokens are temporally pooled into one token and averaged over all validation video data. x axis are different task tokens and y axis corresponds to task prompts. Note that in Figure 1 in the main paper, we average the attention weights of verb and noun for AR and LTA and visualize the resulting 4×4 matrix. Figure 8 reveals inherent task relations and provides an intuitive illustration of how the task-general translator utilizes task tokens differently conditioned on the task of interest (*i.e.*, task prompt). In the left figure, we observe that LAM and ASD features have large attention weights when the task prompt is TTM, indicating that EgoT2-g effectively utilizes the two relevant tasks to improve TTM predictions. On the contrary, when the task prompt is ASD, ASD tokens are largely activated while non-beneficial LAM and ASD tokens are rarely adopted in task translation. This demonstrates that EgoT2-g learns to selectively activate task tokens to mitigate the issue of negative transfer. In the right figure, AR task tokens are more activated given that the task prompt is OSCC rather than PNR. This aligns with our previous finding in EgoT2-s that AR features are beneficial for the OSCC task. Also, when the task of interest is predicting nouns (*i.e.*, task prompt is AR noun or LTA noun), attention weights of PNR and OSCC are very small, which indicates that the two task features do not help in noun prediction. The conclusion is also consistent with EgoT2-s.