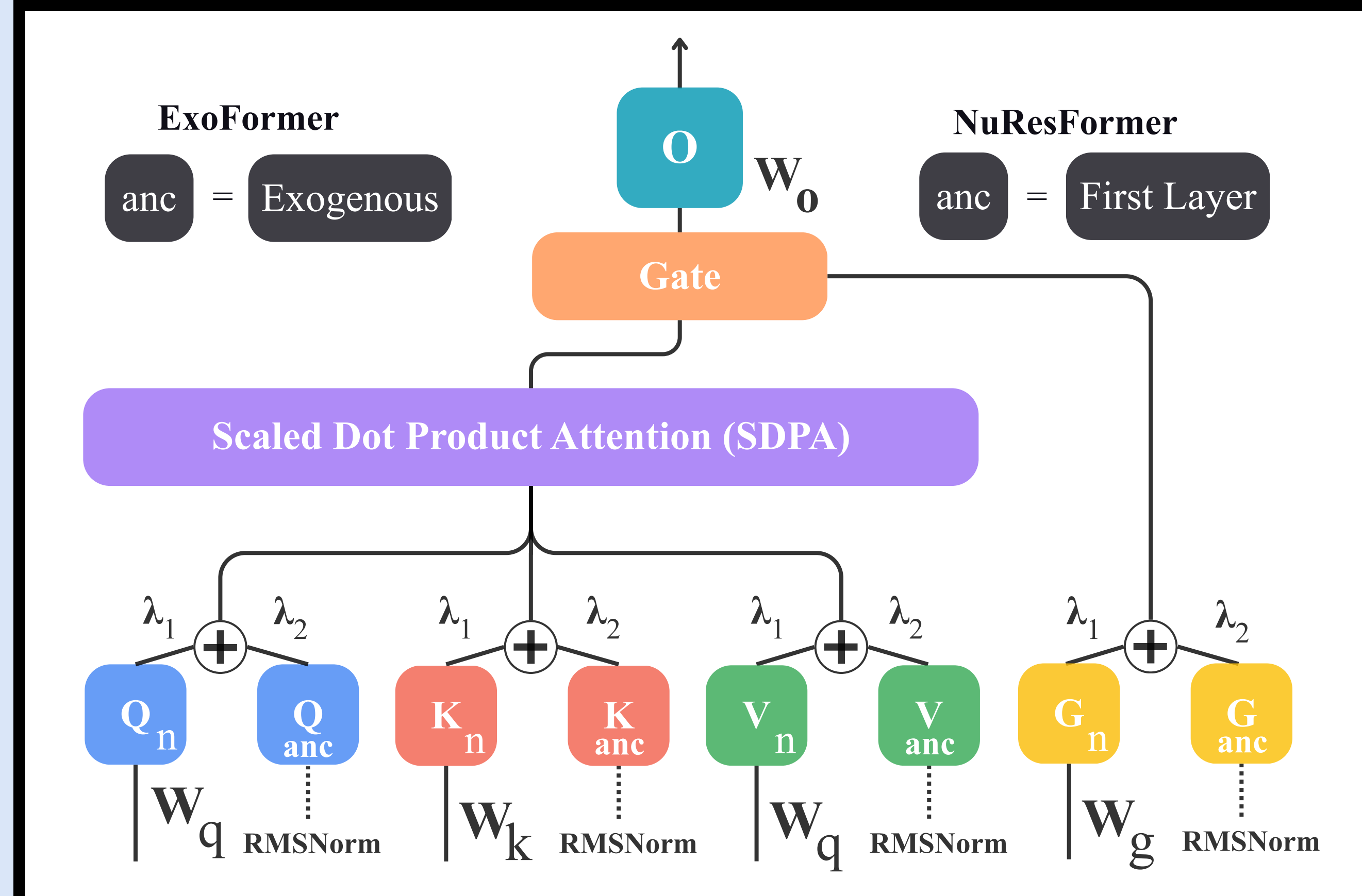
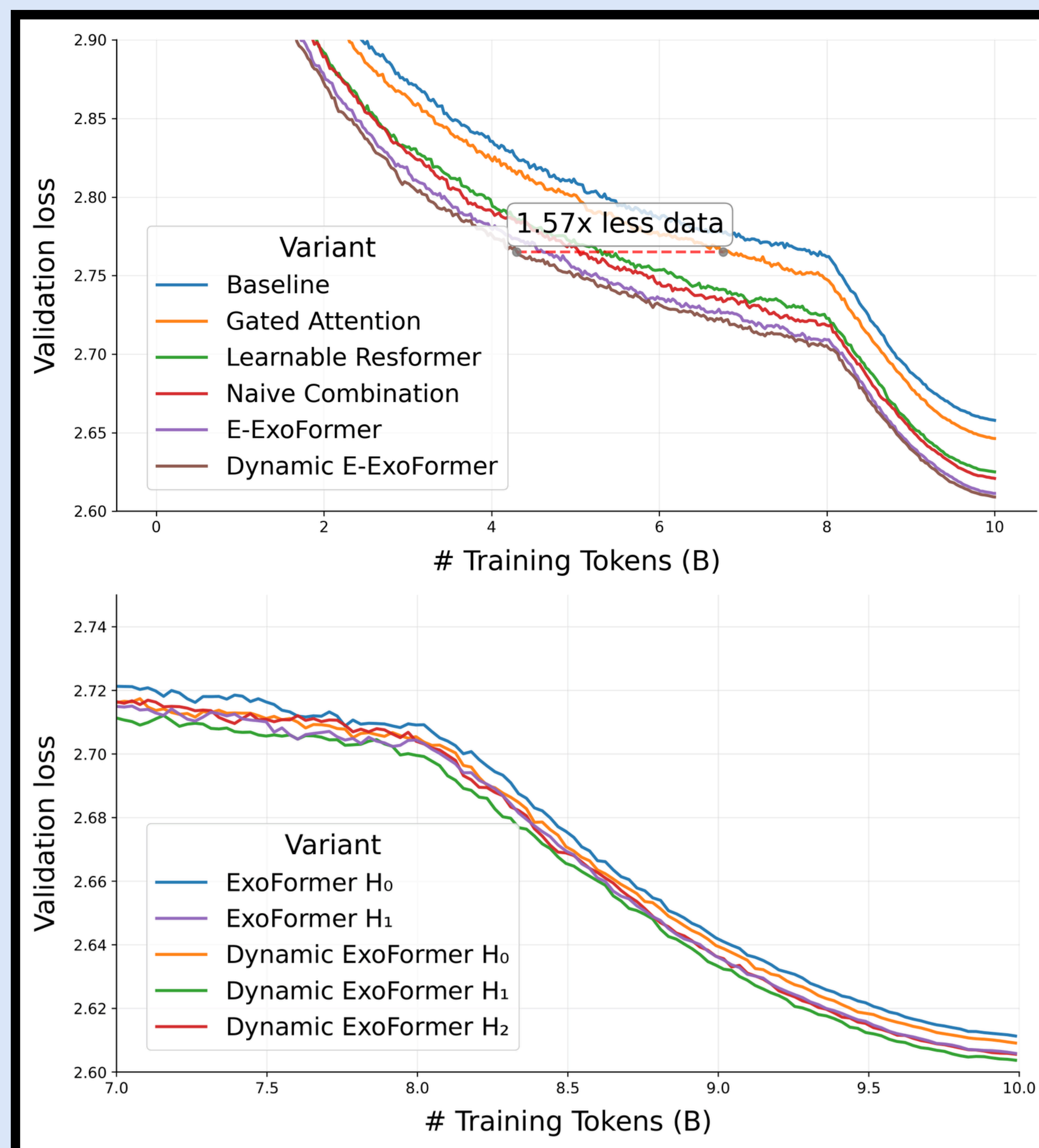


Method



Findings

1. Mixing residual connections across all attention pathways (queries, keys, values, and gating logits) yields gains beyond value residuals alone.
2. RMSNorm on anchor sources stabilizes query/key mixing, resolving the distributional mismatch that previously made Q/K residuals diverge.
3. Exogenous anchors decouple identity preservation from layer-wise computation, improving both perplexity and downstream accuracy with only ~1.3% parameter overhead.
4. Anchoring on H_1 (post-first-layer) rather than H_0 (input embeddings) further improves perplexity while reducing mixing overhead.



Results

MODEL	ARC-C	ARC-E	HELLA.	OBQA	PIQA	WINO.	AVG. ACC	PPL
BASELINES								
BASE TRANSFORMER	30.97	63.38	42.02	33.60	67.30	51.54	48.14	14.79
GATED ATTENTION	30.89	64.69	42.73	34.00	67.14	53.35	48.80	14.64
RESFORMER (VALUE RESIDUAL)	33.62	64.86	43.49	34.40	68.88	52.64	49.65	14.32
NAÏVE COMBINATION	32.94	64.52	43.47	32.60	68.34	52.64	49.09	14.25
INTERNAL ANCHOR								
E-NURESFORMER (ONLY Q/K NORMS)	31.74	64.27	43.83	33.00	68.23	52.41	48.91	14.21
H-NURESFORMER (ONLY Q/K NORMS)	34.73	64.06	44.45	34.20	67.57	53.04	49.68	14.21
S-NURESFORMER (ONLY Q/K NORMS)	32.85	66.08	43.87	34.20	68.28	53.67	49.83	14.22
E-NURESFORMER (NO NORM, {V,G})	32.42	64.94	43.80	35.00	67.52	51.54	49.20	14.24
H-NURESFORMER (NO NORM, {V,G})	31.23	64.18	43.50	34.00	67.68	54.06	49.11	14.22
S-NURESFORMER (NO NORM, {V,G})	31.83	64.44	43.66	34.80	68.44	52.49	49.28	14.24
E-NURESFORMER	33.70	65.07	44.23	33.60	68.34	53.12	49.68	14.15
H-NURESFORMER	33.79	65.11	44.09	33.40	67.41	52.72	49.42	14.17
S-NURESFORMER	32.42	64.02	43.93	33.20	67.90	53.20	49.11	14.17
EXTERNAL ANCHOR								
E-EXOFORMER (NO NORM)	32.08	63.76	43.49	34.40	68.77	55.64	49.69	14.30
DYNAMIC E-EXOFORMER	33.36	65.87	44.54	34.40	68.28	55.17	50.27	14.09
E-EXOFORMER	32.42	64.65	44.28	36.40	67.74	53.59	49.85	14.13
H-EXOFORMER	32.17	65.87	44.00	32.40	68.23	52.72	49.23	14.14
S-EXOFORMER	32.17	65.66	44.33	33.60	68.06	51.14	49.16	14.15

