Supplementary Materials for

Scene-Conditioned General Gaze Following in Static Images with Uncertainty-Preserving Geometry

Metric conventions. Development-validation AUC denotes the mean per-instance ROC-AUC computed at the native 56 × 56 heatmap resolution using one annotated gaze-target location per development instance. This single-annotation metric is used for development-stage model selection and sensitivity analysis and is distinct from both the pooled/global AUC (5 × 5) and the original-resolution per-image AUC used in the official GazeFollow test. Development Dist and Ang are measured against the same single annotated target. Unless otherwise noted, development analyses use the image-disjoint subset of 12,569 instances.

Supplementary Table S1. Development-Stage Control of Decoder-Side Geometry Re-injection.

This development control tests whether the Full-18 probabilistic geometry representation needs to be injected again immediately before the heatmap decoder after already participating in 21-channel image-level early fusion. The two configurations use the same CDDL–PGGM representation, GCViT-T/FPN backbone, initialization, optimizer, training budget, and development-validation selection rule; they differ only in the second geometry pathway.

Configuration	Selected epoch	Development-validation AUC	Dist	Ang
Dual geometry injection	2	0.934566	0.146930	18.117°
Early-only geometry fusion	2	0.934498	0.146903	18.145°
Early-only − dual	–	−0.000067	−0.000026	+0.028°

Note: Early-only geometry fusion follows RGB + Full-18 → 21-channel GCViT-T → FPN → heatmap decoder. Dual geometry injection uses the same input-level fusion and additionally projects Full-18 geometry into a 32-channel decoder-side feature before heatmap decoding. Reported deltas are Early-only minus Dual. The results are numerical development evidence only and are not interpreted as a statistical-equivalence test.

Supplementary Table S2. Standardized Paired Effects and Instance-Level Benefit Coverage

For five principal official-test comparisons, paired standardized mean differences were calculated from 4,782 aligned GazeFollow person instances as the mean signed within-instance difference divided by its sample standard deviation (d_z). For original-resolution per-image AUC, differences were defined as candidate minus reference; for localization errors, they were defined as reference minus candidate, so positive effects indicate improvement. Improved, unchanged, and degraded proportions used a numerical equality tolerance of 10^{-12}. The standardized effect sizes complement the paired bootstrap confidence intervals and Holm-adjusted Wilcoxon tests reported in Table 2b.

Controlled contrast	Metric	Mean benefit	d_z	Improved / equal / degraded
DCLD vs. reference	Dist	0.002757	+0.0620	36.70% / 34.13% / 29.17%
DGQI vs. static DCLD	Original-resolution per-image AUC	0.000515	+0.0864	53.26% / 0.02% / 46.72%
DGQI vs. static DCLD	Dist	0.000061	+0.0017	19.70% / 59.14% / 21.16%
STMRP vs. BCE	Original-resolution per-image AUC	0.001010	+0.1055	61.33% / 0.00% / 38.67%
STMRP vs. BCE	Dist	-0.002071	-0.0441	24.61% / 46.51% / 28.88%

Note: Most standardized effects are small, even when Holm-adjusted p-values are very small. Paired instance-level effect sizes apply to original-resolution per-image AUC and peak-based localization errors; pooled/global AUC is computed after concatenating the full test set and has no corresponding per-instance d_z. For the DGQI-versus-static-DCLD contrast, exactly one person instance had numerically unchanged original-resolution per-image AUC.

Supplementary Table S3. Three-Seed Development Robustness of STMRP

The STMRP training-objective effect was evaluated over three development seeds using the same DGQI + DCLD inference architecture and the same fixed weight of 0.10. The table reports paired changes from BCE-only to STMRP-regularized training. Because only three seeds were examined, these results are descriptive robustness evidence rather than an inferential significance analysis.

Seed	Δ development-validation AUC	Δ Dist	Δ Ang
20260814	+0.000877	+0.000846	+0.058°
20260815	+0.000794	+0.001477	−0.083°
20260816	+0.000542	+0.001350	−0.090°
Mean ± sample SD	**+0.000737 ± 0.000175**	**+0.001224 ± 0.000334**	**−0.038° ± 0.083°**

Note: **Δ** = STMRP-regularized minus BCE-only DGQI + DCLD on the development-validation subset. Positive AUC is favorable; positive Dist indicates a larger localization error; negative Ang is favorable. All three seeds show an AUC increase and a Dist increase, reproducing the discrimination–localization trade-off observed in the official-test comparison.

Supplementary Table S4. Epoch-Matched Development Comparison of STMRP and BCE-Only Training

A matched-epoch comparison examined whether the discrimination–localization trade-off between BCE-only and STMRP-regularized training could be explained solely by different selected training durations. For each of three development seeds, we compared the two training objectives at epoch 6 while keeping the DGQI + DCLD inference architecture unchanged.

Seed	Epoch	Δ development-validation AUC	Δ Dist	Δ Ang (°)
20260814	6	+0.001113	+0.001551	+0.017
20260815	6	+0.001097	+0.001368	+0.031
20260816	6	+0.000869	+0.001495	-0.007

Note: Differences are STMRP minus BCE-only at epoch 6. Development-validation AUC and Dist increased for all three seeds, whereas Ang showed mixed directions. These results support a matched-epoch development-stage trade-off, but they do not constitute an epoch-matched re-evaluation of the four separately trained official-test models.

Supplementary Table S5. Development Sensitivity to the Number of Radial Candidates

The direction–distance candidate grid was evaluated with 18 inherited direction groups and $R \in \{8, 12, 16\}$ radial bins (144, 216, and 288 candidates). All configurations used the same image-disjoint GazeFollow development-validation subset (n = 12,569), random seed 20260814, BCE-only DGQI + DCLD architecture, and a maximum training budget of eight epochs. Model selection was based on development-validation AUC, with Dist and Ang serving only as exact-tie criteria.

Radial bins R	Candidates	Epoch	Development-validation AUC ↑	Dist ↓	Ang (°) ↓
8	144	4	0.933622	0.145263	18.092
12	216	5	0.933874	0.145287	18.019
16	288	5	0.933683	0.145051	18.040

Note: R = 12 achieved the highest development-validation AUC among the three single-seed configurations, whereas R = 16 achieved a slightly lower Dist. These small descriptive differences do not establish statistical superiority or guarantee stability across random seeds or backbone architectures.

Supplementary Table S6. STMRP Loss-Weight Sensitivity

The STMRP regularization weight λ was evaluated at 0, 0.05, 0.10, and 0.20, holding R = 12, the 3 × 3 protected target neighborhood, the frozen reference model, random seed 20260814, the image-disjoint development split, and the eight-epoch training budget fixed. The BCE-only setting ($\lambda = 0$) and the main-study setting ($\lambda = 0.10$) were drawn from the initial development experiments; $\lambda = 0.05$ and $\lambda = 0.20$ provide additional local sensitivity evidence.

Weight λ	Epoch	Development-validation AUC ↑	Dist ↓	Ang (°) ↓
0 (BCE only)	5	0.933874	0.145287	18.019
0.05	6	0.934789	0.145550	18.014
0.10 (main study)	6	0.934750	0.146133	18.077
0.20	8	0.934698	0.146476	18.123

Note: All three positive weights improved development-validation AUC relative to BCE-only training, while larger ranking weights generally increased Dist. In this single-seed comparison, $\lambda = 0.05$ produced marginally higher development-validation AUC and lower Dist than $\lambda = 0.10$. The main-study model retains $\lambda = 0.10$; the additional sensitivity configurations do not replace the models used for official-test evaluation.

Supplementary Table S7. Sensitivity to the Protected Target Neighborhood

With $\lambda = 0.10$ and all other development settings fixed, the target-centered exclusion neighborhood in STMRP was evaluated at 1 × 1, 3 × 3, and 5 × 5. The primary framework uses the 3 × 3 setting. Under the same development-validation AUC selection rule, all three configurations selected epoch 6.

Excluded window	Radius	Epoch	Development-validation AUC ↑	Dist ↓	Ang (°) ↓
1 × 1	0	6	0.934755	0.146179	18.075
3 × 3 (main study)	1	6	0.934750	0.146133	18.077
5 × 5	2	6	0.934785	0.146209	18.083

Note: The numerical variation across the three neighborhood sizes is small and does not establish equivalence or statistically significant superiority of 3 × 3. The primary 3 × 3 setting permits local adjustment around the annotated target while applying selective ranking constraints outside the protected region.

Supplementary Table S8. ResNet-50/FPN Backbone Bridge Control for DCLD

To assess representation dependence, a previously trained ResNet-50/FPN geometry-guided reference was frozen and the same static-geometry DCLD architecture was optimized for up to eight epochs. The experiment used the same image-disjoint development-validation split (n = 12,569), random seed 20260814, and development-validation AUC selection criterion. Directional geometry and the reference encoder and decoder were frozen; only the DCLD parameters were trained. The frozen reference reproduced the established ResNet development-validation result.

ResNet-50/FPN condition	Epoch	Development-validation AUC ↑	Dist ↓	Ang (°) ↓
Frozen reference	0	0.925955	0.154185	18.498
Reference + DCLD (selected)	1	0.926036	0.154430	18.561

Note: Under the predefined development selection rule, adding DCLD to the ResNet-50/FPN reference did not yield a localization improvement: development-validation AUC increased by 0.000081, whereas Dist increased by 0.000245 and Ang increased by 0.063°. Other epochs displayed different discrimination–localization trade-offs, but were not substituted for the selected model. This single-seed bridge control does not isolate a purely causal backbone effect and does not demonstrate that DCLD is universally ineffective with ResNet.

Supplementary Method S1. Inference-Efficiency Measurement Protocol

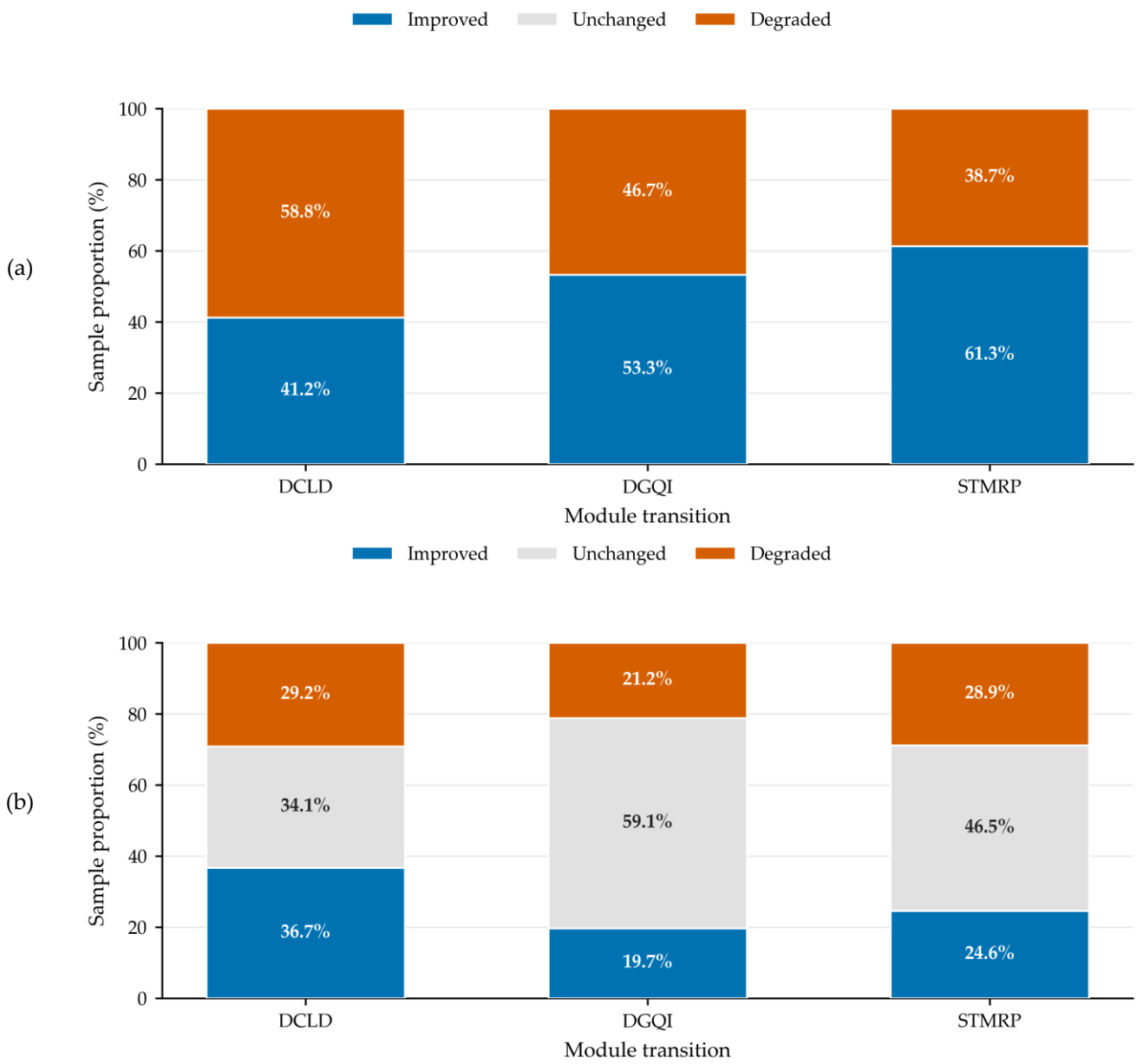
Computational efficiency was measured with synthetic 224×224 FP32 inputs on the same NVIDIA GeForce RTX 5090 D V2 GPU, with autocast and TF32 disabled. Batch size 1 was used for latency; 50 warm-up iterations were followed by five measurement blocks of 100 iterations. Batch size 64 was used for throughput and peak allocated GPU memory; 20 warm-up iterations were followed by five measurement blocks of 20 iterations. Data loading, host-to-device transfer, metric computation, and CPU post-processing were excluded. Latency and throughput in Table 5 are reported as mean (sample standard deviation) over the five measurement blocks. Peak memory is CUDA maximum allocated memory during batch-64 forward inference after the model and fixed input tensors are resident on the GPU.

Supplementary Note S1. Secondary DGQI Coupling Diagnostics

DGQI produces a strong representation-level change, but the magnitude of candidate refinement is not monotonically coupled to the final-output gain. Across the development-validation subset, the Spearman correlation between candidate-distance gain and the net development-validation AUC gain from adding DGQI to DCLD is $\rho = 0.0082$. The correlation between DCLD correction magnitude and the development-validation AUC gain from DGQI is $\rho = -0.0097$. With DGQI active, the selected development model has a mean context-vector norm of 8.5246 and a mean absolute decoder correction of 0.2509. These diagnostics support a nonlinear interpretation in which DGQI robustly changes candidate evidence, while DCLD determines how much of that change is expressed in the final heatmap; they do not establish a causal relation between candidate-distance improvement and final AUC gain.

Supplementary Figure S1. Sample-Level Benefit Coverage of the Controlled Transitions

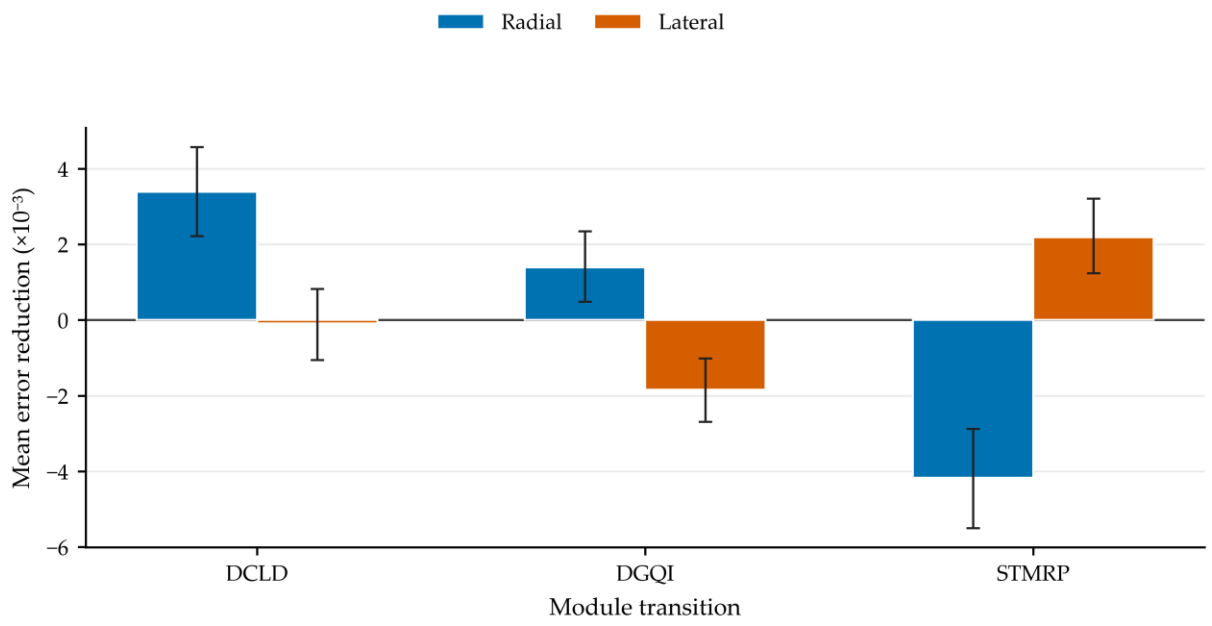
Across the 4,782 official GazeFollow test instances, DCLD with static geometry improves per-image AUC on 41.2% of samples and degrades 58.8%; for Dist, 36.7% improve, 34.1% are unchanged, and 29.2% degrade. Adding DGQI to DCLD broadens the discrimination benefit: per-image AUC improves on 53.3% and degrades on 46.7%, whereas Dist is unchanged for 59.1%, improves for 19.7%, and degrades for 21.2%. STMRP-regularized training yields the broadest per-image AUC coverage, improving 61.3% and degrading 38.7%; for Dist, 24.6% improve, 46.5% are unchanged, and 28.9% degrade. These proportions show that the paired effects summarized in Table 2b are not produced by uniform shifts across all samples.



Supplementary Figure S1. Sample-level benefit coverage of the controlled transitions. (a) Original-resolution per-image AUC: proportions of improved, unchanged, and degraded instances for DCLD, DGQI, and STMRP. (b) Dist: corresponding proportions for the same three transitions.

Supplementary Figure S2. Radial–Lateral Localization-Error Changes for the Controlled Transitions

Prediction error relative to the mean human annotation was decomposed into a radial component parallel to the eye-to-target direction and a lateral component orthogonal to that direction. Positive values denote reductions in absolute error relative to the preceding controlled configuration. DCLD reduces radial absolute error by +0.00339 (95% bootstrap CI: [+0.00222, +0.00458]) while showing no supported lateral change (−0.00009; confidence interval crossing zero), indicating that its localization gain is primarily radial. DGQI reduces radial error by +0.00139 (95% CI: [+0.00048, +0.00235]) but increases lateral error (error-reduction effect −0.00184; 95% CI: [−0.00269, −0.00101]). STMRP shows the reverse pattern: radial error increases (error-reduction effect −0.00417; 95% CI: [−0.00550, −0.00287]), whereas lateral error decreases by +0.00218 (95% CI: [+0.00124, +0.00321]).



Supplementary Figure S2. Radial and lateral localization-error changes for the controlled transitions. Positive values indicate error reduction; negative values indicate error increase. Error bars denote 95% bootstrap confidence intervals.

Supplementary Table S9. Coverage-Limited Zero-Shot ChildPlay Results

The evaluation covers 14,332 inside-visible person-frames from 25 of the 45 official ChildPlay test clips. All four models were trained only on GazeFollow and were fixed before ChildPlay evaluation. Because the locally available data do not cover the complete official test set, these results are supplementary cross-dataset evidence and must not be ranked directly against complete-benchmark ChildPlay results.

(a). Micro person-frame performance on the available subset.

Configuration	ChildPlay AUC ↑	Dist ↓	Short-side Dist ↓	Ang ↓
Geometry-guided reference	0.9493	0.1192	0.1519	**11.99°**
DCLD with static geometry	**0.9512**	**0.1175**	**0.1499**	12.15°
DGQI + DCLD (BCE)	0.9507	0.1182	0.1509	12.28°
Proposed DGQI + DCLD (BCE + STMRP)	0.9496	0.1186	0.1513	12.05°

(b). Clip-cluster AUC contrasts on the available subset.

Controlled contrast	Mean AUC benefit	Clip-bootstrap 95% CI	Holm-adjusted p	Interpretation
DCLD with static geometry vs. reference	+0.001881	[+0.000746, +0.002498]	0.0184	Both summaries support an AUC gain
DGQI + DCLD vs. DCLD with static geometry	−0.000434	[−0.001023, +0.000208]	0.2028	No stable AUC effect
STMRP-regularized vs. BCE-only DGQI + DCLD	−0.001144	[−0.001583, −0.000333]	0.3799	Mean-bootstrap decrease; rank evidence not supported
Proposed vs. reference	+0.000304	[+0.000084, +0.000414]	0.3406	Tiny mean gain; rank evidence not supported

Note: ChildPlay AUC is the dataset-specific per-person-frame tolerant-support AUC. The bootstrap interval summarizes the mean clip-cluster effect, whereas the Holm-adjusted Wilcoxon test is performed on clip means; disagreement between the two is reported explicitly rather than converted into a single binary significance statement. Best values in Table S9a are bolded metric-wise.