OMNI GROUP FUSION: GROUPOID-COMPATIBLE CORRESPONDENCE AND EQUIVARIANT TRANSPORT ACROSS MULTIMODAL SENSOR FRAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni and any-to-any models increasingly ingest text, images, video, audio, depth, and 3D into unified sequence backbones, achieving architectural interface uniformity through unconstrained attention. However, this convenience overlooks a fundamental physical reality: physical sensors and spatial language naturally operate in distinct, uncalibrated, and uncertain local reference frames. Treating directional and geometric quantities as flat, interchangeable scalars forces backbones to memorize pose-dependent statistical co-occurrences, precipitating sample inefficiency, cross-modal gradient starvation, and fragility under novel viewpoints or clock drift—while heuristic point-pose canonicalization breaks discontinuously under physical ambiguity. We introduce **Omni Group Fusion (OGF)**, which grounds multimodal fusion on an *observation groupoid* respecting independent local coordinate charts. OGF operationalizes four transformation-consistent primitives: (i) routing tokens to shared world anchors via group-invariant descriptors, preventing correspondence drift under sensor rotations; (ii) maintaining multi-hypothesis spatial and temporal posteriors to represent frame ambiguity and clock drift without discontinuous collapse; (iii) equivariantly transporting typed representations ($\ell = 1$ vectors, $\ell = 2$ tensors, coordinates) into a common world frame *before* pooling; and (iv) adjudicating consensus to attenuate conflicting sensors, updating anchors via Clebsch–Gordan couplings and reading back into frozen backbones through zero-initialized residuals. We theoretically establish global equivariance, local-gauge independence, stabilizer well-definedness, posterior composition consistency, and streaming stability. Across a controlled physical world, real continuous sensors, and five frozen generative backbones (3B–8B), factorial ablations isolate different claims: removing the group returns chance; typed transport, not orbit coverage, is what continuous sensors require; and, on Qwen2.5-Omni-7B, breaking sensor pairing removes routed synergy on four of the six audio–visual benches.

1 INTRODUCTION

Modern any-to-any and Omni systems have expanded multimodal coverage across text, vision, audio, depth, and 3D within unified sequence backbones (Tang et al., 2023; Wu et al., 2024; Zhan et al., 2024; Tang et al., 2024; Xie et al., 2025; Wu et al., 2025; Ma et al., 2025b; Qu et al., 2025; Fang et al., 2025; Xu et al., 2025). Their fusion primitive, however, predominantly remains learned sequence similarity over flat, untyped embeddings. This achieves an engineered interface uniformity, but sidesteps the physical nature of multimodal observation: *architectural unification* $\neq$ *representational unification* $\neq$ *transformation-consistent fusion*.

The limitation of current architectures is not an inability to fit multimodal training sets. Given massive parameter scale and pretraining data, unconstrained attention readily fits statistical correspondences across diverse streams. The core defect lies in the *inductive bias* of flat token mixing. Physical sensors and spatial language reside in distinct, uncalibrated local reference frames. Consider a scene where an agent perceives a table lamp: a camera displacement vector, an acoustic arrival direction from a sound source, a depth surface normal, and the natural language phrase "behind the lamp" describe aspects of a single physical reality across four separate reference frames. Because flat tokens carry no declared transformation laws, cross-attending or concatenating them treats fundamentally incommensurable

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

geometric quantities as interchangeable scalars. Without transformation-consistent constraints, the model is forced to expend enormous capacity memorizing how coordinates correlate across specific training poses. When evaluated under novel camera rotations, uncalibrated sensor orientations, or clock drift, this memorized association degrades: attention mechanisms suffer from rank collapse, dominant modalities starve the gradients of weaker streams, and the model produces sensory hallucinations. Heuristic workarounds cannot resolve this: camera-pose conditioning presumes known extrinsics, and point-pose canonicalization breaks discontinuously when physical symmetries or occlusions create pose ambiguity.

Real multimodal fusion demands a principled foundation respecting independent local charts while enabling coherent reasoning in a shared world. We introduce **Omni Group Fusion (OGF)**, which grounds multimodal fusion in the mathematics of an *observation groupoid*. Rather than forcing all modalities into a single rigid group or collapsing them into untyped vectors, OGF respects independent local frames ($g'_{j\leftarrow i} = u_j g_{j\leftarrow i} u_i^{-1}$) and operationalizes four core physical decisions: **(i) Invariant candidate correspondence**: group-invariant descriptors route cross-modal features to shared world anchors, guaranteeing candidate matching cannot drift under sensor rotations. **(ii) Multi-hypothesis spatial and temporal posteriors**: rather than forcing a fragile point pose, discrete posteriors represent uncertain relative transformations and temporal warps, capturing physical ambiguity smoothly without discontinuous collapse. **(iii) Equivariant transport before pooling**: directional and geometric features ($\ell = 1$ vectors, $\ell = 2$ tensors, coordinates) are transformed into the world frame via posterior expectation *before* aggregation, ensuring commensurate physical combination. **(iv) Consensus adjudication and Clebsch–Gordan coupling**: modalities disagreeing in the world frame are attenuated to guard against sensor corruption, while typed anchor features interact through symmetry-preserving Clebsch–Gordan couplings (Thomas et al., 2018) and read back into frozen backbones through zero-initialized residuals without overwriting open-world knowledge.

The empirical question is whether this physical contract is *necessary* once ordinary mixing, latent resampling, matched adapters (Hu et al., 2022), and data augmentation receive a fair budget. We formulate fusion success operationally: a multimodal system demonstrates genuine fusion only if it achieves strictly positive synergy over all independently trained modality subsets ($\mathrm{Syn} > 0$, §3.2) *and* predictably collapses when cross-modal pairings, coordinate frames, or clocks are broken. Section 6 evaluates this in three stages: (1) a controlled physical world decomposing group action versus optimal transport; (2) real continuous sensor features under camera rotation, cyclic audio-visual drift, and 3D spatial referring; and (3) lightweight residual fusion modules (sidecars) across five frozen generative backbones (3B–8B). Factorial ablations, geometric order truncations ($\ell = 0, 1, 2$), and pairing probes isolate the mechanism: on the controlled world, removing the group returns chance; on continuous sensors, typed transport, not orbit coverage, is required; on Omni-7B, pairing shuffle removes routed synergy on four of six audio–visual benches.

Contributions.

1. **Observation groupoids for multimodal fusion.** We cast fusion across uncalibrated sensor frames as computation on an observation groupoid and prove global equivariance, local-gauge independence, stabilizer well-definedness, posterior composition consistency, and streaming stability with state linear in the number of anchors.
2. **A transformation-consistent fusion architecture.** Invariant routing to world anchors, multi-hypothesis spatial and temporal frame posteriors, typed transport before pooling, and consensus adjudication with Clebsch–Gordan coupling form a residual sidecar that reads back into frozen generative backbones through a zero-initialized path.
3. **A fusion criterion and its mechanism.** Fusion must beat every independently trained modality subset and collapse when pairing, frames, or clocks break; frozen trunks use a routed variant. Controlled factorial, typed-order, and power-set ablations show that group transport, not matching alone, unlocks synergy, that flat fusion fails by coupling-rank collapse and gradient starvation, and that frame posteriors stay calibrated under ambiguity and accurate under conflict.
4. **Evidence from real sensors to generative Omni models.** On frozen real features, typed transport is exact under continuous camera rotation and carries 3D referring to unseen viewpoints where augmentation fails; temporal posteriors keep event detection under clock shifts; post-transport agreement flags cross-sensor conflict; and moment retrieval improves. On five frozen generative trunks with a nested mixture, pairing shuffle largely removes routed audio–visual synergy, and the sidecar leads a slightly larger LoRA on most benches as training data grow.

2 Related Work

OGF changes the fusion law, not the codec or decoder. Any-to-any and Omni models broaden coverage (Tang et al., 2023; Wu et al., 2024; Zhan et al., 2024; Chameleon Team, 2024; Xie et al., 2025; Wu et al., 2025; Xu et al., 2025); concurrent benches target any-to-any interleaving, omni-language QA, and cross-modal direction consistency (Li et al., 2026b; 2025c; Wang et al., 2026b), whereas OGF targets covariance under world transformations. Cross-attention, Perceiver, Flamingo, and Q-Former (Jaegle et al., 2021; Alayrac et al., 2022; Li et al., 2023) are untyped under sensor-frame changes. Equivariant nets supply laws on one known group (Cohen et al., 2019; Thomas et al., 2018; Fuchs et al., 2020; Satorras et al., 2021; Bronstein et al., 2021); frame averaging and canonicalization (Puny et al., 2022; Kaba et al., 2023) act on a single input under known symmetry. Neither resolves correspondence across several uncertain frames that coexist with open semantics. Entropic OT and Kabsch (Cuturi, 2013; Kabsch, 1976) supply the concrete estimator of the OGF frame posterior, not a competing fusion law.

3 Problem Formulation

A camera vector, an acoustic direction, a surface normal, and the phrase "behind the lamp" describe one physical relation in four uncalibrated frames. This section defines those frames, the arrows between them, and the chart-change law that makes the four quantities comparable.

3.1 Observations, local frames, and typed features

Let a latent world state be $w \in \mathcal{W}$. Modality m produces $x_m \sim p_m(x_m \mid w, c_m)$, where c_m includes its sensor frame, clock, sampling process, resolution, visibility, and nuisance variables. We make no assumption that two raw observation spaces are isomorphic.

Definition 1 (Observation groupoid)**.** *An observation groupoid $\mathcal{G}$ has local sensor or event frames $\mathcal{F}_i$ as objects. A morphism $g_{j\leftarrow i} : \mathcal{F}_i \to \mathcal{F}_j$ exists when a spatial relation between the two frames is meaningful; composition is defined only for compatible arrows. A change of local coordinate convention is a gauge transformation at an object. Temporal alignment is represented separately by a posterior over offsets, rate changes, or local warps, avoiding the assumption that spatial and temporal nuisance variables form one universal group.*

Under local chart changes u_i and u_j, the same physical relation is represented by

$$g'_{j\leftarrow i} = u_j \, g_{j\leftarrow i} \, u_i^{-1}. \tag{1}$$

A valid fusion operator is independent of the source chart and covariant only in its target chart. A groupoid, not a single ambient group, is the right object: arrows that do not exist cannot be filled in by $\mathrm{SE}(3)$.

A token or object/event slot is

$$\tau_i = \left(m_i, s_i, \{h_i^{(\ell)}\}_{\ell=1}^{L}, p_i, t_i, r_i, \eta_i \right), \tag{2}$$

where $s_i \in \mathbb{R}^{d_s}$ is semantic content that transforms trivially under *geometric frame changes only*; $h_i^{(\ell)} \in \mathbb{R}^{c_\ell} \otimes V_\ell$ is a typed feature; p_i, t_i are local space/time coordinates; r_i contains semantic and structural reliability; and η_i is an entity/event identity distribution. We use $L = 2$: $\ell = 1$ for directions, velocities, normals, and displacements, $\ell = 2$ for anisotropy and shape.

3.2 What counts as successful fusion?

For modalities M, define performance $P(A)$ for every independently trained strict subset $A \subsetneq M$. The fusion synergy is

$$\mathrm{Syn}(M) = P(M) - \max_{A \subsetneq M} P(A). \tag{3}$$

A model establishes fusion only if synergy is positive under a pre-specified confidence interval and if breaking pairings by shuffling, swapping, wrong frames, or time misalignment causes the predicted degradation. Masking a modality only at test time is insufficient because the full model may never have learned an optimal subset predictor. Generative-trunk experiments later report a routed variant $\mathrm{Syn}_{\text{route}}$, because the frozen trunk always sees every stream.

4 OMNI GROUP FUSION

The architecture implements the four decisions of §1 under the chart-change law (1). Two later instantiations share these equations: a small head on frozen encoder features, and a zero-initialized sidecar inserted once into a frozen generative trunk (fig. 1).

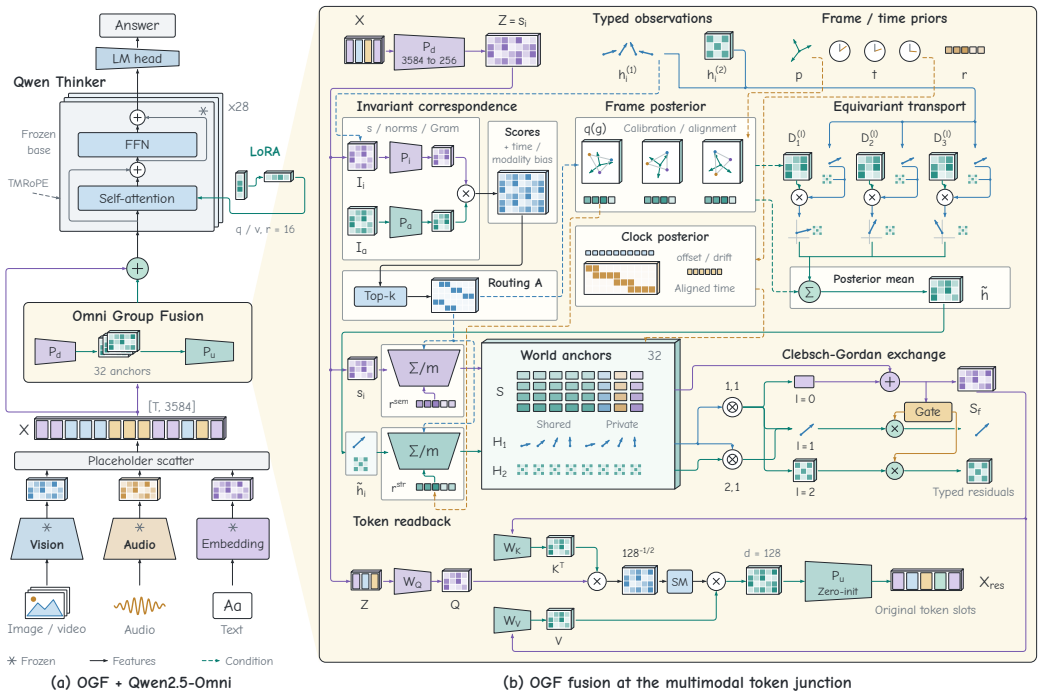


Figure 1: Residual sidecar on a frozen trunk: invariant routing, a frame posterior, equivariant transport, and typed read-back.

4.1 NATIVE FRONTENDS AND OBJECT/EVENT SLOTS

Each modality uses a strong native encoder. Text contributes semantics and entity, event, and relation slots; spatial language is typed only relative to an identified frame or an unresolved global gauge. Images and video contribute region/track semantics, motion, and camera-frame geometry; audio contributes content, event time, source identity, and spatial cues only when observable (monaural audio never fabricates a direction); depth and points provide coordinates, normals, and symmetric-traceless tensors. Slots carry the identities and transformations on which cross-modal matching is tractable.

4.2 INVARIANT CANDIDATE CORRESPONDENCE

Each token is summarized by a descriptor that no rotation can change,

$$I_i = \Big[\, s_i + e_{m_i},\ \{\|h_{ic}^{(1)}\|\}_c,\ \mathrm{Gram}(h_i^{(1)}),\ \{\|h_{ic}^{(2)}\|_F\}_c,\ \{v_i^\top h_{ic}^{(2)} v_i\}_c \,\Big], \tag{4}$$

with a modality embedding e_{m_i} and v_i the first vector channel. R world anchors have learned keys κ_a and fixed times $\tau_a \in [0,1]$. Token i is routed by

$$\begin{aligned} \alpha_{ia} &= w_i \cdot \mathrm{softmax}_{a \in \mathrm{TopK}}\Big(\beta \cos\big(\phi(I_i), \kappa_a\big) - \lambda\,|t_i - \tau_a|\Big), \\ w_i &= \mathbb{1}[\mathrm{valid}_i]\ c_i\ \sigma\big(\psi(s_i)\big), \end{aligned} \tag{5}$$

where ϕ maps the invariant descriptor I_i into the anchor-key space, β, λ are learned scales, c_i is the frontend confidence, and $\sigma(\psi(\cdot))$ a learned token reliability. Routing reads only invariants, so a common world rotation cannot rematch the mention "lamp" to a different region; time enters as

proximity to an anchor, not as a coordinate to be rotated. When a one-to-one matching is required, entropic OT (Sinkhorn) replaces the softmax, and approximate nearest-neighbor top-k retrieval keeps routing scalable; no $N_m \times N_n$ cross-modal matrix is formed.

4.3 Spatial and temporal posteriors

Each modality's local-to-world arrow is a discrete posterior

$$q_m(g) = \sum_{h=1}^{H} \pi_{mh}\, \delta(g - g_{mh}), \qquad \pi_{mh} \propto \exp(-e_{mh}/\sigma), \tag{6}$$

with an observability and stabilizer label. Calibration metadata can supply a sharp prior. Otherwise, invariant matching between paired geometric evidence gives one entropic-OT plan per temperature, each plan gives a weighted Kabsch hypothesis g_{mh} with residual e_{mh}, residuals are refined, and hypotheses that break cycle consistency across compatible frames are pruned. A rank test marks collinear or too-sparse evidence as unobservable, and the frame confidence $c_m^{\mathrm{fr}} = 1 - H(\pi_m)/\log H$ vanishes there. Symmetric or partially visible observations therefore keep a broad or quotient-valued posterior instead of a fictitious pose; hard canonicalization is the one-hypothesis special case. Time has its own posterior $q(\delta, \gamma, \omega)$ over offset, rate drift, and local warp, estimated by coarse event alignment followed by banded fine alignment; aligned timestamps then enter the router through the anchor-proximity term of (5).

4.4 Posterior equivariant transport and world anchors

A typed value is moved into the world frame *before* anything is added:

$$\widetilde{h}_i^{(\ell)} = \mathbb{E}_{g \sim q_{m_i}} \left[\mathrm{D}^{(\ell)}(g) h_i^{(\ell)} \right], \tag{7}$$

i.e. $\sum_h \pi_{mh} R_h v$ for vectors, $\sum_h \pi_{mh} R_h T R_h^\top$ for tensors, and $\sum_h \pi_{mh}(R_h p + t_h)$ for positions. Each modality m then fills every anchor with its own summary, semantic and structural mass kept apart:

$$S_{ma} = \frac{\sum_{i \in m} \alpha_{ia} s_i}{\mu_{ma}^{\mathrm{sem}}}, \qquad H_{ma}^{(\ell)} = \frac{\sum_{i \in m} \alpha_{ia} c_m^{\mathrm{fr}} \widetilde{h}_i^{(\ell)}}{\mu_{ma}^{\mathrm{str}}}, \tag{8}$$

where μ are the corresponding masses. An unobservable frame ($c_m^{\mathrm{fr}} = 0$) therefore removes only that modality's structure, never its language, identity, or acoustic content. Modalities are then combined by agreement. Each summary is compared with the anchor's cross-modal consensus, semantically by $1 - \cos$ and structurally by the normalized typed difference after transport, which is counted only when at least two modalities carry structure:

$$\rho_{ma} = \exp\big(-\eta \, [\, d_{ma}^{\mathrm{sem}} + \sigma(\xi)\, d_{ma}^{\mathrm{str}} \,] \big), \qquad S_a = \frac{\sum_m \rho_{ma} \mu_{ma}^{\mathrm{sem}} S_{ma}}{\sum_m \rho_{ma} \mu_{ma}^{\mathrm{sem}}}, \tag{9}$$

and likewise for $H_a^{(\ell)}$ with structural mass. A sensor that keeps object identity but reports an incompatible pose is thus down-weighted after transport; this is the group-consistency signal of the later RGB–depth conflict test. The gate $\sigma(\xi)$ starts small, with learned temperature $\eta > 0$; neither symbol is the time rate γ or the tensor T. A small gate keeps complementary modalities from being treated as contradictory, and a KL term toward uniform anchor occupancy prevents dead anchors.

4.5 Typed anchor blocks and read-back into a trunk

Anchors are processed by typed blocks whose attention logits use scalars and invariant contractions, while values keep their type. Cross-type updates use legal Clebsch–Gordan (CG) paths: $1 \otimes 1 \to 0$ (dot), $1 \otimes 1 \to 1$ (cross), $1 \otimes 1 \to 2$ (symmetric-traceless outer product), and $2 \otimes 1 \to 1$. Invariant CG outputs update the semantic channels, and semantic scalars gate typed residuals. A map that is not an intertwiner (an equivariant linear map) would break the theorems: an MLP on concatenated typed channels treats a vector as an untyped embedding. A task readout then consumes the anchors. A pretrained trunk keeps open semantics, knowledge, and instruction following, and exchanges information with the anchors every K layers. Its tokens write into the anchors through (5)–(9)

and read them back: each token's down-projected state queries the fused anchor semantics S_a by cross-attention, invariant CG outputs enter the same path, and a zero-initialized up-projection adds the result to that token. Training therefore starts exactly at the pretrained model, and the trunk need not make every hidden channel an irreducible representation.

4.6 Two instantiations, objective, and complexity

The experiments use two implementations of the same equations. *Small heads* on frozen encoder features run the pipeline end to end: typed frontends, frame posteriors, transport, typed blocks, and a task readout. *The residual sidecar* (fig. 1) inserts the same fusion module—typed channels, a frame posterior, transport, and typed blocks—at the junction between Qwen2.5-Omni's frozen Thinker (multimodal encoder) and the language-model trunk, with one modality batch per video, audio, and text span; the sidecar and LoRA (q, v) are trained. The objective is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_c \mathcal{L}_{\text{corr}} + \lambda_f \mathcal{L}_{\text{frame}} + \lambda_t \mathcal{L}_{\text{time}} + \lambda_e \mathcal{L}_{\text{equiv}} + \lambda_r \mathcal{L}_{\text{reliability}} + \lambda_o \mathcal{L}_{\text{occupancy}}, \quad (10)$$

with each term active where its supervision exists; equivariance losses monitor approximate frontends and do not replace equivariant parameterization. With R anchors, routing costs $O(\sum_m N_m R)$ ($O(\sum_m N_m k)$ with top-k) and streaming keeps $O(Rd)$ state per modality.

5 Theoretical Properties

The four steps of §4 are the maps that satisfy the chart-change law (1). We state the assumptions here and give complete proofs in section P. For each morphism, $\mathrm{D}^{(\ell)}$ is a bounded representation; local frontends are equivariant in typed channels; routing weights are invariant; frame posteriors transform by pushforward; and backbone maps are intertwiners or invariant gates.

Theorem 1 (Global equivariance)**.** *Applying a common world transformation u to all observable modalities gives $F(\{\rho_m(u)x_m\}) = \rho_{\text{out}}(u)F(\{x_m\})$ for typed outputs, and leaves invariant outputs unchanged.*

Theorem 2 (Local-gauge independence)**.** *Changing the coordinate convention independently at each local frame, while pushforward-transforming its arrow posterior and typed features, leaves semantic and invariant predictions unchanged and transforms equivariant outputs only in the requested output gauge.*

Theorem 3 (Stabilizer-aware well-definedness)**.** *If an observation is identifiable only modulo a stabilizer H, posterior transport averaged over the corresponding coset is independent of the chosen canonical representative, and any component not fixed by H vanishes rather than being assigned an arbitrary orientation.*

Theorem 4 (Posterior composition consistency)**.** *For compatible frames i, j, k, deterministic transports satisfy $\mathcal{T}_{i\leftarrow j}\mathcal{T}_{j\leftarrow k} = \mathcal{T}_{i\leftarrow k}$. For uncertain arrows, the discrepancy is bounded by the Wasserstein distance between the direct posterior and the convolution of the two posterior measures, scaled by the representation Lipschitz constant and feature norm.*

Theorem 5 (Stability and streaming complexity)**.** *Assume feature norm at most B, anchor mass at least $\mu > 0$, and ℓ-representation Lipschitz constant L_ℓ. The anchor error is bounded by*

$$\|\widehat{H}^{(\ell)} - H_*^{(\ell)}\| \le C_A \|\widehat{\alpha} - \alpha_*\|_1 + C_F L_\ell W_1(\widehat{q}, q_*) + C_X \|\widehat{h} - h_*\|, \quad (11)$$

with constants scaling as $O(B/\mu)$. Routing costs $O(\sum_m N_m R)$, or $O(\sum_m N_m k)$ with top-k retrieval, and fusion keeps $O(Rd)$ streaming state updated chunk by chunk.

Complete assumptions and proofs are in section P (§P.1–§P.6). Three no-go statements delimit the claim and are stated as propositions in section Q, with the stabilizer argument in section P and the text-gauge and raster arguments in section R: relational text fixes axes only up to a global gauge; a symmetric observation fixes pose only modulo its stabilizer; and exact arbitrary-angle rotation on a finite raster needs a band-limited signal model.

6 EXPERIMENTS: NECESSITY, REAL GROUPS, AND ROUTED SYNERGY ON FROZEN TRUNKS

The three stages train different models and hold out different nuisances (pose or clock); each supports a claim the others do not. Stage I is *OGF-World*: four sensors, unknown $\mathrm{SO}(3)$, a four-class relative-pose label, and a mechanism suite (group×OT factorial, typed-order cuts $\ell = 0, 1, 2$, coupling spectra, a 15-subset lattice, and the frame posterior). Stage II is a small head on frozen real features: RefCOCO, CREMA circular time, TUM pose and yaw, ScanRefer $\mathrm{SE}(3)$, AVE cyclic time, QVHighlights, and RGB–depth conflict, with no generative trunk. Stage III is one residual sidecar on five frozen trunks (3B–8B) trained on Core96K, our nested mixture 12K $\subset$ 24K $\subset$ 48K $\subset$ 96K (composition in App. H), and 98 trunk–task pairs ($\approx$ 644K evaluation items) across audio–visual, spatial, and video benches. Matched comparisons and ablations run in every stage, named below. Stages I–II report Syn (3); Stage III reports $\mathrm{Syn}_{\text{route}}$ (section 6.3).

6.1 STAGE I: CONTROLLED PHYSICS, COUPLING RANK, AND THE FRAME POSTERIOR

Experimental protocol. Four sensors, unknown $\mathrm{SO}(3)$, two inner-product signs (4 classes, chance 0.25), no true frame as input; 10 seeds, 600 steps. Train and IID share the identity rotation; OOD draws an independent rotation per sensor. Arms: late fusion, dense cross-attention, Perceiver, identity-frame Sinkhorn, and hard canonicalization (true frame); interventions: pairing shuffle and a corrupted frame (table 1). App. D records the group×OT factorial, coupling spectra and the audio-to-vision gradient-norm ratio $\gamma_{\mathrm{av}} = \|\nabla_{\mathrm{aud}}\|/\|\nabla_{\mathrm{vis}}\|$, the 15-subset lattice with Shapley attributions, test-time dropout, and frame-hypothesis poisoning (fig. 8); $\ell = 0/1/2$ (fig. 5) and the frame posterior versus the MAP (hard argmax) frame (fig. 7) are in App. C; per-seed values in fig. 2.

Table 1: Stage I, OGF-World (10-seed means, 600 steps). OGF Syn is OOD minus best subset. For untyped and hard-canonicalization rows, Syn uses a matched-capacity subset model. App. C.

Method	IID	OOD	Best subset	Syn	Shuffle gap	Wrong frame
Late fusion	–	0.262	0.269	−0.007	0.009	–
Dense cross-attention	0.856	0.266	0.266	−0.008	0.004	–
Perceiver latent	–	0.253	0.257	−0.004	−0.001	–
No group (identity frames)	–	0.251	0.267	−0.016	−0.002	–
Hard canon. (true frame)	0.926	0.927	0.506	+0.419	0.455	–
OGF	0.919	0.915	0.499	**+0.416**	**0.464**	0.251

Necessity of a shared frame. Untyped mixers and identity-frame Sinkhorn stay at OOD chance (0.251–0.266, $\mathrm{Syn} \le 0$); dense attention fits training poses (IID = 0.856) and collapses (OOD = 0.266). OGF reaches OOD = 0.915 and mean $\mathrm{Syn} = +\mathbf{0.416}$ (one of the ten seeds has $\mathrm{Syn} = 0$), with no statistical gap versus hard canonicalization (0.927, $p = 1.0$). Pairing shuffle drops OGF to 0.450; a corrupted frame or a no-group ablation both return 0.251. After eight AdamW steps the Kabsch rotation residual is 6.66×10^{-16}. The label is learned only as a relation in a shared frame.

Mechanisms, failure modes, and calibration. A group×OT factorial on independent frames (table 9a) isolates group frames: identity frames stay at chance (0.245–0.250); group-frame means are 0.773–0.903, with two of five OGF seeds near 0.47. Independent-frame $\ell = 0/1/2$ give 0.259/0.777/0.882 (App. C). Untrained low-rank mixers (gated multimodal units, Perceiver, Q-Former, linear attention) collapse the coupling's effective rank to 1.05–1.57, with 93–99% of the energy in the top two components; Sinkhorn stays rank 3.84 with doubly-stochastic residual $< 10^{-7}$. Dense rank 4.04 and Sinkhorn rank 3.84 sit on different spectral scales. Dense attention starves vision on 100% of 150 steps ($\gamma_{\mathrm{av}} = 1.55\times10^{6}$); Sinkhorn's mean starvation is 11.6% (one of five seeds), and late fusion and GMU also end near $\gamma_{\mathrm{av}} = 1$, so balance alone is not the group effect. The 15-subset lattice full set is 0.963 with $\mathrm{Syn} = +\mathbf{0.475}$ and Shapley attributions 0.199/0.179/0.174/0.161 (text/vision/audio/point); non-equivariant full sets stay at chance (Dense 0.235, Perceiver 0.243, local MLP 0.254). Text+Vision is 0.429; Audio+Point is 0.332, and its interval includes chance. Dropping modalities at test time retains a missing-modality area of 0.536 versus dense 0.249; poisoning frame hypotheses leaves posterior entropy at 0.34, so a broken frame is not flagged (App. D). The posterior ties the MAP frame when the pose is determinate (0.915 vs. 0.927, $p = 1.0$), is less accurate and better calibrated under ambiguity (0.254 vs. 0.284; calibration error ECE 0.010 vs. 0.282), and wins

both under conflict (0.384 vs. 0.300; ECE 0.052 vs. 0.282). Optimal transport is not the accuracy source on this inner-product label; an MLP on already-transported vectors reaches 0.903.

6.2 Stage II: Continuous Geometry on Real Sensor Features

Experimental protocol. Frozen encoder features, small head, no generative trunk. table 2 isolates one group per row; CREMA scale, SigLIP2, discrete C_4 (90° orbit), six held-out frozen-encoder tasks, LongVALE, and the capacity-fair TUM / AVE retention protocols are in App. F (figs. 11 and 13) (Yu et al., 2016; Radford et al., 2021; Cao et al., 2014; Sturm et al., 2012; Chen et al., 2020a; Tian et al., 2018; Lei et al., 2021; Tschannen et al., 2025; Wang et al., 2024; Geng et al., 2025). Arms: Q-Former, dense, no-group, $\mathrm{SO}(3)$ augmentation, pose-concat (pose appended to features), naive time fusion, Kabsch, a larger regressor, and true-frame heads.

Table 2: Stage II, strongest competitor per row. hard-hybrid: single-hypothesis RGB–D pose; sel. err: selective error at 50% coverage; C_{10}: ten cyclic shifts; MR-full-mAP: official moment-retrieval mAP; early: early fusion; gap: change in error under a test rotation. App. F.

Task (seeds)	Metric	Strongest baseline	OGF
RefCOCO, CLIP (10)	val IoU	Q-Former 0.392	**0.419**
CREMA-D circular time (10)	AUROC / miss-event AP	true-frame 1.000 / 0.971	**1.000 / 0.978**
TUM metric pose (10)	joint / sel. error	hard-hybrid 0.872 / 0.022	**0.878 / 0.011**
TUM $\mathrm{SO}(3)$ yaw bins (5)	antipodal acc / geo. err	$\mathrm{SO}(3)$ aug. 0.8129 / 2.7°	**0.9908 / 0°**
ScanRefer $\mathrm{SE}(3)$ (5)	OOD err (m) / gap	regressor 0.968 / 2.345	**0.728 / 0**
AVE C_{10} OOD (5)	OOD AP	aug. 0.503	**0.548**
QVHighlights official (5)	MR-full-mAP	dense 10.51	**16.81**
QVHighlights 300M (5)	MR-full-mAP	early 10.85	**15.66**
TUM RGB–depth conflict (10)	AUROC	hard-hybrid 0.953	**0.959**

Continuous spatial transport. On TUM yaw bins the typed head is 0.9908 with geodesic error 0.0° in every 60° bin, including the antipode. Kabsch is also 0.0° but only 0.8989 accurate (registration without a typed readout). $\mathrm{SO}(3)$ augmentation stays in 0.81–0.85 (antipodal 0.8129, 2.7°); pose-concat 0.78–0.80 (antipodal 0.8021, 13.2°); no-augmentation falls to 0.3399 (165.0°). RefCOCO gains $+10.46$pp over its best subset, and TUM joint pose $+48.4$pp over the strongest unimodal subset. On ScanRefer, training never sees $|\mathrm{yaw}| \geq 90^\circ$: OGF OOD error is 0.728 m with rotation gap 0 against a larger regressor at 0.968 m and rotation gap 2.345 (tables 22 and 23). Continuous invariance is a property of typed transport, not of covering the orbit by data. With a smaller OGF head, accuracy on the held-out full-circle sequence is still 0.9931 against $8\times$ augmentation 0.9296 (App. F).

Cyclic time and temporal posteriors. CREMA circular sync is AUROC 1.000 against single streams ≤ 0.506; missing-event AP is 0.978 versus hard 0.971. On AVE, held-out OOD AP is 0.548 versus augmentation 0.503 and naive 0.422—and naive leads when nothing is shifted (clean acc 0.656 vs. 0.568). The eleven-shift mean event AP is 0.5457 versus 0.5131 / 0.5036; at ± 5 s OGF stays at 0.555 while naive falls to 0.404. The posterior's top offset is correct in at most 15% of clips. Official QVHighlights moment-retrieval mAP is 16.81 versus dense 10.51; a 300M-parameter encoder scores 15.66 versus early fusion 10.85 (no-group 10.58). Highlight Hit@1 versus dense has an interval that includes 0. On LongVALE the same pairing and time shifts lower F1 and IoU.

Conflict detection and scalability. On the same TUM sequence, agreement after transport flags mismatched RGB at AUROC 0.9592 against single-stream confidence 0.7135 and hard transport 0.953 (table 26). CREMA heads at 98M / 301M / 1.00B keep missing-AP gains $+1.03$ / $+0.76$ / $+1.70$pp. Frozen SigLIP2 referring IoU agrees on clean and 90°-rotated charts to 4×10^{-6} and exceeds a C_4-augmented Q-Former by $+0.0051$; without compilation, inference takes 1.22–$1.70\times$ the compiled runtime. Discrete C_4 on frozen Qwen2-VL is exact (val IoU 0.453 on both charts) and ties a C_4-augmented Q-Former ($p = 0.29$). On six held-out frozen-encoder tasks the mean drop versus the matched head is 0.321pp; the largest is CREMA audio-only average precision, 1.42pp.

6.3 Stage III: Routed Synergy on Five Frozen Trunks, Nested Scale, and Broad Benches

Stage III asks whether one residual sidecar, inserted once into each of five *frozen* generative trunks, is used by the trunk and, where audio exists, yields routed synergy that pairing shuffle removes on four

of the six audio–visual benches. The trunk always sees the full input and question; a V-route (A-route) sends only vision (audio) into the sidecar: $\mathrm{Syn}_{\text{route}} = P(\text{full}) - \max(P(\text{V-route}), P(\text{A-route}))$.

Experimental protocol. The same sidecar attaches to five frozen trunks (Xu et al., 2025; OpenGVLab Team, 2024; Li et al., 2025a; Qwen Team, 2025): Omni-7B, Omni-3B, LLaVA-OV, InternVL2, and Qwen3-VL-4B, on the Core96K mixture at every nested scale ($6{,}000$ steps, batch 64, 384K views). Omni-7B and Omni-3B run six audio–visual benches (OmniBench, WorldSense without subtitles, Daily-Omni, XModBench-Lite (a $6{,}000$-item subset), AVE, AV-Odyssey) (Li et al., 2025c; Hong et al., 2026; Zhou et al., 2025; Wang et al., 2026b; Tian et al., 2018; Gong et al., 2024) and the six XMod directions A→V, V→A, T→A, A→T, T→V, V→T; the vision–language trunks run only T→V and V→T; all five also run Video-Holmes, RefCOCO/ + /g, 3DSRBench, VSI-Bench, All-Angles, ViewSpatial, SPBench-SI, CrossView, NExT-QA, Perception Test, and Charades-STA (Cheng et al., 2025; Yu et al., 2016; Mao et al., 2016; Ma et al., 2025a; Yang et al., 2025; Yeh et al., 2026; Li et al., 2026a; 2025b; Wang et al., 2026a; Xiao et al., 2021; Pătrăucean et al., 2023; Gao et al., 2017) (98 pairs, ≈ 644K items). Each comparison reports a frozen trunk, LoRA-only, OGF, and a V-route; an A-route where audio exists; pairing shuffle and a 30° rotation of the recorded relative pose (image and label unchanged); time shift where a synchronized AV clock exists; and three retrained ablations (no typed transport, equal frame-hypothesis weights, no correspondence) on Omni-7B. Omni uses the Thinker–LM junction of Qwen2.5-Omni; the vision–language trunks have no Thinker. Omni-7B at 96K trains these five arms at two seeds; other trunks and scales use one.

Table 3: Omni-7B at 96K, mean of two seeds. Accuracy (%). A superscript is routed synergy (pp): one number when the seeds agree to 0.1pp, otherwise the two seeds, with the minus kept on both ends of a negative range. †: intervention on that seed's checkpoint.

Condition	OmniB.	WorldS.	Daily	XMod	AVE	AV-Od.
Frozen trunk	56.1	45.4	62.1	60.2	68.5	45.5
LoRA-only	58.6	47.5	64.5	61.8	69.6	47.0
OGF (V-route)	58.1	46.8	63.8	61.1	68.7	46.6
OGF (A-route)	57.4	46.0	63.2	61.0	68.6	46.2
OGF full	$\mathbf{60.0}^{+1.8–2.0}$	$\mathbf{48.5}^{+1.7}$	$\mathbf{65.9}^{+2.0–2.2}$	$\mathbf{62.7}^{+1.5–1.6}$	$\mathbf{69.9}^{+1.2}$	$\mathbf{48.3}^{+1.7–1.8}$
Pairing shuffle†	$56.2^{+0.4}$	$45.0^{-0.4}$	$60.9^{-1.2–-1.1}$	$60.4^{+0.7}$	$66.7^{-0.8–-0.7}$	$45.8^{-0.4}$
Time shift†	–	$47.2^{+0.5–0.6}$	$63.9^{+0.5–0.7}$	–	$68.0^{-0.2}$	$48.0^{+1.4–1.5}$
Wrong frame†	$59.7^{+1.8}$	$48.3^{+1.5–1.6}$	$65.8^{+2.1}$	$62.6^{+1.5–1.6}$	$69.8^{+1.1–1.2}$	$48.3^{+1.7–1.8}$
No transport	$59.5^{+1.6–1.7}$	$47.7^{+1.4}$	$65.3^{+1.7–1.8}$	$62.3^{+1.2}$	$69.7^{+1.0}$	$47.8^{+1.3–1.4}$
Uniform posterior	$59.7^{+1.8}$	$48.1^{+1.3–1.4}$	$65.4^{+1.8}$	$62.4^{+1.2–1.3}$	$69.7^{+1.0}$	$48.1^{+1.6}$
No correspondence	$58.5^{+0.6}$	$46.9^{+0.5–0.6}$	$63.2^{+0.7–0.8}$	$61.7^{+0.5}$	$69.0^{+0.3–0.4}$	$47.3^{+0.8–0.9}$

Cross-trunk scaling. At 96K, OGF exceeds LoRA on $15/19$, $19/20$, $16/16$, $12/16$, and $16/16$ benches on Omni-7B, Omni-3B, LLaVA-OV, InternVL2, and Qwen3-VL-4B (tables 36 and 37). At 12K, OGF is below LoRA on 85 of 87 trunk–benchmark pairs (table 38). The mean gap turns positive at 24K on Omni-7B, LLaVA-OV, and Qwen3-VL (-0.09/$+0.08$/$+0.18$/$+0.21$pp from 12K to 96K), and at 48K on Omni-3B and InternVL2.

Sources of routed synergy. On Omni-7B at 96K, OGF has 9.20M trainable parameters and LoRA-only 9.58M (16.745 versus 16.744 GiB per GPU). Both seeds keep $\mathrm{Syn}_{\text{route}}$ positive on all six audio–visual benches ($+1.2$ to $+2.2$pp). Pairing shuffle makes it negative on four and leaves OmniBench and XModBench-Lite near $+0.4$ and $+0.7$pp. On Daily-Omni the gain sits on audio–visual items ($+3.15$pp) and event items ($+3.10$pp) (table 31).

7 Conclusion

Unifying sequence interfaces does not make local sensor charts commensurate. OGF does so on an observation groupoid: invariant correspondence routes evidence to shared anchors, a relative-frame posterior absorbs pose and clock uncertainty, and typed transport writes every stream in one frame before pooling. Across three stages, one criterion finds fusion only where this holds: removing the group returns chance; continuous sensors need typed transport, not orbit coverage; on frozen Omni trunks, routed audio–visual synergy holds across two seeds and largely vanishes when pairing breaks. Transport carries the spatial gain, since retraining without it falls below LoRA there, and the sidecar leads LoRA on most vision–language benches as data grow.

AI USE STATEMENT

This section is required by ICLR 2027 and does not count toward the page limit. We used generative AI tools only to polish the wording of the manuscript. We did not use them to invent claims, results, proofs, figures, or code. The authors reviewed every AI-assisted edit and take full responsibility for the final text.

REPRODUCIBILITY STATEMENT

This optional statement is recommended by ICLR 2027. It is not part of the main text and does not count toward the page limit. We have aimed to make the claims checkable by pointing to the parts of the paper, appendix, and supplementary material that contain the details, rather than restating those details here. For the architecture and algorithm, the fusion maps and the two instantiations are in section 4; the input fields and numerical settings are in section T. For the theoretical results, the assumptions and complete proofs of the five theorems are in section P, the three no-go propositions in section Q, and the typed construction in section O. For the experiments, metric definitions are in section A; Stage I protocols in sections C and D; Stage II protocols in sections E and F; the Core96K mixture, processing, optimization, and evaluation matrix in section H; the evaluation protocol in section U; and the matched baselines and ablations in section K. Per-seed results, negative results, and definitions of the reported quantities are in sections I, L and N. Anonymous source code and the data split will be submitted as supplementary material.

REFERENCES

Jean-Baptiste Alayrac et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022.

Michael Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. In *European Conference on Computer Vision*, 2020a.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020b.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-Holmes: Can MLLM think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

Taco Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant CNNs on homogeneous spaces. In *Advances in Neural Information Processing Systems*, 2019.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. LLaMA-Omni: Seamless speech interaction with large language models. In *International Conference on Learning Representations*, 2025.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. SE(3)-transformers: 3D roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems*, 2020.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *IEEE International Conference on Computer Vision*, 2017.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. LongVALE: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, and Xiangyu Yue. AV-Odyssey bench: Can your multimodal LLMs really understand audio-visual information? *arXiv preprint arXiv:2412.02611*, 2024.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating real-world omnimodal understanding for multimodal LLMs. In *International Conference on Learning Representations*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, 2021.

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *International Conference on Machine Learning*, 2023.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32:922–923, 1976.

Jie Lei, Tamara L. Berg, and Mohit Bansal. QVHighlights: Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems*, 2021.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025a.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. ViewSpatial-Bench: Evaluating multi-perspective spatial localization in vision-language models. In *European Conference on Computer Vision*, pp. 95–111, 2026a.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. SpatialLadder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*, 2025b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023.

Yanlin Li, Minghui Guo, Kaiwen Zhang, Shize Zhang, Yiran Zhao, Haodong Li, Congyue Zhou, Weijie Zheng, Yushen Yan, Shengqiong Wu, Wei Ji, Lei Cui, Furu Wei, Hao Fei, Mong-Li Lee, and Wynne Hsu. UniM: A unified any-to-any interleaved multimodal benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15902–15911, 2026b.

Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Yidan Wen, Yanghai Wang, Shihao Li, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. OmniBench: Towards the future of universal omni-language models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025c.

Xiulong Liu, Zhikang Dong, and Peng Zhang. Tackling data bias in MUSIC-AVQA: Crafting a balanced dataset for unbiased question-answering. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4478–4487, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3DSRBench: A comprehensive 3D spatial reasoning benchmark. In *IEEE/CVF International Conference on Computer Vision*, 2025a.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. JanusFlow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

OpenGVLab Team. InternVL2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. OpenGVLab technical report, 2024.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, et al. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023.

Omri Puny, Matan Atzmon, Edward J. Smith, Ishan Misra, Aditya Grover, Heli Ben-Hamu, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations*, 2022.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K. Du, Zehuan Yuan, and Xinglong Wu. TokenFlow: Unified image tokenizer for multimodal understanding and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Machine Learning*, 2021.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. In *Advances in Neural Information Processing Systems*, 2023.

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. CoDi-2: In-context, interleaved, and interactive any-to-any generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *European Conference on Computer Vision*, 2018.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Wei Wang, Yuqian Yuan, Tianwei Lin, Wenqiao Zhang, Siliang Tang, Jun Xiao, and Yueting Zhuang. CrossView suite: Harnessing cross-view spatial intelligence of MLLMs with dataset, model and benchmark. *arXiv preprint arXiv:2605.18621*, 2026a.

Xingrui Wang, Jiang Liu, Chao Huang, Xiaodong Yu, Ze Wang, Ximeng Sun, Jialian Wu, Alan Yuille, Emad Barsoum, and Zicheng Liu. XModBench: Benchmarking cross-modal capabilities and consistency in omni-language models. In *International Conference on Learning Representations*, 2026b.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal LLM. In *International Conference on Machine Learning*, 2024.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in MLLMs. In *AAAI Conference on Artificial Intelligence*, 2026.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, 2016.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. AnyGPT: Unified multimodal LLM with discrete sequence modeling. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-Omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.

A Experimental Appendix

Each subsection states one protocol and its results. Notation and proofs follow. Stage I means in table 1 are repeated here with confidence intervals and every seed.

Figure	Contents
figs. 2 to 4	Stage I fusion seeds, intervals, and known subsets
figs. 5 and 6	Sensitivity and coupling, then dropout and spectra
fig. 7	Ambiguity, conflict, robustness areas, and frame poisoning
fig. 8	Independent-frame factorial, lattice centers, and Shapley values
figs. 9 and 10	Stage II frozen-encoder task heads, then RGB–depth conflict
figs. 11 and 12	TUM yaw and ScanRefer, then six frozen trunks
fig. 13	AVE retention and CLIP C_4
figs. 14 and 15	Orbit aggregates and the unmodified Omni-7B frame probe
fig. 16	Omni-7B Core96K profile and per-task gains
figs. 17 to 20	Ablations, routes, categories, and Omni-3B
figs. 21 and 22	Intervention and route scores
figs. 23 and 24	The other four trunks and XMod directions
figs. 25 to 28	Nested scale means, CDFs, and the 348 nested-scale cells
figs. 29 and 30	Negative results and boundaries

Metric definitions used everywhere. $\mathrm{Syn} = P(M) - \max_{A \subsetneq M} P(A)$ uses independently trained strict subsets; $v(\emptyset) = 0.25$ on the four-class world. Shuffle gap is the drop after a pairing permutation. Equivariance gap is the change in a score after a random group element is applied to the test inputs. Geodesic error is the residual rotation in degrees after a declared registration. Routed synergy $\mathrm{Syn}_{\text{route}}$ is defined in section 6.3: the trunk always sees every stream, and only the sidecar is restricted. Retention AUC on AVE is the arithmetic mean of event AP over eleven circular shifts. Missing-modality area averages accuracy over $\{4, 3, 2, 1\}$ present modalities. Wrong-group area averages accuracy over wrong-MAP fractions $\{0, 0.25, \ldots, 1\}$.

When two protocols share a dataset. Some datasets support more than one question. The two protocols are never averaged. When one dataset supports two questions, each question keeps its own protocol and its own table.

Experiment	First protocol	Second protocol
Spectra / γ_{av}	untrained coupling + 150-step gradients	—
group×OT factorial	independent-frame factorial	shared-frame capacity check
Lattice / Shapley / dropout	independent-frame power set	—
Order ℓ on OOD frames	independent-frame $\ell = 0/1/2$ (0.259 at $\ell = 0$)	5-seed sensitivity medians (0.506 at $\ell = 0$)
TUM continuous yaw	five-arm yaw bins	capacity-fair three-arm full-circle sequence
AVE shift retention	11-point curve	held-out OOD AP
AVE family comparison	AP 0.662 vs. 0.611	separate from held-out and retention
Posterior calibration	10-seed ambiguity / conflict	continuous-α ECE (supplementary)
CLIP RefCOCO vs. Qwen2-VL C_4	CLIP referring regression	discrete C_4 token referring
CREMA / metric TUM / ScanRefer / QVHighlights / conflict / scale	the protocol named in that subsection	—

B Experiment Catalog

The catalog names the question answered by each subsection. Each row points to the appendix subsection that states the protocol.

Table 4: Experiment catalog.

Question	Recorded in
End-to-end equivariance and optimizer-step preservation	App. C
Groupoid composition and inverse cycles	App. C
Fusion necessity (synergy) on the controlled world	App. C
High-confidence / wrong-group conflict	App. C
Wrong-group control and misspecification curve	App. C
Gauge independence and stabilizer/quotient (proofs)	App. P
Text-direction gauge (identifiable only up to a global Q)	App. R
Temporal offset and drift on AVE	App. F
Order $\ell = 0/1/2$ and posterior versus MAP	App. C
Group $\times$ OT factorial, spectra, lattice, dropout, poisoning	App. D
Image–text referring on frozen CLIP (Radford et al., 2021)	App. F
Discrete C_4 referring on frozen Qwen2-VL tokens (Wang et al., 2024)	App. F
Circular time-gauge event alignment on CREMA-D	App. F
CREMA heads at 98M / 301M / 1.00B	App. F
TUM metric pose and five-arm yaw bins	App. F
TUM capacity-fair up/down (a different question from the yaw bins)	App. F
TUM RGB–depth conflict detection	App. F
AVE held-out C_{10} and eleven-shift retention	App. F
AVE family comparison (a different metric from the clock splits)	App. F
Official-feature and 300M moment retrieval on QVHighlights	App. F
LongVALE mechanism controls (F1/IoU does not improve)	App. F
Continuous 3D referring under camera-viewpoint OOD	App. F
Same small head on six frozen vision–language feature families	App. F
Gain persistence across frozen SigLIP2 scales	App. F
Frozen-encoder retention on six tasks	App. F
Omni-7B C_4 orbit module (rotation probe)	App. G
400-step frozen-CLIP head and unmodified OmniBench frame probe	App. G
Core96K nested mixture on frozen Omni and VL trunks	App. H

C Controlled World (Stage I): Full Records for Protocol and Ambiguity

This section is the 10-seed world behind table 1. The independent-frame factorial, spectra, and lattice are in App. D; they share the world but not the protocol.

C.1 World, label, and default model

Four modalities each observe one unit vector $w_m \in \mathbb{R}^3$. Training and IID share one identity frame $R = I$ across modalities, so the two inner products are visible without transport. OOD draws an independent $R_m \in \mathrm{SO}(3)$ per modality: $v_m = R_m^\top w_m$, plus a set of frame hypotheses; the true R_m is never an input. The label is $\mathbf{1}[w_0 \cdot w_1 > 0] + 2 \cdot \mathbf{1}[w_2 \cdot w_3 > 0]$ (chance 0.25). Prediction on OOD is possible only after the four local vectors have been written in one frame.

The default OGF uses semantic width 32, four vector channels and one tensor-2 channel, one anchor, correspondence width 16, one backbone layer with four heads, three frame hypotheses with temperatures $(0.03, 0.08, 0.2)$, fp32, and no dropout (88,644 parameters). Matched baselines: local and hard canonicalization 40,324, frame-conditioned 47,236, dense 68,436, Perceiver latent 85,332. Fusion, no-group, ambiguity, and robustness runs use seeds 0–9 and 600 steps on CPU. The sensitivity sweep below is a separate protocol (5 seeds, 220 steps, batch 128). OOD uses world orientations unseen in training. Shuffle permutes which observations are paired. Wrong frame supplies an incorrect reference; the entry is accuracy under that reference.

C.2 EQUIVARIANCE RESIDUALS

With a C_4-equivariant RGB frontend, Sinkhorn matching, and Kabsch alignment, the equivariance residuals are 1.39×10^{-17} / 7.48×10^{-18} / 8.52×10^{-18} for semantic, vector, and tensor-2 channels; Sinkhorn row / column errors 1.11×10^{-16} / 5.55×10^{-17}; Kabsch R / t errors 6.66×10^{-16} / 2.22×10^{-16} with zero residual. After eight AdamW steps the same three equivariance errors are 3.47×10^{-18} / 6.83×10^{-18} / 1.48×10^{-18}. The same residuals hold for deterministic transport, discrete posterior transport, invariant descriptors, groupoid composition and inverse cycles, stability bounds, anchor complexity, fixed streaming state, and the wrong-frame non-identifiability check, in float64, float32, and bfloat16.

C.3 STAGE I FUSION MEANS AND THE RETAINED FAILURE

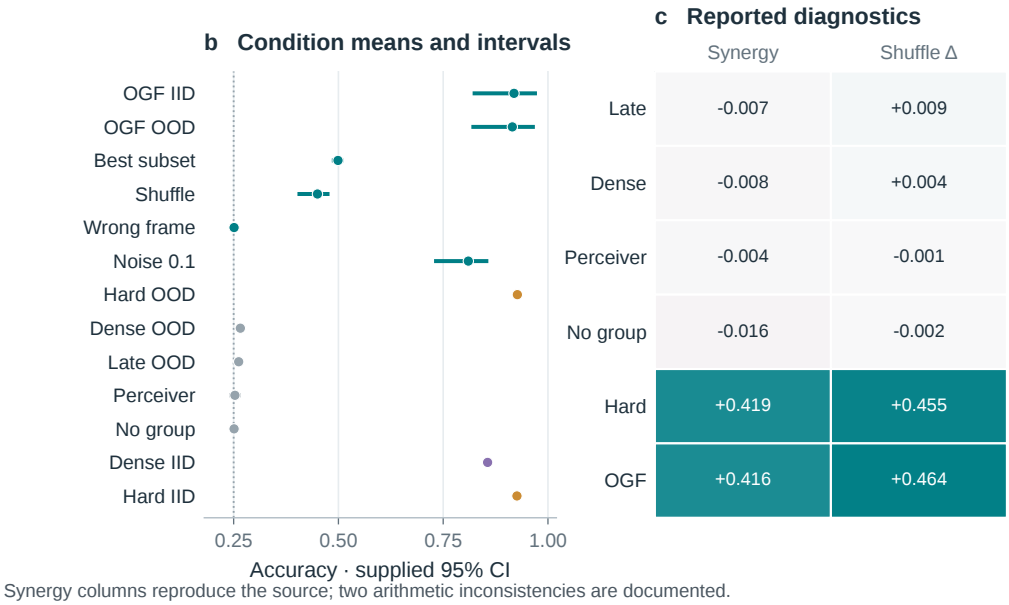
a Every controlled seed

	0	1	2	3	4	5	6	7	8	9
IID	0.960	0.511	0.961	0.969	0.939	0.977	0.967	0.959	0.980	0.966
OOD	0.953	0.505	0.966	0.968	0.946	0.967	0.958	0.954	0.971	0.959
Subset	0.472	0.505	0.498	0.502	0.503	0.490	0.517	0.516	0.475	0.510
Shuffle	0.467	0.264	0.471	0.468	0.481	0.468	0.465	0.474	0.481	0.466
Wrong frame	0.248	0.242	0.248	0.252	0.254	0.253	0.254	0.252	0.253	0.251
Hard OOD	0.927	0.930	0.923	0.918	0.928	0.921	0.930	0.928	0.926	0.940
Dense OOD	0.269	0.269	0.277	0.278	0.258	0.259	0.273	0.278	0.256	0.241

Failed seed 1 is retained. Dense IID 0.856 and hard IID 0.926 remain in the source ledger.

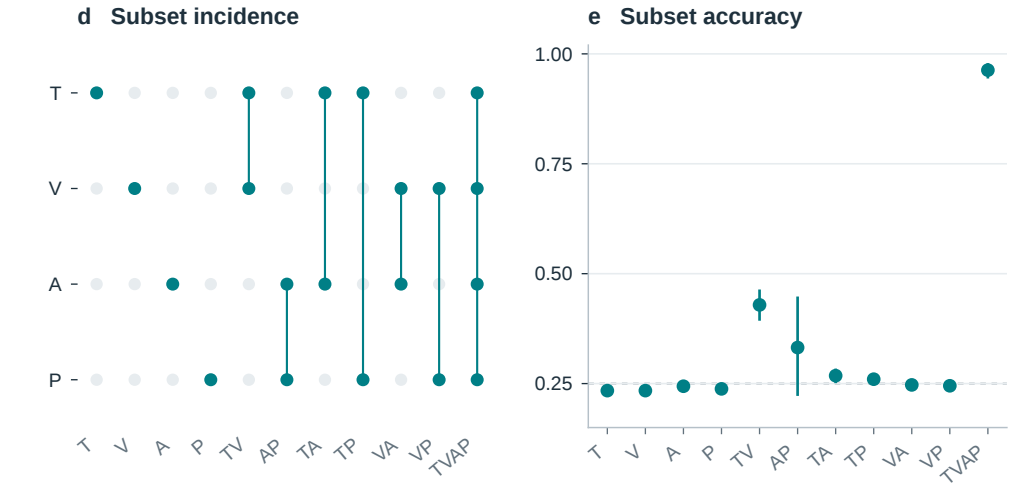
Figure 2: Stage I seeds. The failed seed is retained.

The 10-seed means in table 1 are repeated here with intervals and every seed. Seed 1 fits neither IID (0.511) nor OOD (0.505) and stays in the mean. The converged median (0.959, $\mathrm{Syn} = +0.466$) is kept separate from the 10-seed mean $\mathrm{Syn} = +0.416$.



Synergy columns reproduce the source; two arithmetic inconsistencies are documented.

Figure 3: Stage I intervals and reported synergy.



T/V/A/P: text / vision / audio / point. Triple subsets: only range 0.416–0.475 is known.

Figure 4: Stage I independently trained subsets.

Table 5: Stage I fusion, 10 seeds, 600 steps. Intervals are 95% CIs on the listed mean. Hard canon. is given the true frame; No group keeps Sinkhorn but replaces every frame by the identity. Hard canonicalization receives the true frame, so its synergy is an oracle. For dense attention and hard canonicalization, synergy uses a matched-capacity subset model.

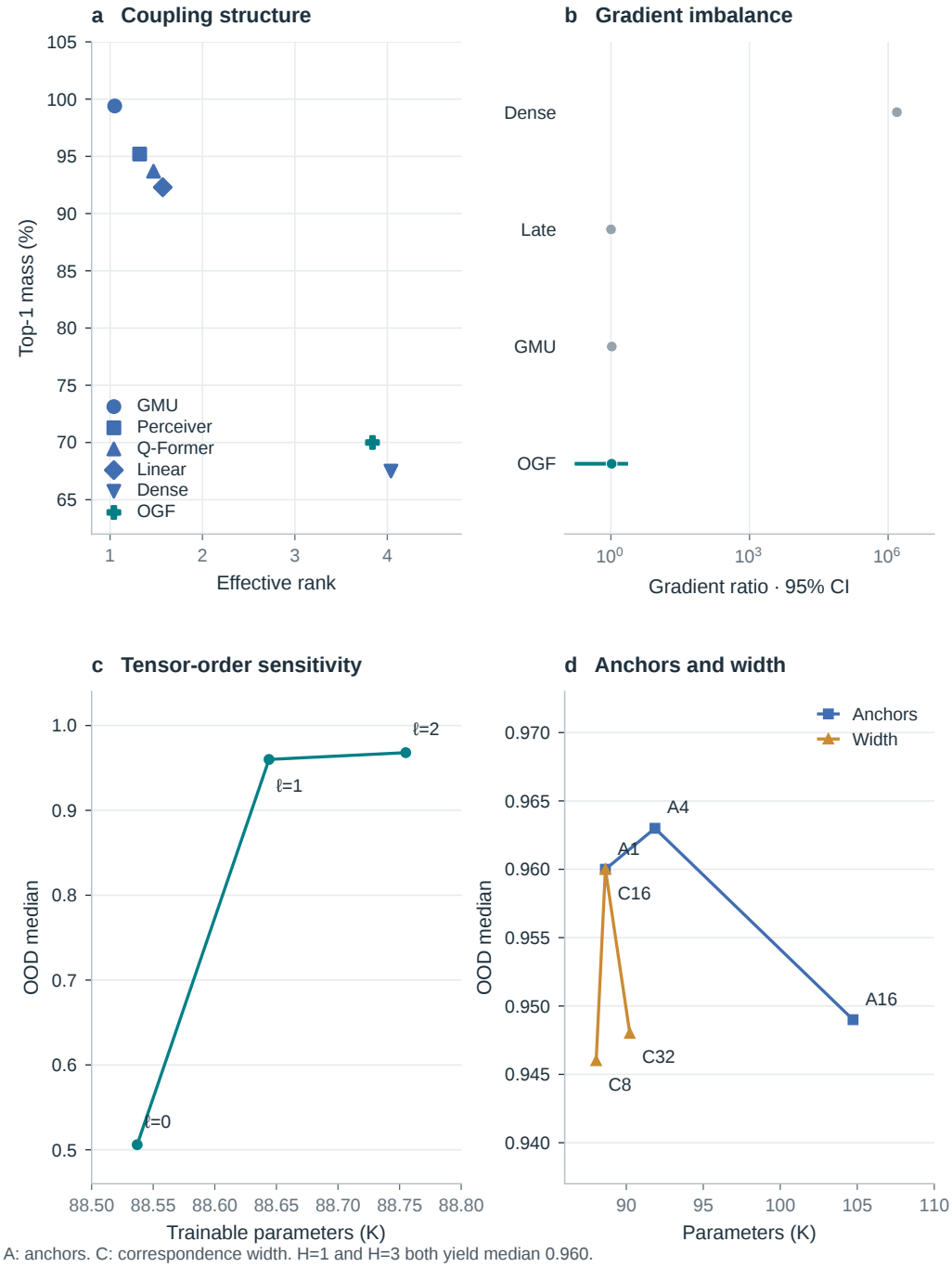
Condition	Mean	95% CI	Syn	Shuffle gap
OGF IID	0.919	$[0.825, 0.969]$	—	—
OGF OOD	0.915	$[0.822, 0.964]$	$+0.416$	0.464
OGF best subset	0.499	$[0.489, 0.507]$	—	—
OGF shuffled pairs	0.450	$[0.407, 0.474]$	—	—
OGF wrong frame	0.251	$[0.248, 0.253]$	—	—
OGF noise 0.1	0.810	$[0.733, 0.853]$	—	—
Hard canon. OOD	0.927	$[0.924, 0.931]$	$+0.419$	0.455
Dense attn. IID / OOD	0.856 / 0.266	OOD $[0.258, 0.272]$	-0.008	0.004
Late fusion OOD	0.262	$[0.260, 0.264]$	-0.007	0.009
Perceiver OOD	0.253	$[0.244, 0.262]$	-0.004	-0.001
No group OOD	0.251	$[0.248, 0.253]$	-0.016 $[-0.019, -0.014]$	-0.002

Table 6: OGF and matched arms, every seed. Seed 1 is the retained collapse.

Seed	IID	OOD	Subset	Shuffle	Wrong	Canon OOD	Dense OOD
0	0.960	0.953	0.472	0.467	0.248	0.927	0.269
1	0.511	0.505	0.505	0.264	0.242	0.930	0.269
2	0.961	0.966	0.498	0.471	0.248	0.923	0.277
3	0.969	0.968	0.502	0.468	0.252	0.918	0.278
4	0.939	0.946	0.503	0.481	0.254	0.928	0.258
5	0.977	0.967	0.490	0.468	0.253	0.921	0.259
6	0.967	0.958	0.517	0.465	0.254	0.930	0.273
7	0.959	0.954	0.516	0.474	0.252	0.928	0.278
8	0.980	0.971	0.475	0.481	0.253	0.926	0.256
9	0.966	0.959	0.510	0.466	0.251	0.940	0.241

C.4 Sensitivity sweep and the order ablation

A one-factor-at-a-time sweep (5 seeds, 220 steps, batch 128) reports medians, because a single seed skews means (table 7). Only the typed order matters on this determinate-frame task: the sweep median at $\ell_{\max} = 0$ is 0.506, not chance, and $\ell \geq 1$ reaches 0.960–0.968. This sweep is not the independent-frame ℓ claim. More anchors, a second or third frame hypothesis, or wider correspondence do not move the score; those axes are for the ambiguity regimes below. The primary ℓ claim trains and tests on independent random frames: $\ell = 0/1/2$ give 0.259 / 0.777 / 0.882 (CI $[0.241, 0.279]$ / $[0.587, 0.966]$ / $[0.693, 0.981]$). A shared-frame capacity check on the same orders is 0.248 / 0.777 / 0.868.



A: anchors. C: correspondence width. H=1 and H=3 both yield median 0.960.

Figure 5: Coupling rank, γ_{av}, and the one-factor sensitivity sweep.

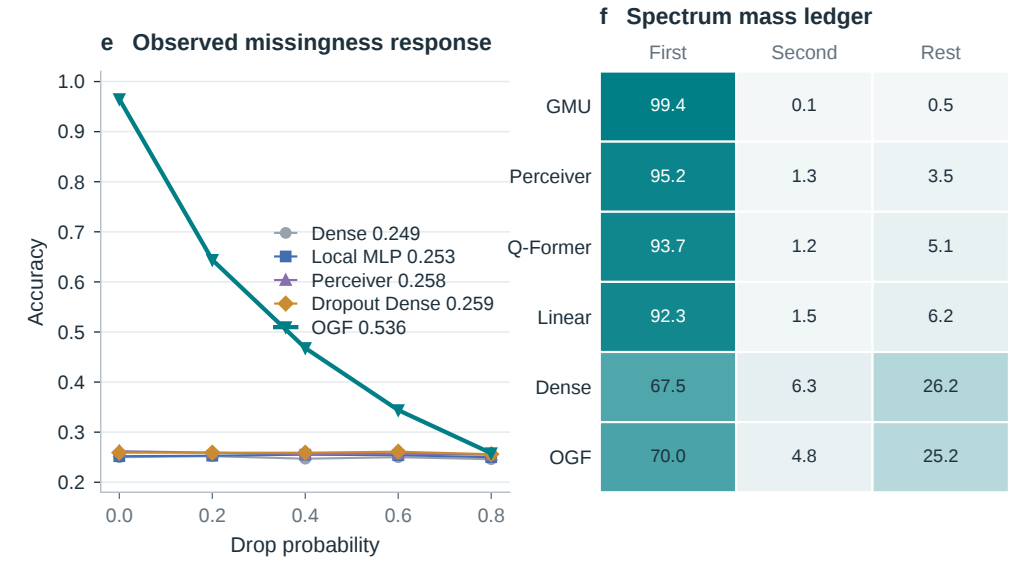


Figure 6: Observed dropout response and leading spectral mass.

Table 7: Sensitivity sweep (median OOD, 5 seeds, 220 steps; wrong-frame medians ≈ 0.25 throughout).

Axis	Value	OOD median	Params
$\ell_{\max}$ (tensor order)	0 / 1 / 2	0.506 / 0.960 / 0.968	88,537 / 88,644 / 88,755
World anchors	1 / 4 / 16	0.960 / 0.963 / 0.949	88,644 / 91,860 / 104,724
Frame hypotheses	1 / 3	0.960 / 0.960	88,644
Correspondence width	8 / 16 / 32	0.946 / 0.960 / 0.948	88,044 / 88,644 / 90,228

C.5 AMBIGUITY AND CONFLICT

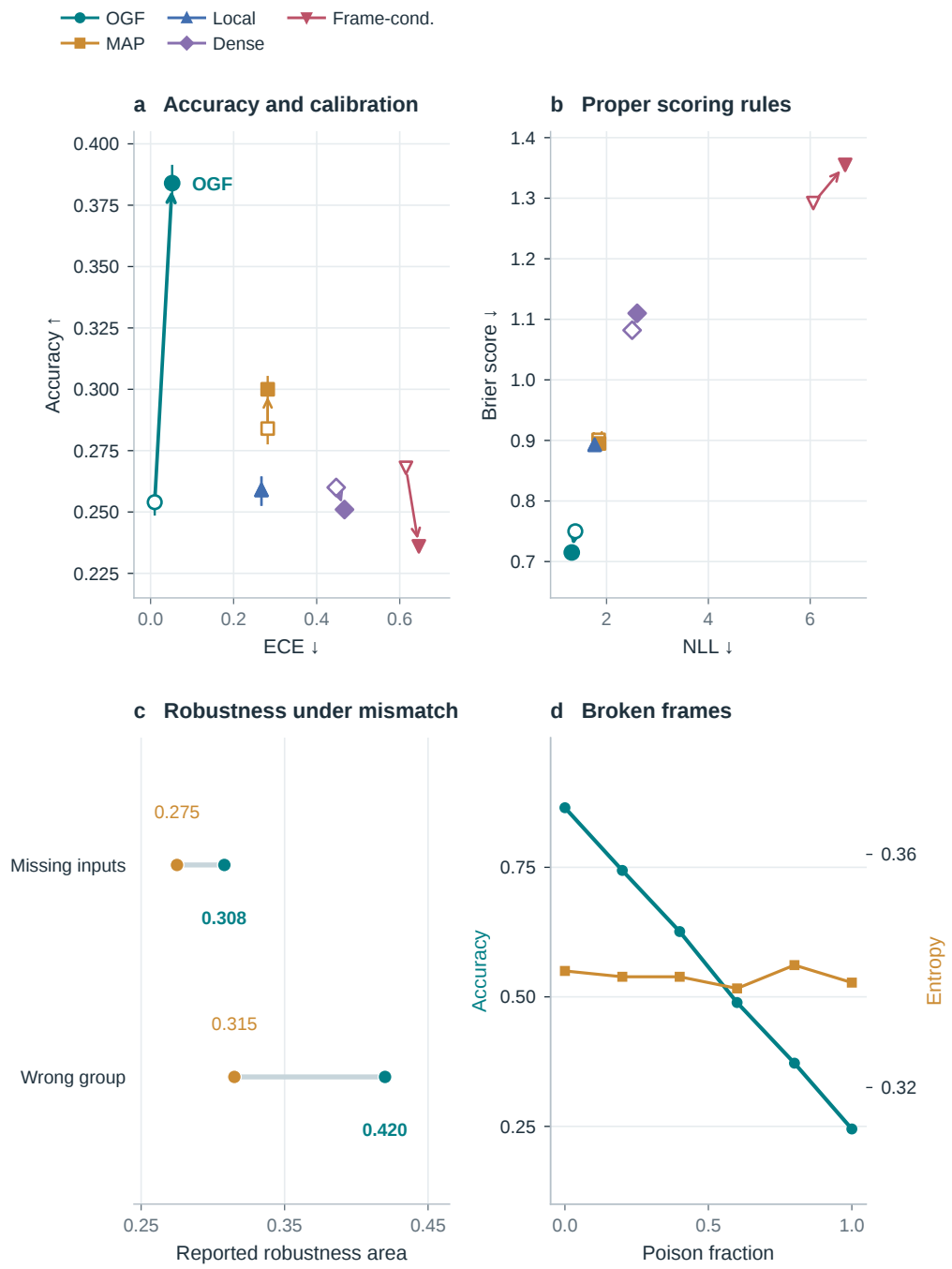


Figure 7: Ambiguity (open) versus conflict (filled). OGF is less accurate than MAP under pure ambiguity and better calibrated under conflict.

The ambiguity regime makes the true frame and a decoy indistinguishable (symmetric, conflict probability 0.5). The conflict regime keeps posterior mass on the true hypothesis while a confident decoy competes. Hard canonicalization receives the MAP frame; OGF averages a two-hypothesis posterior (table 8). This 10-seed table is the primary calibration claim. The no-group model has

ambiguity-regime NLL 1.397 and ECE 0.042: calibrated by ignorance, because it cannot use frame structure at all. Its informative separation from OGF is the accuracy collapse of the fusion table, not this ECE. A five-seed discrete world gives the same split (ambiguity accuracy 0.251 versus MAP 0.317; conflict 0.301 versus 0.250). A continuous-α sweep is supplementary: at full symmetry OGF ECE is 0.092 against 0.173 for hard MAP and 0.059 for a temperature-scaled MAP, so a tuned temperature is a competitive calibrator there. The calibration claim is the 10-seed table above.

Table 8: Ambiguity and conflict regimes (10 seeds, 600 steps). Accuracy cells include 95% CIs. OGF versus hard canonicalization paired $p = 0.00195$ for accuracy, NLL, ECE, and Brier in both regimes. Robustness areas: missing-modality averages $4, 3, 2, 1$ present modalities; wrong-group averages wrong-MAP fractions $0, 0.25, \ldots, 1$. OGF missing 0.308 $[0.304, 0.310]$ vs. hard 0.275 $[0.272, 0.279]$; misspec. 0.420 $[0.412, 0.426]$ vs. 0.315 $[0.313, 0.317]$.

Regime	Arm	Acc	NLL	ECE	Brier
Ambiguity	OGF	0.254 $[0.249, 0.258]$	1.386	**0.010**	0.750
	Hard canon. (MAP)	$\mathbf{0.284}$ $[0.278, 0.293]$	1.844	0.282	0.901
	Local	0.259 $[0.253, 0.264]$	1.766	0.267	0.893
	Dense / frame-cond.	0.260 / 0.268	2.502 / 6.061	0.447 / 0.615	1.082 / 1.292
Conflict	OGF	$\mathbf{0.384}$ $[0.377, 0.391]$	**1.316**	**0.052**	**0.715**
	Hard canon. (MAP)	0.300 $[0.295, 0.305]$	1.852	0.282	0.895
	Local	0.259 $[0.253, 0.264]$	1.766	0.267	0.893
	Dense / frame-cond.	0.251 / 0.236	2.601 / 6.687	0.467 / 0.646	1.110 / 1.355

Area	Hard canon.	OGF
Missing-modality ↑	0.275	**0.308**
Wrong-group ↑	0.315	**0.420**

Canonicalization commits to an increasingly wrong MAP as the conflict fraction grows, whereas OGF keeps mass on the remaining hypotheses; the missing area favors OGF for the same reason. OGF remains 1.24 nats above the oracle-frame NLL (0.151).

D Mechanism Dissection (Stage I): Factorial, Spectra, and Lattice Full Records

The factorial and lattice train and test on independent random frames (1024 train / 512 test per seed, 5 seeds). Spectra are one untrained forward; γ_{av} uses a shared-frame set. fig. 8 plots the factorial and the lattice; fig. 5 plots sensitivity and the coupling spectrum. An earlier factorial trained with a shared identity frame, where the sensors are already commensurate; a small MLP then fit the two dot products (0.90) faster than full OGF (0.77). That comparison is a capacity artifact of a world with no incommensurable frames.

Table 9: Mechanism tables (5 seeds). (a) Group×OT factorial on independent random frames. (b) Correspondence-matrix spectrum on one untrained forward. GMU: gated multimodal unit.

(a)

Group	Arm	Acc.
$+$	Full OGF (OT)	0.773
$+$	Group Transformer	0.773
$+$	Group MLP	**0.903**
$-$	Anchor-OT	0.250
$-$	Naive Transformer	0.245
$-$	Naive MLP	0.250

(b)

Coupling	Eff. rank	Top-1	Top-2
GMU gating	1.05	99.4%	99.5%
Perceiver (8 q.)	1.32	95.2%	96.5%
Q-Former (16 q.)	1.47	93.7%	94.9%
Linear attention	1.57	92.3%	93.8%
Dense cross-attn	4.04	67.5%	73.8%
OGF Sinkhorn	3.84	70.0%	74.8%

D.1 Factorial (group × OT)

Question. The no-group arm in the 10-seed world fell to chance. Is Sinkhorn enough without the group? *Arms* (250 steps). Full OGF $[+G, +OT]$; Anchor-OT $[-G, +OT]$ (same model, identity

frames, Sinkhorn intact); group Transformer $[+G, -OT]$ (vectors transported, then dense Transformer); naive Transformer $[-G, -OT]$; group MLP and naive MLP.



Controlled protocols are separate from Core96K. Triple subset: reported range only.
Panel d: signed Shapley terms; no claim that they sum to the reported synergy.

Figure 8: Independent-frame factorial, lattice centers, and Shapley values.

Per-seed accuracies. Full OGF $0.469, 0.979, 0.977, 0.479, 0.961$ (mean 0.773, CI $[0.571, 0.974]$); Anchor-OT $0.258, 0.254, 0.260, 0.225, 0.256$ (mean 0.250); group Transformer $0.785, 0.846, 0.770, 0.834, 0.633$ (mean 0.773); naive Transformer $0.270, 0.232, 0.234, 0.275, 0.211$ (mean 0.245); group MLP $0.885, 0.920, 0.922, 0.881, 0.908$ (mean 0.903, CI $[0.888, 0.918]$); naive MLP $0.240, 0.223, 0.254, 0.260, 0.275$ (mean 0.250). $p(\text{OGF} > \text{Anchor-OT}) = 0.0625$; OGF versus group Transformer $p = 1.0$. *Reading.* Every identity-frame arm is at chance. OT adds no accuracy on this inner-product label. Full OGF is bimodal (two seeds ≈ 0.47, three ≈ 0.97); the mean 0.773 has a wide interval. The group MLP's 0.903 is the task-specific ceiling once vectors already live in one frame.

D.2 Power-set lattice, Shapley values, and knockouts

Each of the 15 non-empty subsets is trained independently (250 steps, batch 128, AdamW 4×10^{-3}; $v(\emptyset) = 0.25$). Backbones: OGF; Dense Transformer (width 48, 4 heads, 2 layers); Perceiver (4 latents); Local MLP (hidden 64, frames included).

Table 10: OGF power set on independent frames (5 seeds). Non-equivariant full-set means stay at chance: Dense 0.235, Perceiver 0.243, Local MLP 0.254.

Subset	Acc.	95% CI	Reading
Text / Vision / Audio / Point	$0.234/0.234/0.244/0.238$	chance	singletons carry no bit
Text+Vision	0.429	$[0.396, 0.461]$	first geometric pair
Audio+Point	0.332	$[0.225, 0.445]$	second pair; interval includes chance
Text+Audio	0.268	$[0.254, 0.281]$	non-geometric pair
Text+Point / Vis+Aud / Vis+Pt	$0.260/0.247/0.245$	near chance	same
Triples	0.416–0.475	—	above the best pair, below full
All four	0.963	$[0.947, 0.976]$	both relations at once

Synergy and exact Shapley values are table 11. Knockouts on the trained full model (clean 0.9688 $[0.956, 0.980]$): shuffling vision pairings gives 0.455 and identity frames 0.243 ($p = 0.0625$ each). Untyped full sets never leave chance, so their Syn is negative.

Table 11: Lattice synergy and exact Shapley values (5 seeds; $p = 0.0625$). OGF Syn CI $[+0.464, +0.488]$.

Backbone	Syn	ϕ_{Text}	ϕ_{Vis}	ϕ_{Aud}	ϕ_{Pt}
Local MLP	-0.020	$+0.004$	$+0.004$	-0.003	-0.001
Perceiver	-0.032	-0.005	$+0.001$	-0.004	$+0.002$
Dense Transformer	-0.038	-0.002	$+0.003$	-0.008	-0.007
OGF	$\mathbf{+0.475}$	$\mathbf{+0.199}$	$\mathbf{+0.179}$	$\mathbf{+0.174}$	$\mathbf{+0.161}$

Table 12: Test-time modality dropout. Retention AUC is the mean of the five accuracies. Dropout-trained Dense zeros one or two streams at train time.

Arm	AUC	p=0	0.2	0.4	0.6	0.8
Dense Transformer	0.249	0.250	0.252	0.247	0.250	0.246
Local MLP	0.253	0.252	0.253	0.255	0.254	0.250
Perceiver	0.258	0.262	0.259	0.256	0.257	0.255
Dropout-trained Dense	0.259	0.259	0.259	0.259	0.261	0.256
OGF	**0.536**	**0.965**	**0.644**	**0.468**	**0.344**	**0.258**

Table 13: Frame poisoning on the trained full OGF. The poison-0 mean 0.865 includes one collapsed seed; knockout clean is 0.9688. Entropy does not approach $\ln 4 \approx 1.386$.

Poison fraction	0	0.2	0.4	0.6	0.8	1
Accuracy	0.865	0.744	0.626	0.489	0.372	0.245
Posterior entropy	0.340	0.339	0.339	0.337	0.341	0.338

D.3 Correspondence spectrum and gradient competition

Spectrum. Random tokens ($B = 256$, 16 tokens per side, $d = 48$), one untrained forward; effective rank $\exp H(\sigma/\sum\sigma)$. Row entropy is ≈ 2.76 for every arm and does not discriminate. OGF's leading singular value is the uniform mass $1/16$, so its spectrum has a different scale from softmax attention: "more full-rank than Dense" is not a meaningful comparison (Dense rank $4.04 > 3.84$). Sinkhorn doubly-stochastic residual is $\max 6.33 \times 10^{-8}$, mean 6.11×10^{-8}. OGF versus Q-Former

on rank and top-2 energy has $p = 0.0625$. *Gradients.* 150 steps on a shared-frame dataset (to isolate competition from extrapolation), batch 64, AdamW 3×10^{-3}; final γ_{av} averages the last 20 steps. Dense 1.55×10^6 (starved on 100% of steps, CI $[1.51, 1.58] \times 10^6$); late fusion 1.006 $[0.987, 1.025]$; GMU 1.045 $[1.007, 1.102]$; OGF 1.038 $[0.180, 2.151]$ with 46.7% of steps in $[0.5, 2]$. Individual OGF final-20 means are 0.055, 1.19, 0.74, 3.18, 0.023. One seed has $87/150$ steps with $\gamma_{\text{av}} < 0.05$; averaging the five seed-wise step fractions $0, 0, 0, 0, 58\%$ gives the 11.6% step-level starvation rate.

Table 14: Input-gradient ratio $\gamma_{\text{av}} = \|\nabla_{\text{aud}}\|/\|\nabla_{\text{vis}}\|$ over 150 shared-frame steps. Starvation is $\gamma_{\text{av}} < 0.05$ or $\gamma_{\text{av}} > 20$.

Arm	Final γ_{av} [95% CI]	Starved	Balanced $[0.5, 2]$
Dense cross-attn	1.55×10^6 $[1.51, 1.58] \times 10^6$	100%	—
Late fusion	1.006 $[0.987, 1.025]$	0%	—
GMU	1.045 $[1.007, 1.102]$	0%	—
OGF Sinkhorn	1.038 $[0.180, 2.151]$	11.6%	46.7%

Reading. Resamplers are low-rank bottlenecks and dense attention starves a modality at Dense scale. OGF removes that 10^6 starvation and the query-rank collapse; it is not uniquely balanced, and one seed still spends 58% of steps below 0.05. Dense rank 4.04 is *higher* than Sinkhorn 3.84; the spectra are on different scales. Balance of γ_{av} is shared with late fusion and GMU.

E Frozen-Encoder Protocol Index

Full protocols are in App. F, one subsection each. The table names which Stage II question isolates which group.

Protocol	Group or clock	Primary table	Other protocol
CLIP RefCOCO referring	frozen CLIP referring	table 15	discrete C_4 token referring
Qwen2-VL C_4 referring	discrete 4-orbit	App. F	C_4-augmented Q-Former
TUM metric pose	RGB–depth relative pose	table 17	own table, apart from yaw bins
TUM yaw bins	$\mathrm{SO}(3)$ geodesic residual	table 18	own table, apart from capacity-fair
TUM up/down, capacity-fair	$\mathrm{SO}(3)$	table 19	own table, apart from yaw bins
TUM RGB–depth conflict	mismatched RGB	table 26	conflict AUROC, not pose error
CREMA circular time	circular sync / missing AP	table 16	width varied at fixed steps
AVE held-out OOD	cyclic C_{10}	table 20	clean alignment, no shift
AVE eleven-shift retention	same features	table 21	own table, apart from held-out OOD AP
AVE family comparison	family AP 0.662 vs. 0.611	App. F	own table, apart from either AVE clock
ScanRefer viewpoint-OOD	continuous $\mathrm{SE}(3)$ centres	tables 22 and 23	trunk stays frozen
QVHighlights official / 300M	moment retrieval	tables 24 and 25	LongVALE F1/IoU does not improve
SigLIP2 / six frozen-encoder tasks	frozen-encoder retention	App. F	matched head on the same six tasks

F REAL FROZEN FEATURES (STAGE II): FULL RECORDS FOR SENSORS AND GROUPS

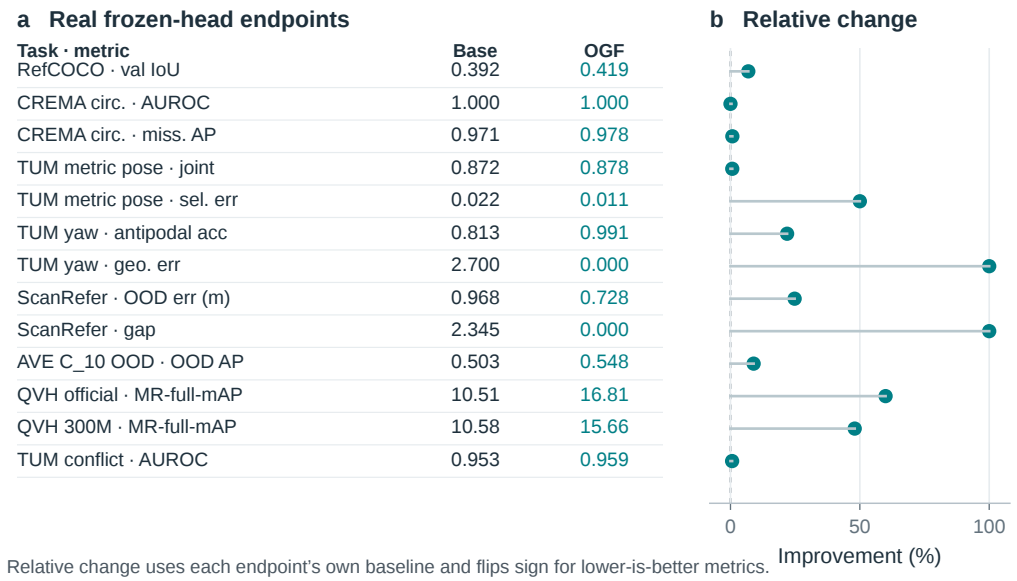


Figure 9: Stage II frozen-encoder task heads. Each relative change is computed inside that task.

Each subsection states one protocol. Protocols that share a dataset and ask different questions stay in separate tables. table 2 gives one strongest competitor per row. Every arm for those nine rows is in tables 15 to 18, 20, 22 and 24 to 26. Camera-frame points p^{cam} and a pose-derived gravity g satisfy $(p_A^{\text{cam}} - p_B^{\text{cam}}) \cdot g \equiv z_A^{\text{world}} - z_B^{\text{world}}$, so an $\mathrm{SO}(3)$-invariant $\ell = 1$ contraction recovers the label at every camera orientation. Discrete C_4 is the complementary regime: the orbit is cheap to cover. Continuous $\mathrm{SO}(3)/\mathrm{SE}(3)$ and cyclic time are the regimes in which exact transport, rather than orbit coverage, is the source of the gain.

F.1 IMAGE–TEXT REFERRING (REFCOCO, FROZEN CLIP VIT-B/32)

Question. On frozen CLIP patches (Yu et al., 2016; Radford et al., 2021), does a typed head still beat a matched Q-Former and the best independently trained subset? *Protocol.* OGF typed 727,230 parameters, Q-Former 760,580, dense 727,173, semantic-only 693,765; 10 seeds; 10,000 steps; batch 64. Independently trained subsets cross input mode (image or text) with the four arms. *Result.* Strongest subset is text×OGF (val IoU 0.314); image-only arms 0.273–0.276. Full-modality val IoU / Acc@0.5: OGF 0.419 / 0.373, Q-Former 0.392 / 0.326, dense 0.399 / 0.325, semantic 0.397 / 0.326; testA 0.471 versus 0.442; testB 0.378 versus 0.358. Paired gains, all $p = 0.00195$: over the strongest subset +10.46pp [10.13, 10.74]; over Q-Former +2.68 [2.46, 2.91]; Acc@0.5 +4.70; patch recall@1 +9.46. Per-seed val IoU: 0.412, 0.421, 0.424, 0.422, 0.418, 0.417, 0.421, 0.420, 0.417, 0.419. On the same CLIP family, a smaller head trained for 400 steps loses to 4× test-time augmentation (TTA; App. G).

Table 15: RefCOCO, frozen CLIP (10 seeds). TestA / TestB are given for OGF versus Q-Former. Image-only arms span 0.273–0.276.

Arm	val IoU	Acc@0.5	TestA / TestB
OGF	**0.419**	**0.373**	0.471 / 0.378
dense	0.399	0.325	–
semantic	0.397	0.326	–
Q-Former	0.392	0.326	0.442 / 0.358
text×OGF (strongest subset)	0.314	–	–

F.2 AUDIO–VISUAL CIRCULAR TIME (CREMA-D)

Question. After a circular shift policy on CREMA-D (Cao et al., 2014) removes a pairwise shortcut, does group time still beat single streams and stay non-inferior on clean AP? *Protocol.* OGF with banded temporal alignment and hard canonical 1,716,788 parameters, dense 1,717,291, late 1,323,049; 10 seeds. *Result.* Sync AUROC: OGF 1.000, hard 1.000, dense 0.9998, late 0.500; single streams ≤ 0.506 ($+49.45$pp over the strongest subset). Missing-event AP: OGF 0.978, hard 0.971, dense and late 0.949 ($+0.64$ $[0.58, 0.69]$ over hard). Clean AP versus hard -0.24 $[-0.26, -0.21]$, inside the pre-specified -1pp non-inferiority margin. This comparison does not measure robustness to white-noise SNR.

Table 16: CREMA-D circular time (10 seeds). Clean AP is OGF$-$hard -0.24pp, inside the -1pp non-inferiority margin.

Arm	Sync AUROC	Missing-event AP
OGF	**1.000**	**0.978**
hard	1.000	0.971
dense	0.9998	0.949
late	0.500	0.949
single streams	≤ 0.506	–

Under white noise, sync AUROC for OGF with banded temporal alignment stays near chance: 0.510 / 0.516 / 0.527 / 0.542 / 0.544 / 0.545 at $-10/-5/0/+5/+10/+20$ dB. The -10 dB interval is $[0.438, 0.569]$ and includes 0.5. Dense is 0.503–0.510; late and uniform stay at 0.500; dual-encoder cosine inverts (0.439 down to 0.362). The circular-sync result is table 16.

F.3 METRIC POSE (TUM RGB-D)

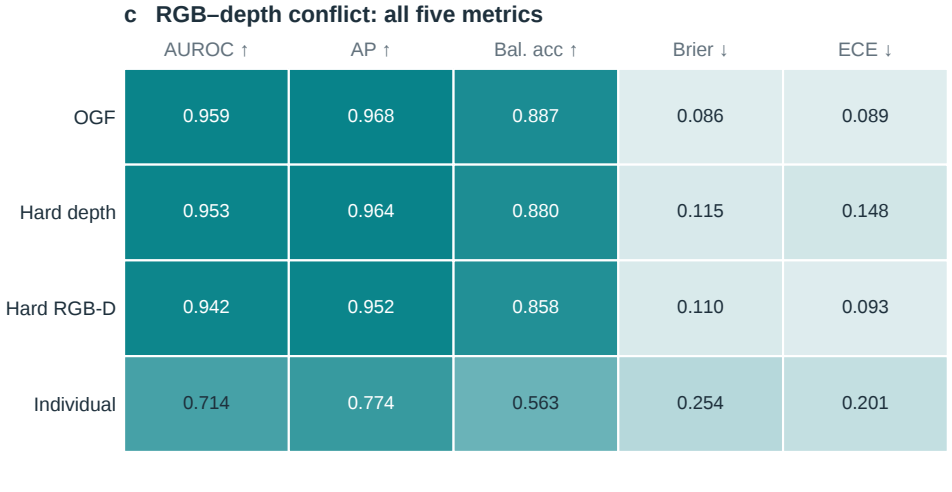


Figure 10: TUM RGB–depth conflict: AUROC, average precision, balanced accuracy, Brier score, and calibration error.

Question. On held-out TUM RGB-D frames (Sturm et al., 2012), does a posterior head beat the strongest unimodal subset and stay close to a hard RGB-D hybrid? *Protocol.* Relative pose with 5° / 5 cm success under a mixture of clean, RGB-occluded, and depth-dropout inputs; 10 seeds. *Result.* Joint success: OGF posterior 0.878, hard hybrid 0.872, RGB-D SIFT-PnP 0.872, RGB-only 0.433, depth-only flow-PnP 0.394 (strongest subset), depth-only ICP 0.256, learned depth geometry 0.234. Selective error at 50% coverage: 0.011 versus 0.022. OGF scores 1.0 on clean frames in all ten seeds (joint / rotation / translation). Relative to the strongest subset the joint gain is $+48.4$pp; relative to hard hybrid $+0.61$pp. Hard hybrid has slightly better Brier (-0.0026, $p = 0.0039$).

Table 17: TUM metric pose (10 seeds). Selective error at 50% coverage is recorded for OGF and hard hybrid only.

Arm	Joint success	Sel. err (50%)
OGF posterior	**0.878**	**0.011**
hard hybrid	0.872	0.022
RGB-D SIFT-PnP	0.872	–
RGB-only	0.433	–
depth-only flow-PnP (strongest subset)	0.394	–
depth-only ICP	0.256	–
learned depth geometry	0.234	–

F.4 Continuous SO(3): yaw bins and a capacity-fair counterpart

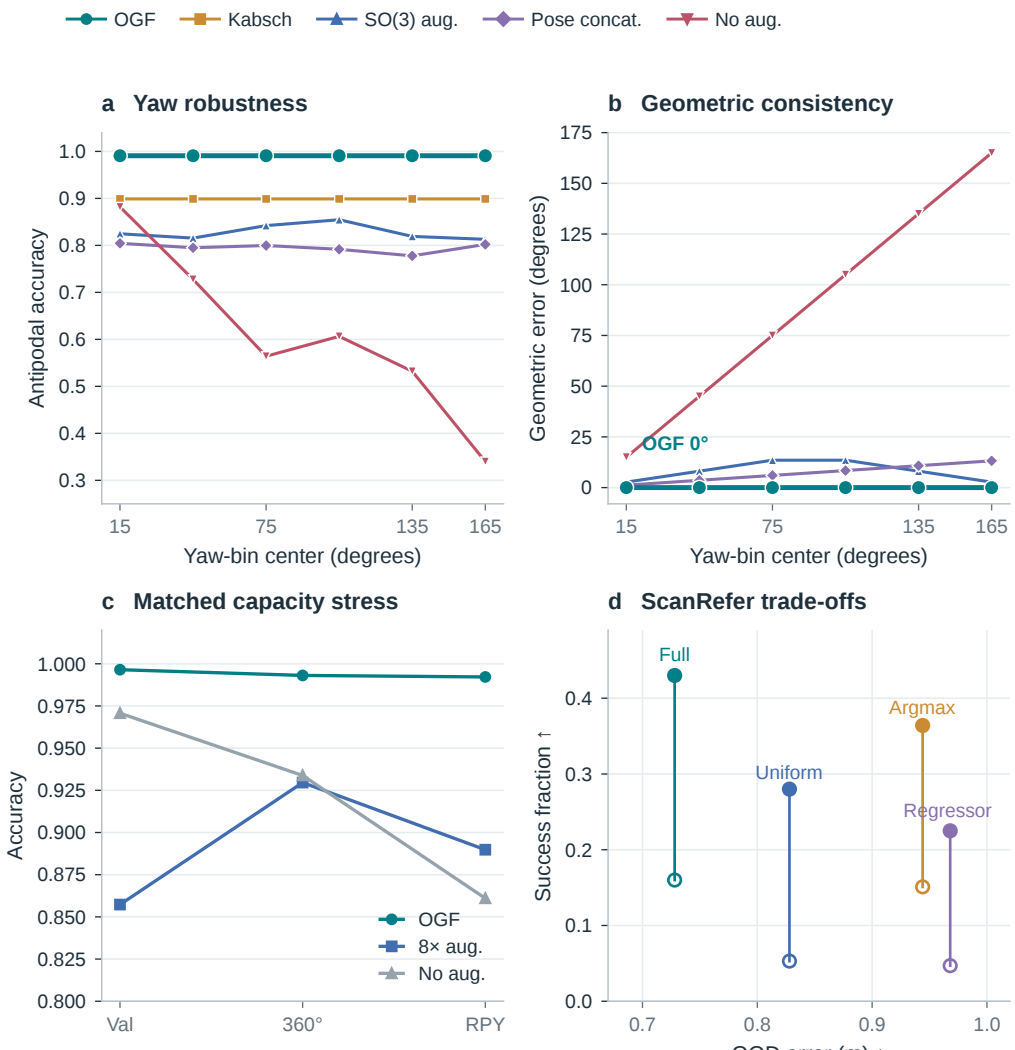


Figure 11: TUM yaw, capacity-fair centers, and ScanRefer OOD.

Two protocols share TUM RGB-D and do not share a question. The five-arm protocol asks whether geodesic error is identically zero across yaw. The three-arm protocol asks whether a deliberately

smaller OGF head still beats a capacity-favored augmented regressor on the held-out full-circle sequence.

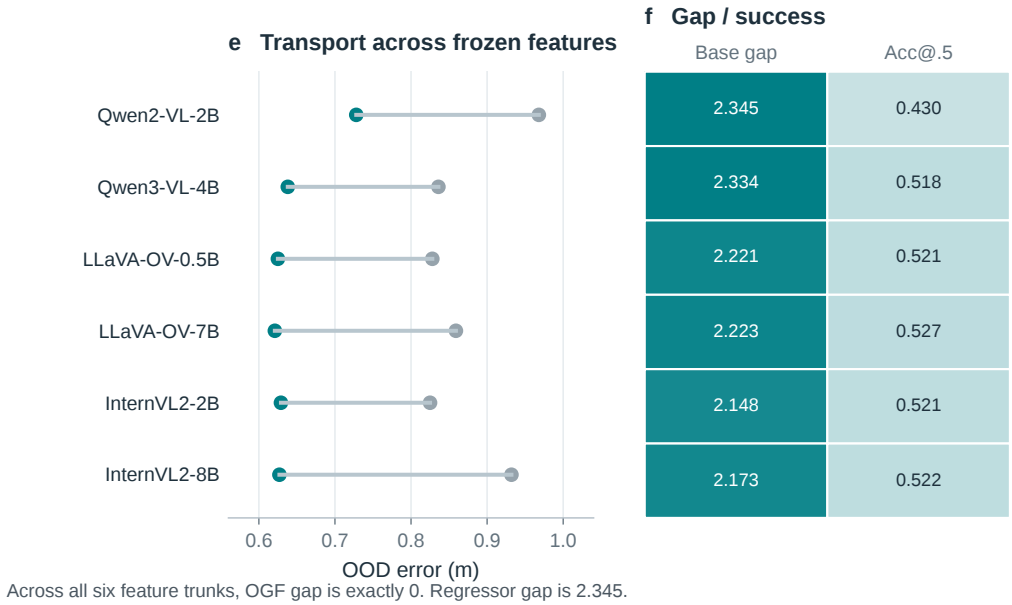


Figure 12: Six frozen-feature trunks: OOD error, base gap, and success.

Five-arm yaw bins. Sequences `rgbd_dataset_freiburg1_{desk,360}`. Inputs: camera-frame point pairs, gravity, and grid-mean colour. Label: world Δz in three bins (threshold 0.05 m). Train: desk, 200 frames $\times$ 40 pairs, 85% train, 1500 steps. Test: 360 sequence, 150 frames $\times$ 40 pairs, six 60° yaw bins. Arms: OGF typed $\ell = 1$ (width 64); no-augmentation and $\mathrm{SO}(3)$-augmented MLPs (width 256); pose concatenation (explicit R); Kabsch oracle (ideal pre-registration, then classification).

Table 18: TUM yaw bins: accuracy (geodesic error). OGF is identical in every bin because the typed $\ell = 1$ invariant is exactly $\mathrm{SO}(3)$-invariant.

Yaw bin	OGF	Kabsch	$\mathrm{SO}(3)$ aug.	Pose concat.	No aug.
0°–30°	**0.9908** (0.0°)	0.8989 (0.0°)	0.8247 (2.7°)	0.8044 (1.2°)	0.8818 (15.0°)
30°–60°	0.9908 (0.0°)	0.8989 (0.0°)	0.8155 (8.1°)	0.7950 (3.6°)	0.7277 (45.0°)
60°–90°	0.9908 (0.0°)	0.8989 (0.0°)	0.8421 (13.5°)	0.7997 (6.0°)	0.5641 (75.0°)
90°–120°	0.9908 (0.0°)	0.8989 (0.0°)	0.8546 (13.5°)	0.7917 (8.4°)	0.6064 (105.0°)
120°–150°	0.9908 (0.0°)	0.8989 (0.0°)	0.8191 (8.1°)	0.7777 (10.8°)	0.5321 (135.0°)
150°–180°	**0.9908** (0.0°)	0.8989 (0.0°)	0.8129 (2.7°)	0.8021 (13.2°)	0.3399 (165.0°)

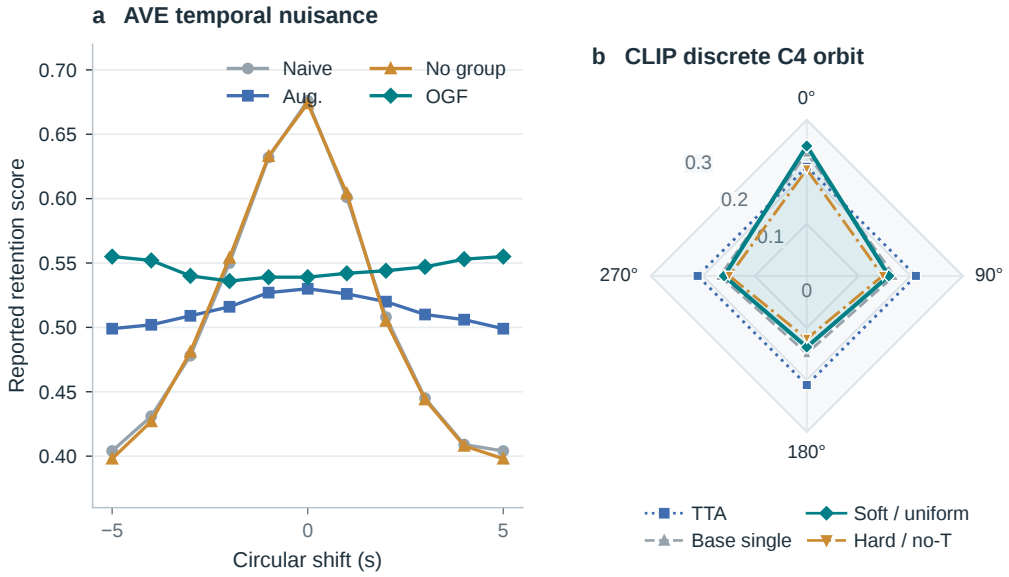
Kabsch matches OGF's 0° residual and still loses 9 points of accuracy: the oracle registers and then classifies without a typed invariant readout.

Capacity-fair three-arm protocol. Task: world up/down from a pose-derived gravity axis. OGF: width 64, 5,123 parameters, 1500 steps, 200 train frames, 40 pairs. Baselines: width 512, 8000 steps, with and without $\mathrm{SO}(3)$ augmentation. Five seeds.

Table 19: Capacity-fair TUM protocol (mean±std). 360-seq. is accuracy on the held-out full-circle sequence; RPY is held-out roll–pitch–yaw. Equivariance gap is the score change under a random test $SO(3)$.

Arm	Val	360-seq.	RPY	Equiv. gap
OGF	0.9965 ± 0.0021	$\mathbf{0.9931 \pm 0.0006}$	0.9922 ± 0.0008	$\mathbf{0.0000}$
$8\times$ $SO(3)$ aug.	0.8573 ± 0.0764	0.9296 ± 0.0305	0.8898 ± 0.0288	0.0194 ± 0.0156
No aug.	0.9707 ± 0.0105	0.9338 ± 0.0019	0.8610 ± 0.0012	0.4671 ± 0.1790

F.5 Cyclic time: held-out OOD and eleven-shift retention



AVE source does not name its retention metric. CLIP radii share a zero-based 0–0.30 IoU scale.

Figure 13: AVE eleven-shift retention and the CLIP C_4 radar.

Frozen CLIP visual and Audio Spectrogram Transformer (AST) audio per-second features on AVE (Tian et al., 2018). The group is C_{10}. The two protocols share those features and do not share a metric.

Held-out OOD shifts. Train shifts $\{-2, \ldots, 2\}$, OOD $\{3, 4, 5\}$; 5 seeds. Clean AP / acc.: OGF $0.533/0.568$, aug. $0.521/0.562$, naive $0.679/0.656$, no-group $0.516/0.555$. OOD AP / acc.: $0.548/0.576$, $0.503/0.551$, $0.422/0.481$, $0.502/0.546$. Expected AP: 0.540, 0.507, 0.507, 0.506. OGF offset recovery 0.163. When streams are already aligned, naive clean accuracy is 0.656 and OGF is 0.568.

Table 20: AVE held-out C_{10} OOD (5 seeds). Naive leads on clean accuracy; OGF leads on held-out OOD AP. Offset recovery for OGF is 0.163.

Arm	Clean AP / acc.	OOD AP / acc.	Expected AP
OGF	0.533 / 0.568	$\mathbf{0.548}$ / 0.576	0.540
aug.	0.521 / 0.562	0.503 / 0.551	0.507
naive	$\mathbf{0.679}$ / $\mathbf{0.656}$	0.422 / 0.481	0.507
no-group	0.516 / 0.555	0.502 / 0.546	0.506

Eleven-shift retention. Partial events of length 2–8 s; train 1036 / test 128; 1200 steps, batch 128, Adam 10^{-3}, BCE; OGF and aug. train only on $\{-2, \ldots, +2\}$; five seeds. Retention AUC is the mean of the eleven APs.

Table 21: AVE event AP by circular shift. Retention AUC: OGF 0.5457 $[0.536, 0.555]$, aug. 0.5131, naive 0.5036, no-group 0.5024 ($p = 0.0625$ vs. each).

Δt (s)	Naive	Aug.	No group	OGF
-5 / $+5$	0.404	0.499	0.398	**0.555**
-4 / $+4$	0.431 / 0.409	0.502 / 0.506	0.427 / 0.408	**0.552** / **0.553**
-3 / $+3$	0.478 / 0.445	0.509 / 0.510	0.481 / 0.444	**0.540** / **0.547**
-2 / $+2$	0.550 / 0.508	0.516 / 0.520	0.554 / 0.505	0.536 / **0.544**
-1 / $+1$	0.632 / 0.601	0.527 / 0.526	0.633 / 0.604	0.539 / 0.542
0	**0.676**	0.530	0.674	0.539

OGF offset hit-rate by Δt: 0.147, 0.084, 0.083, 0.102, 0.088, 0.147, 0.084, 0.083, 0.102, 0.088, 0.000 from -5 to $+5$. Offset MAE at $|\Delta t| = 5$ is 4.5–5.5 s. Flat event AP does not recover the offset. A separate AVE family comparison (AP 0.662 vs. 0.611) misses a pre-specified recall@1 threshold ($0.829 < 0.90$).

F.6 3D REFERRING (SCANREFER SE(3))

On ScanRefer (Chen et al., 2020a), frozen Qwen2-VL-2B (Wang et al., 2024) merged vision tokens on a common 8×8 grid (image size 224), depth-lifted to camera-frame points; frozen LM text states; the trunk is not trained. Train only on $|\mathrm{yaw}| < 90^\circ$ (9.7k); evaluate in-distribution 9.6k and OOD 10.0k. OGF 786,945 parameters; matched regressor 920,580. Five seeds. Main backbone (mean±std): OGF IND err. 0.786 ± 0.004, OOD 0.728 ± 0.004, acc@$0.5 = 0.430$, rotation gap 0, 8-view orbit err. 0.728. Regressor: IND 1.002 ± 0.008, OOD 0.968 ± 0.008, acc@$0.5 = 0.225$, rotation gap 2.345 ± 0.220, orbit 2.116. Accuracies at $0.25/0.5/1.0$ m: $0.160/0.430/0.773$ versus $0.047/0.225/0.627$. Component ablations and the six-family sweep are tables 22 and 23. Relative OOD-error cuts: 24.8, 23.7, 24.4, 27.7, 23.8, 32.7%. The same small head reads frozen features; those trunks stay frozen.

Table 22: ScanRefer viewpoint-OOD variants (5 seeds, matched budget). Gap is the change in localization error under a test rotation. The last row is a matched non-equivariant regressor.

Variant	OOD err ↓	acc@.25	acc@.5	Gap ↓
OGF (full)	**0.728**	**0.160**	**0.430**	**0.000**
− posterior (argmax)	0.944	0.151	0.364	**0.000**
− semantic corr. (uniform)	0.828	0.053	0.280	**0.000**
Non-equiv. regressor	0.968	0.047	0.225	2.345

Table 23: ScanRefer viewpoint-OOD with the same small head on six frozen vision–language feature families (common 8×8 grid, 5 seeds). The trunks stay frozen.

Frozen features	OGF err ↓	Base err ↓	OGF gap ↓	Base gap	acc@.5
Qwen2-VL-2B	0.728	0.968	**0.000**	2.345	0.430
Qwen3-VL-4B	0.638	0.836	**0.000**	2.334	0.518
LLaVA-OV-0.5B	0.625	0.828	**0.000**	2.221	0.521
LLaVA-OV-7B	0.621	0.859	**0.000**	2.223	0.527
InternVL2-2B	0.629	0.825	**0.000**	2.148	0.521
InternVL2-8B	0.627	0.932	**0.000**	2.173	0.522

F.7 MOMENT RETRIEVAL (QVHIGHLIGHTS)

The official-feature protocol and the 300M end-to-end protocol are two rows (Lei et al., 2021). Neither is a 7B Omni result.

Official Moment-DETR features. Only the official feature family; the official Moment-DETR evaluator; 1550 validation queries; five seeds; 10,000 bootstrap resamples; Holm correction against matched baselines. MR-full-mAP: OGF 16.81 (Recall@1 at IoU 0.5 is 21.25, at 0.7 is 12.41, HL-mAP 31.98, HL-Hit@1 49.69), dense 10.51, early 10.11, video-only 9.88, Q-Former 9.75, late 7.00, text-only 6.95. OGF − dense +6.30 $[5.25, 7.37]$, Holm $p = 4 \times 10^{-5}$; − video-only +6.93 $[6.17, 7.71]$; − early / Q-Former / late $+6.70/ + 7.06/ + 9.81$. Highlight Hit@1 − dense +0.31 $[-0.57, 1.13]$ (interval includes 0). Perturbations on this protocol are reported as *drops* from clean: query shuffle −0.123, wrong time −0.175.

Table 24: QVHighlights official Moment-DETR features (5 seeds). OGF also records Recall@1 at IoU $0.5/0.7$ $21.25/12.41$, HL-mAP 31.98, HL-Hit@1 49.69.

Arm	MR-full-mAP
OGF	**16.81**
dense	10.51
early	10.11
video-only	9.88
Q-Former	9.75
late	7.00
text-only	6.95

300**M-parameter encoder.** Five seeds. MR-full-mAP: OGF 15.66, early fusion 10.85, world anchors without group transport 10.58, dense 10.49, Q-Former/Perceiver 10.48, same trunk without OGF 10.11, video-only 9.38, late 6.64, text-only 6.53 (Holm $p = 1.7 \times 10^{-4}$). Relative to dense +5.16pp, early fusion +4.81, no-group +5.07, video-only +6.27, text-only +9.13. These perturbations are absolute MR-full-mAP: query shuffle stays at 15.66; query swap falls to 14.80; wrong time rises to 18.84, above the clean 15.66.

Table 25: QVHighlights with a 300M-parameter encoder (5 seeds). No-group is world anchors without group transport.

Arm	MR-full-mAP
OGF	**15.66**
early	10.85
no-group	10.58
dense	10.49
Q-Former / Perceiver	10.48
same trunk without OGF	10.11
video-only	9.38
late	6.64
text-only	6.53

On LongVALE (Geng et al., 2025) the same pairing and time shifts lower the score, and downstream F1 and IoU do not improve. Moment mAP scores highlight alignment; it is not open-vocabulary localization.

F.8 Real sensor conflict (TUM RGB–depth)

Same sequence; depth and ground-truth pose fixed; RGB mismatched; the task is to detect the conflict. Ten seeds.

Table 26: TUM RGB–depth conflict detection (10 seeds).

Arm	AUROC	AP	Bal. acc.	Brier	ECE
OGF group consistency	**0.9592**	**0.9681**	0.887	**0.086**	**0.089**
Hard depth transport	0.9529	0.9635	0.880	0.115	0.148
Hard RGB-D transport	0.9421	0.9521	0.858	0.110	0.093
Individual quality	0.7135	0.7745	0.563	0.254	0.201

OGF AUROC by seed: 0.973, 0.979, 0.979, 0.962, 0.905, 0.918, 0.967, 0.981, 0.970, 0.957. Together with the controlled wrong-group area (App. C) and lattice poisoning (App. D), this is three levels of the same claim. It does not cover arbitrary semantic fabrication.

F.9 Frozen-encoder scale

These three checks ask whether a frozen-encoder gain is an accident of one width.

CREMA 98**M /** 301**M /** 1.00**B.** Fixed budget, five seeds each, against the validation-selected matched fuser. Missing-AP $+1.03$, $+0.76$, $+1.70$pp (intervals exclude 0); clean -0.11 to -0.13pp; runtime ratios 1.100, 1.063, 1.031. The comparison is a fixed step budget at three widths.

Frozen SigLIP2 scales. Widths 375M, 882M, 1B (Tschannen et al., 2025); 45 training runs. OGF clean and C_4 IoU agree to 4×10^{-6}. Pooled over scales, OGF exceeds C_4-augmented Q-Former by $+0.0051$ $[0.0023, 0.0082]$ and a four-view orbit by $+0.018$. Clean IoU is within 0.003 of standard Q-Former. Uncompiled inference is slower (1.43, 1.22, $1.70\times$ a compiled forward); the compiled path reaches 1.0035.

Six frozen-encoder tasks. Ten paired seeds each. Mean drop 0.321pp $[0.271, 0.370]$; the largest drop is CREMA audio-only average precision (1.42pp). OGF improves text grounding by 6.3pp. These checks use frozen-encoder heads. Generative-trunk changes are reported with Core96K.

F.10 Discrete C_4 on frozen Qwen2-VL tokens

Complementary regime: the orbit has four elements and augmentation can tile it. Frozen Qwen2-VL-2B merged vision tokens (row-major 8×8, dim 1536) and frozen LLM text states; 10 seeds, 29k train examples, 8000 steps. OGF val IoU is 0.453 on both clean and C_4 charts (drift $\leq 4 \times 10^{-8}$; a paired check differs by 5×10^{-9}). A C_4-augmented Q-Former reaches 0.435–0.437. OGF minus that Q-Former is $+1.75$pp with an interval that crosses zero ($p = 0.29$). Remaining claims that survive the tie: exact invariance, a single forward that beats a four-view orbit ensemble by $+1.4$ to $+2.2$ IoU at one quarter the inference compute, and chart-invariant calibration ($\Delta\mathrm{ECE} = 0.0000$). When augmentation can tile the orbit, a tie is the prediction.

G Generative Embedding Probes and the Native Frame Check

These runs test whether a module or an unmodified trunk changes its output under rotation.

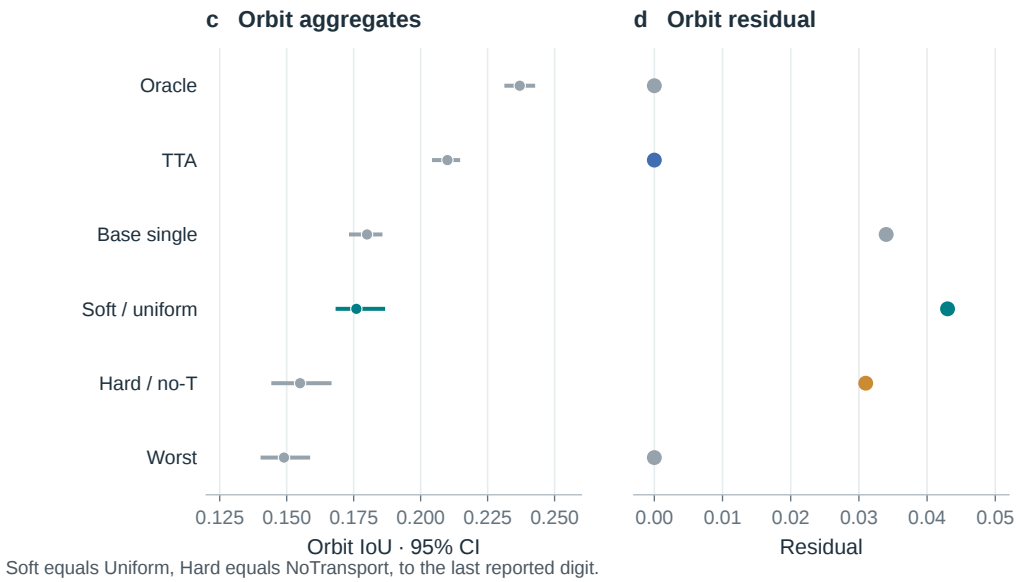


Figure 14: Frozen-CLIP orbit aggregates and residuals.

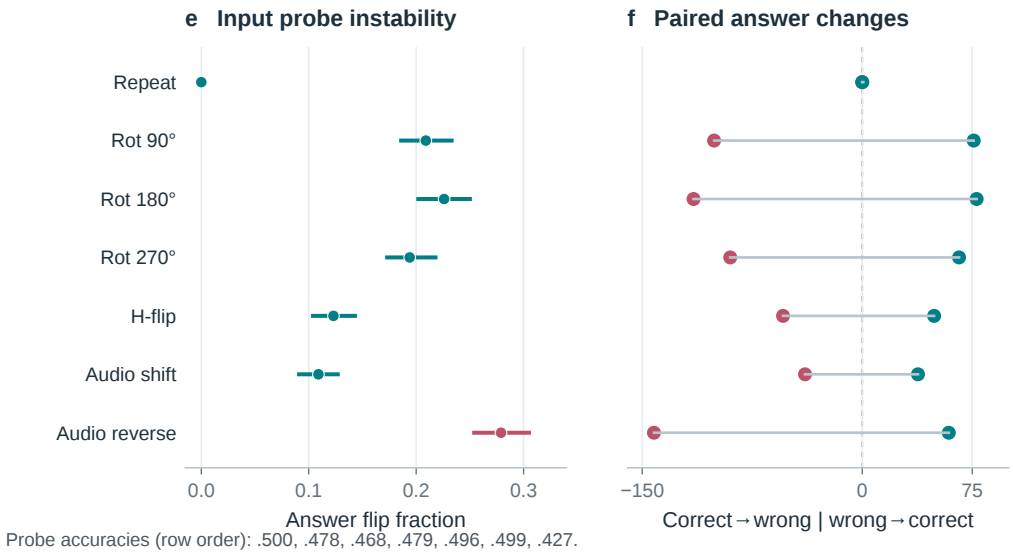


Figure 15: Native-frame probe flips and paired answer changes.

G.1 OMNI-7B C_4 ORBIT MODULE

Both arms fine-tune Qwen2.5-Omni-7B with LoRA ($r = 16$, $\alpha = 32$, dropout 0.05, q, v) for 4000 RefCOCO steps; OGF adds an orbit module (21.4M versus 7.67M trainable). On val ($n = 10{,}834$), clean IoU is 0.755 (OGF) versus 0.771; under a $90°$ rotation it is 0.447 versus 0.186 (Acc@0.5 $0.877 \rightarrow 0.069$ for LoRA). Clean-to-rotated IoU drops by 0.308.

G.2 Frozen single-forward geometry on CLIP

Question. If CLIP is frozen and only a 400-step head is trained, can a soft C_4 posterior beat $4\times$ test-time augmentation? *Protocol.* Cached CLIP ViT-B/32 7×7 patches; 1000 train / 1709 eval from a 2709-item pool (expressions that state a spatial relation, 1417 in the pool, and expressions that state appearance, 1292; slices are on the eval subset); 5 seeds; 400 steps; batch 64; AdamW 3×10^{-3}; L1 box regression. Oracle / Worst are the same Base head's max / min over the four C_4 rotations $k \in \{0, 1, 2, 3\}$, not independent models.

Table 27: Orbit-mean IoU on frozen CLIP (5 seeds). Soft (posterior-averaged C_4) matches uniform weights to the printed decimals, and hard (single frame, no transport) matches those same decimals.

Arm	Orbit IoU	95% CI	σ_{orbit}	0°	90°	$180/270$
Base Oracle	0.237	$[0.232, 0.242]$	0	—	—	—
Base+4× TTA	**0.210**	$[0.205, 0.214]$	≈ 0	0.210	0.210	0.210 / 0.210
Base single	0.180	$[0.174, 0.185]$	0.034	0.237	0.168	0.149 / 0.167
OGF Soft / Uniform	0.176	$[0.169, 0.186]$	0.043	0.250	0.158	0.137 / 0.159
OGF hard / no transport	0.155	$[0.145, 0.166]$	0.031	0.205	0.146	0.121 / 0.149
Base Worst	0.149	$[0.141, 0.158]$	0	—	—	—

OGF Soft vs. TTA $p = 0.0625$ (TTA higher); vs. Base single $p = 0.5625$. On the eval subset at 90°, spatial-relation OGF 0.133 vs. Base 0.144 ($\Delta = -0.012$, $p = 0.25$); appearance 0.187 vs. 0.193 ($\Delta = -0.007$, $p = 0.375$). There is no advantage on spatial-relation expressions. On this frozen CLIP head, $4\times$ test-time augmentation reaches higher orbit IoU than the soft C_4 posterior (table 15).

G.3 Native Omni-7B frame consistency on OmniBench

Question. Does the unmodified generative trunk change its answer under meaning-preserving nuisances? OGF is not in this path. *Setup.* Frozen Qwen2.5-Omni-7B Thinker (multimodal encoder), fp16, greedy decoding, all 1142 items, prompt restricted to A/B/C/D; one modality is changed at a time; item bootstrap and McNemar tests (table 28). *Result.* Clean accuracy is 0.500 under this A/B/C/D prompt, below the official ≈ 0.56. A repeated clean run flips 0 answers. All four C_4 states agree on 66.3% of items. Spatial-relation items flip less under rotation (0.182) than other items (0.214). Audio reversal is driven by speech (speech flip 0.344, accuracy -10.1pp; music 0.094, $+0.9$pp) and is not a meaning-preserving nuisance. The circular audio shift flips 10.9% at a net -0.09pp ($0.500 \to 0.499$): aggregate accuracy hides frame inconsistency.

Table 28: Native Omni-7B on OmniBench ($n = 1142$): answer flips relative to clean.

Condition	Flip rate [95% CI]	Acc	Right→wrong / wrong→right	McNemar p
Clean repeat	0.000	0.500	0 / 0	1
Image rot 90°	0.209 $[0.186, 0.233]$	0.478	101 / 76	0.071
Image rot 180°	0.226 $[0.202, 0.250]$	0.468	115 / 78	0.009
Image rot 270°	0.194 $[0.173, 0.218]$	0.479	90 / 66	0.065
Image h-flip	0.123 $[0.104, 0.143]$	0.496	54 / 49	0.694
Audio circular shift	0.109 $[0.091, 0.127]$	0.499	39 / 38	1
Audio reverse	0.279 $[0.254, 0.305]$	0.427	142 / 59	4×10^{-9}

H Core96K (Stage III): Full Records and Comprehensive Benchmarks

This section is the Core96K protocol for table 3. Accuracies are written as full / V-route / A-route so that $\text{Syn}_{\text{route}}$ can be recomputed. Subsequent subsections localize the gain, then cover visual, spatial, and video benches, the other four trunks, and nested scale. A wrong-frame intervention on the audio–visual benches is a near-tie on those benches.

H.1 PROTOCOL AND EVALUATION MATRIX

Where the sidecar reads the trunk. After Qwen2.5-Omni's Thinker (the multimodal encoder) scatters audio, image, and video features into its input embeddings and before its language model, the token identifiers that bound video, audio, and text spans split the sequence. Each segment is down-projected to width 256 and packed as one modality batch: semantic channels from the projection; coordinates spread along the token layout; timestamps sequential in $[0, 1]$; typed $\ell = 1$ (4×3) and $\ell = 2$ ($2 \times 3 \times 3$) channels; and a frame posterior. Pairing shuffle and no-correspondence change the scores in table 3. The fusion module produces 32 anchors; tokens cross-attend to them and a zero-initialized up-projection adds the result to their own embeddings. The frontends and language model are frozen; the sidecar and LoRA (q, v) are trained. Core96K is nested ($12\text{K} \subset \cdots \subset 96\text{K}$), every scale runs 6000 steps at global batch 64 (384K views) in three stages (sidecar only, sidecar with low-rate LoRA, robustness), and vision-only trunks resample only the rows they can ingest. OmniBench items are held out but the checkpoint is not zero-shot.

Conditions. A V-route / A-route sends one stream into the sidecar while the trunk keeps the full input and question. Pairing shuffle mismatches the sidecar's media instance and the current item. Time shift mislabels one sensor's relative time by 1 s without wrap-around where synchronized clocks exist. Wrong frame rotates the recorded relative pose by $30°$ where it exists; image, label, and trunk input are unchanged. No transport, uniform posterior (uniform frame-hypothesis weights), and no correspondence are full retrains. A separate switch disables transport at inference on the checkpoint trained with it, without retraining. $\mathrm{Syn}_{\mathrm{route}}$ is always computed from the condition's own V/A routes.

Evaluation matrix. Qwen2.5-Omni-7B and -3B run six audio–visual benches (OmniBench 1142 (Li et al., 2025c), WorldSense without subtitles 3172 (Hong et al., 2026), Daily-Omni 1197 (Zhou et al., 2025), XModBench-Lite 6000 (Wang et al., 2026b), AVE per-second 4020 (Tian et al., 2018), AV-Odyssey 4555 (Gong et al., 2024)) and all six XModBench directions ($10{,}220$ each). All five trunks run WorldSense, Video-Holmes (1837) (Cheng et al., 2025), RefCOCO/+/g val (Yu et al., 2016; Mao et al., 2016) ($10{,}834$/$10{,}758$/4896), 3DSRBench (Ma et al., 2025a), VSI-Bench (Yang et al., 2025), All-Angles (2132) (Yeh et al., 2026), ViewSpatial (5712) (Li et al., 2026a), SPBench-SI (1009) (Li et al., 2025b), CrossView ($17{,}000$) (Wang et al., 2026a), NExT-QA val (4996) (Xiao et al., 2021), Perception Test val ($19{,}140$) (Pătrăucean et al., 2023), and Charades-STA R@1 (3720) (Gao et al., 2017); the vision–language trunks also run XModBench T→V and V→T. This gives 98 trunk–task entries and about 644K distinct items. Conditions: frozen trunk, LoRA-only, OGF, and a V-route on every entry; an A-route wherever audio exists; pairing shuffle and wrong frame on every entry; time shift only where a clock pair exists (WorldSense, Daily-Omni, AVE, AV-Odyssey, and Charades-STA on every trunk); the three retrained ablations in table 3. Every trunk is trained at all four nested scales. For the other four trunks, table 36 reports OGF−LoRA and the intervention deltas. Throughout, – marks a condition that does not apply (no synchronized clock pair).

Mixture. The 96K pool is a fixed quota, and each smaller pool is its nested prefix: OmniInstruct-v1 train $30{,}000/15{,}000/7{,}500/3{,}750$ (Li et al., 2025c), LongVALE official train $20{,}000/10{,}000/5{,}000/2{,}500$ (Geng et al., 2025), VGGSound train $18{,}000/9{,}000/4{,}500/2{,}250$ (Chen et al., 2020b), AudioSet train $10{,}000/5{,}000/2{,}500/1{,}250$ (Gemmeke et al., 2017), MUSIC-AVQA-v2 official train $8{,}000/4{,}000/2{,}000/1{,}000$ (Liu et al., 2024), and $10{,}000/5{,}000/2{,}500/1{,}250$ further examples disjoint from those five sources and from the evaluation sets. When a source has fewer examples than its quota, the unused slots stay empty, and evaluation items stay out of training. OmniBench, XModBench, and the LongVALE evaluation split are held out of training.

Optimization. Every scale uses AdamW (Loshchilov & Hutter, 2019) (weight decay 0.01, cosine decay, gradient clip 1.0), global batch 64, and 6000 steps (384K views). Three stages: 3000 steps train only the sidecar (learning rate 8×10^{-5}); 2000 steps add LoRA on query and value projections (4×10^{-5} and 5×10^{-6}); 1000 steps keep both at 2×10^{-5} and 2×10^{-6} while up-weighting shuffle, time shift, conflict, and missing modalities. Checkpoints are chosen on a held-out development split, not on the reported benchmarks. Vision-only trunks resample, from this same pool, only the rows they can answer.

H.2 OMNI-7B AUDIO–VISUAL RECORDS

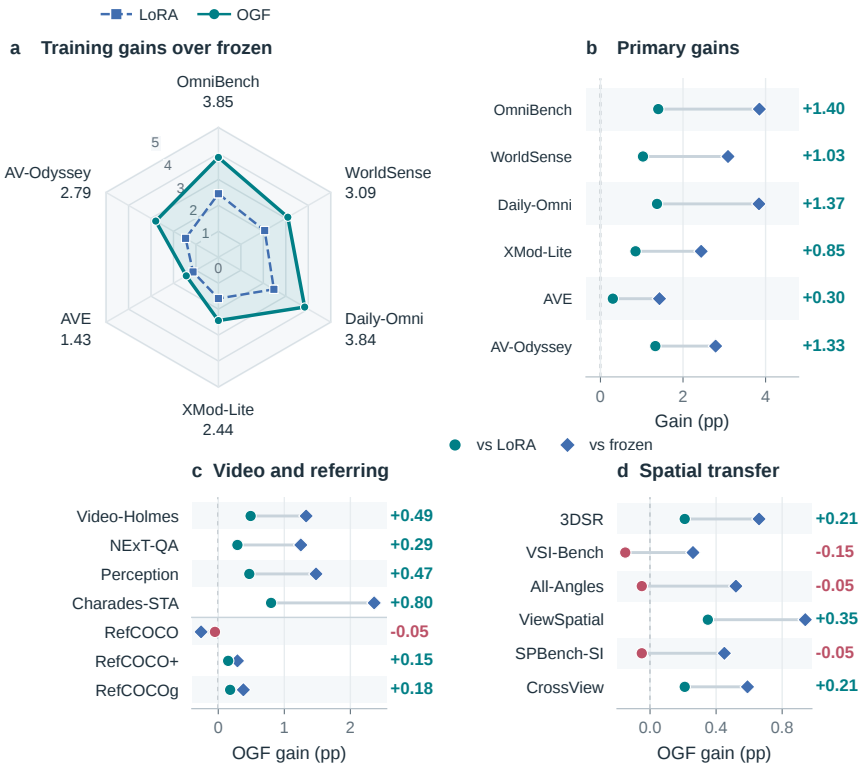


Figure 16: Omni-7B at Core96K: gains versus the frozen trunk and versus LoRA.

Table 29: Omni-7B 96K, two-seed mean accuracy (%) on the six audio–visual benches. Each condition is followed by its own V-route and A-route means. Denominators: OmniBench 1142, WorldSense 3172, Daily-Omni 1197, XModBench-Lite 6000, AVE 4020, AV-Odyssey 4555. $\text{Syn}_{\text{route}}$ in table 3 uses each seed's own three counts. – : no synchronized clock pair.

Condition	OmniB.	WorldS.	Daily	XMod	AVE	AV-Od.
Frozen trunk	56.13	45.40	62.07	60.22	68.48	45.51
LoRA-only	58.58	47.46	64.54	61.81	69.61	46.97
OGF full	59.98	48.49	65.91	62.66	69.91	48.30
V-route	58.06	46.77	63.83	61.14	68.72	46.55
A-route	57.44	45.96	63.24	60.99	68.63	46.17
Pairing shuffle†	56.17	45.03	60.86	60.41	66.72	45.83
V-route	55.78	45.44	61.99	59.72	67.45	46.21
A-route	55.60	44.78	61.32	59.41	67.23	46.06
Time shift†	–	47.23	63.91	–	68.03	47.99
V-route	–	46.67	63.32	–	68.25	46.55
A-route	–	45.89	62.57	–	68.08	46.04
Wrong frame†	59.68	48.27	65.75	62.58	69.81	48.25
V-route	57.84	46.69	63.66	61.08	68.66	46.53
A-route	57.22	45.87	63.07	60.91	68.57	46.14
No transport	59.50	47.73	65.29	62.25	69.65	47.83
V-route	57.88	46.36	63.58	61.08	68.69	46.49
A-route	57.36	45.78	62.99	60.93	68.62	46.09
Uniform posterior	59.72	48.09	65.41	62.39	69.70	48.14
V-route	57.88	46.75	63.58	61.12	68.67	46.56
A-route	57.36	46.04	63.03	60.88	68.54	46.16
No correspondence	58.54	46.89	63.24	61.68	68.99	47.28
V-route	57.75	46.34	62.53	61.14	68.63	46.33
A-route	57.92	46.22	62.32	61.18	68.56	46.43

Three readings follow from these accuracies. First, every intervention and ablation has its own routed synergy, because each condition's V-route and A-route are re-evaluated rather than borrowed from the clean model. Second, under pairing shuffle the V-route and A-route can score above the full model (Daily-Omni $61.99/61.32$ versus 60.86): mismatched media hurt most when both streams are fused. Third, the no-correspondence retrain loses more on its full model than on its best route (WorldSense 1.60 versus 0.43pp, Daily-Omni 2.67 versus 1.30pp), so what the ablation removes is mainly the fused part. Fourth, the no-transport retrain lowers these six accuracies by 0.26 to 0.76pp and keeps its own routed synergy positive, from $+0.95$ to $+1.75$pp. It stays above LoRA on all six; the margin is 0.92, 0.27, 0.75, 0.44, 0.04, and 0.86pp, so AVE is only just above.

H.3 Where the gain lives

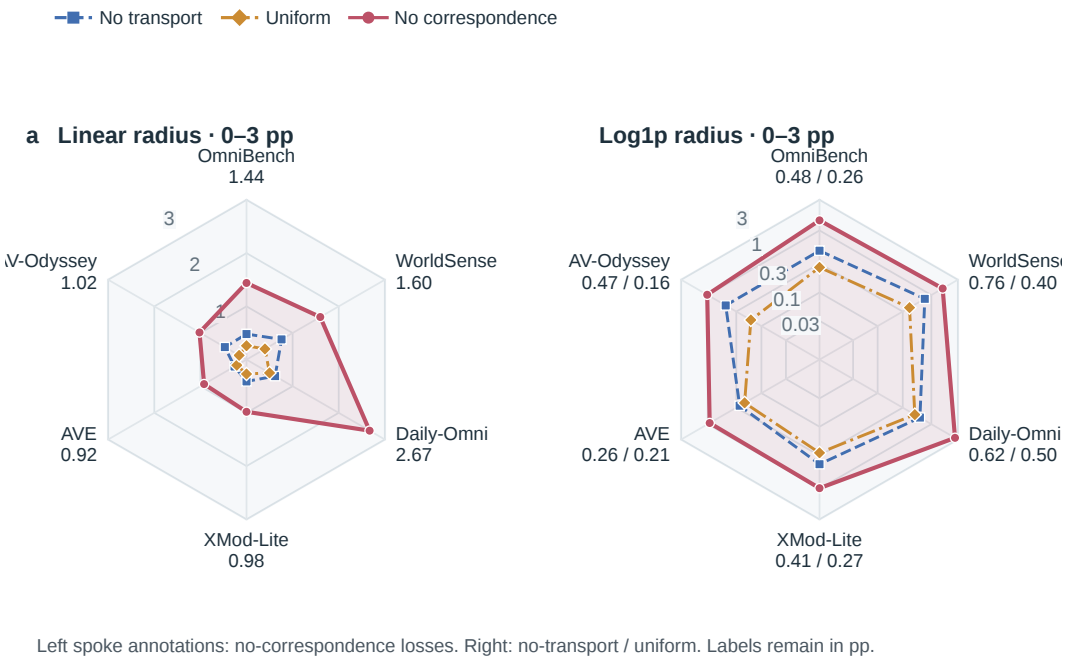
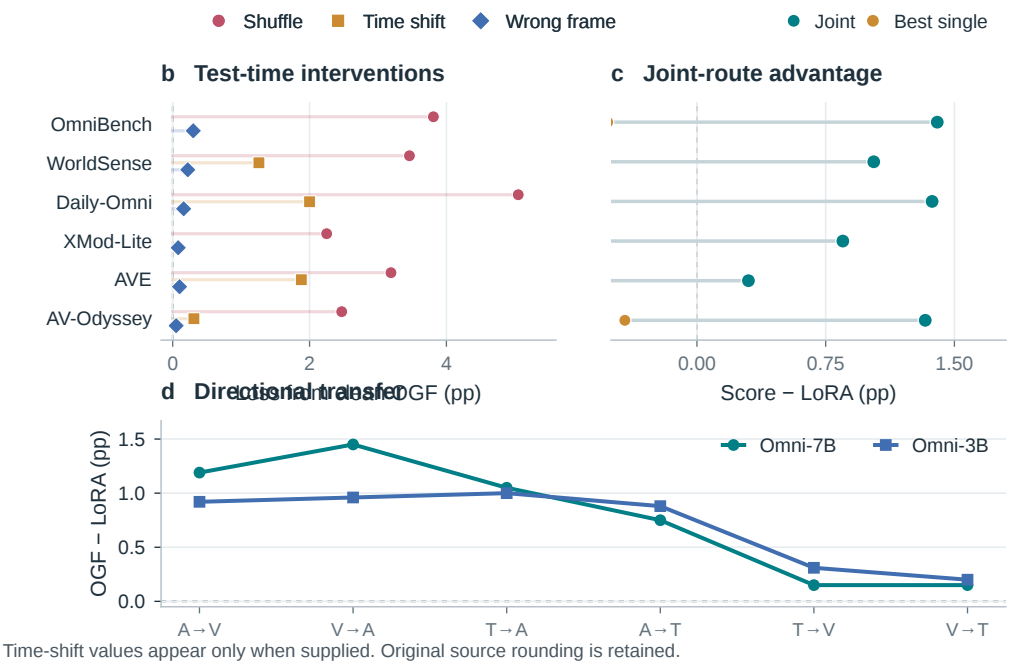


Figure 17: Omni-7B retrained ablations, plotted on a linear radial axis and again with radius $\log(1+x)$.



Time-shift values appear only when supplied. Original source rounding is retained.

Figure 18: Test-time interventions, joint-route advantage, and XMod direction gains.

Sub-items. Relative to LoRA, the two-seed means gain on OmniBench speech $+0.97$pp, sound $+2.64$pp, and music $+1.42$pp. On Daily-Omni the gain is $+3.15$pp on audio–visual items and $+3.10$pp on event items; comparison is 0.00pp, context $+0.52$pp, and inference and reasoning are -0.33 and -0.57pp (table 31). On WorldSense the largest gains are performance $+2.25$pp, music $+1.97$pp, and daily life $+1.14$pp. Pairing shuffle removes the OmniBench and WorldSense category gains (table 30).

a Fine-grained category effects

	Frozen	V route	A route	Joint	Shuffle	No corr.	n
OB · speech	-2.21	-0.39	-0.97	+0.97	-1.17	-0.13	771
OB · sound	-2.83	-0.94	-1.13	+2.64	-5.28	-0.19	265
OB · music	-3.30	-0.47	-2.36	+1.42	-4.24	+0.95	106
WS · tech	-1.63	-0.10	-2.15	+0.81	-1.74	-0.10	490
WS · culture	-1.13	-0.65	-2.43	+0.65	-1.62	-0.32	309
WS · daily life	-2.58	-0.46	-1.82	+1.14	-2.89	-0.46	658
WS · film	-1.71	-0.92	-1.84	+0.66	-1.97	-0.52	379
WS · performance	-2.44	-0.75	0.00	+2.25	-4.12	-1.31	267
WS · games	-1.50	-0.22	-1.72	0.00	-0.22	-0.22	233
WS · sports	-1.74	0.00	-2.21	+0.58	-2.21	-0.70	430
WS · music	-3.20	-2.58	+0.74	+1.97	-3.94	-1.11	406
Daily · Audio–visual	-3.57	–	–	+3.15	–	–	238
Daily · Event	-2.78	–	–	+3.10	–	–	306
Daily · Comparative	-3.43	–	–	0.00	–	–	131
Daily · Context	-2.85	–	–	+0.52	–	–	193
Daily · Inference	-0.98	–	–	-0.33	–	–	154
Daily · Reasoning	-0.57	–	–	-0.57	–	–	175

−5.0 −2.5 0.0 2.5

Score − LoRA (pp)

Every cell uses its row-specific LoRA reference. Daily-Omni route cells remain missing.

Figure 19: Category-level contrasts versus matched LoRA.

b Omni-3B replication

	Frozen	V	A	Joint	S	T	F
OmniBench	-2.89	-0.44	+0.35	+1.40	-2.63	–	+1.22
WorldSense	-1.83	+0.12	+0.09	+0.97	-2.59	-0.32	+0.66
Daily-Omni	-2.26	-0.17	-0.33	+1.25	-3.68	-0.67	+1.09
XMod-Lite	-1.79	-0.29	+0.05	+0.78	-2.05	–	+0.70
AVE	-0.90	-0.72	-0.05	+0.37	-2.66	-1.40	+0.27
AV-Odyssey	-1.54	-1.08	+0.26	+0.72	-1.63	+0.46	+0.66

S/T/F: test-time shuffle / time shift / wrong frame. Missing comparisons remain blank.

Figure 20: Omni-3B route and intervention contrasts versus matched LoRA.

Table 30: Omni-7B 96K sub-items, two-seed mean accuracy (%). Categories are the benchmarks' own; $S^{\dagger}$ is pairing shuffle, −C is the no-correspondence retrain.

Category	n	Frozen	LoRA	V-route	A-route	OGF	$S^{\dagger}$	−C
OmniBench speech	771	55.25	57.46	57.07	56.49	58.43	56.29	57.33
OmniBench sound	265	60.00	62.83	61.89	61.70	65.47	57.55	62.64
OmniBench music	106	52.83	56.13	55.66	53.77	57.55	51.89	57.08
WorldSense tech	490	47.76	49.39	49.29	47.24	50.20	47.65	49.29
WorldSense culture	309	49.84	50.97	50.32	48.54	51.62	49.35	50.65
WorldSense daily life	658	43.62	46.20	45.74	44.38	47.34	43.31	45.74
WorldSense film	379	43.80	45.51	44.59	43.67	46.17	43.54	44.99
WorldSense performance	267	48.31	50.75	50.00	50.75	53.00	46.63	49.44
WorldSense games	233	39.06	40.56	40.34	38.84	40.56	40.34	40.34
WorldSense sports	430	43.49	45.23	45.23	43.02	45.81	43.02	44.53
WorldSense music	406	47.29	50.49	47.91	51.23	52.46	46.55	49.38

Table 31: Daily-Omni sub-items (Omni-7B, 96K), two-seed means: accuracy (%) and the OGF−LoRA gap; cited in section 6.3.

Sub-item	n	Frozen	LoRA	OGF	OGF−LoRA (pp)
Audio–visual	238	48.32	51.89	55.04	+3.15
Event	306	58.17	60.95	64.05	+3.10
Comparative	131	69.47	72.90	72.90	0.00
Context	193	58.55	61.40	61.92	+0.52
Inference	154	76.62	77.60	77.27	−0.33
Reasoning	175	73.14	73.71	73.14	−0.57

Table 32: XModBench full set by direction (condition → answer modality; $n = 10{,}220$ each): accuracy (%) and OGF − LoRA (pp). Omni-7B is a two-seed mean. The vision–language trunks can run only T→V and V→T.

Omni trunks Trunk / arm	A→V	V→A	T→A	A→T	T→V	V→T
Omni-7B frozen	48.00	50.50	55.40	62.01	59.61	76.30
Omni-7B LoRA	50.55	52.94	56.79	63.54	60.91	77.04
Omni-7B OGF	51.74	54.39	57.84	64.29	61.06	77.19
Omni-7B OGF−LoRA (pp)	+1.19	+1.45	+1.05	+0.75	+0.15	+0.15
Omni-3B frozen	45.87	48.16	50.33	57.15	57.69	73.52
Omni-3B LoRA	48.18	50.33	52.25	58.81	58.98	74.76
Omni-3B OGF	49.10	51.29	53.25	59.69	59.30	74.95
Omni-3B OGF−LoRA (pp)	+0.92	+0.96	+1.00	+0.88	+0.31	+0.20

Vision–language trunks (T→V / V→T) Trunk / arm	InternVL2-8B		LLaVA-OV-7B		Qwen3-VL-4B	
	T→V	V→T	T→V	V→T	T→V	V→T
Frozen	46.89	72.04	44.21	68.92	61.36	75.48
LoRA	48.41	73.00	45.64	70.14	62.66	76.42
OGF	48.92	73.26	46.26	70.47	62.86	76.58
OGF−LoRA (pp)	+0.51	+0.25	+0.63	+0.33	+0.20	+0.16

The gain is largest for audio–visual correspondence (A→V, V→A) and smaller for text–audio and text–visual matching. The six-way means are 61.08 (OGF), 60.29 (LoRA), and 58.64 (frozen) on Omni-7B, and 57.93, 57.22, and 55.45 on Omni-3B. The Omni-7B row is a two-seed mean; Omni-3B remains one seed.

H.4 VISUAL, SPATIAL, AND VIDEO BENCHES (OMNI-7B)

a Complete conditional-route scores

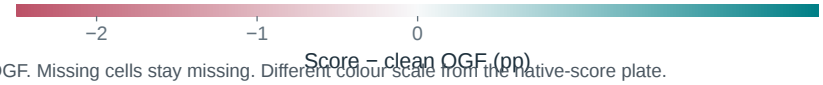
	Omni	World	Daily	XMod	AVE	Odyssey
Clean · joint	59.98	48.49	65.91	62.66	69.91	48.30
Clean · V route	58.06	46.77	63.83	61.14	68.72	46.55
Clean · A route	57.44	45.96	63.24	60.99	68.63	46.17
Shuffle · joint	56.17	45.03	60.86	60.41	66.72	45.83
Shuffle · V route	55.78	45.44	61.99	59.72	67.45	46.21
Shuffle · A route	55.60	44.78	61.32	59.41	67.23	46.06
Time shift · joint	–	47.23	63.91	–	68.03	47.99
Time shift · V route	–	46.67	63.32	–	68.25	46.55
Time shift · A route	–	45.89	62.57	–	68.08	46.04
Wrong frame · joint	59.68	48.27	65.75	62.58	69.81	48.25
Wrong frame · V route	57.84	46.69	63.66	61.08	68.66	46.53
Wrong frame · A route	57.22	45.87	63.07	60.91	68.57	46.14
No transport · joint	59.50	47.73	65.29	62.25	69.65	47.83
No transport · V route	57.88	46.36	63.58	61.08	68.69	46.49
No transport · A route	57.36	45.78	62.99	60.93	68.62	46.09
Uniform · joint	59.72	48.09	65.41	62.39	69.70	48.14
Uniform · V route	57.88	46.75	63.58	61.12	68.67	46.56
Uniform · A route	57.36	46.04	63.03	60.88	68.54	46.16
No corr. · joint	58.54	46.89	63.24	61.68	68.99	47.28
No corr. · V route	57.75	46.34	62.53	61.14	68.63	46.33
No corr. · A route	57.92	46.22	62.32	61.18	68.56	46.43

Native scores. Each intervention has joint, visual-sidecar and audio-sidecar rows.

Figure 21: Conditional-route scores. Inapplicable conditions are left blank.

b Visual and spatial interventions

	Shuffle	Time	Frame	No T	Unif.	No C
Video-Holmes	-1.41	–	-0.05	-0.43	-0.16	-0.68
RefCOCO	-0.80	–	-0.77	-0.19	+0.05	-0.01
RefCOCO+	-1.09	–	-0.41	-0.24	-0.04	-0.19
RefCOCOg	-1.24	–	-0.68	-0.29	-0.06	-0.25
3DSR	-0.85	–	-0.68	-1.49	+0.03	-0.21
VSI-Bench	-1.04	–	-0.68	-1.08	+0.14	-0.22
All-Angles	-1.13	–	-0.96	-1.81	+0.07	-0.09
ViewSpatial	-0.95	–	-1.27	-1.87	-0.13	-0.41
SPBench-SI	-0.99	–	-0.69	-1.98	0.00	-0.25
CrossView	-0.65	–	-0.77	-1.52	-0.07	-0.25
NExT-QA	-1.59	–	-0.11	-0.48	-0.03	-0.50
Perception	-1.45	–	-0.16	-0.51	-0.10	-0.43
Charades-STA	-2.40	-1.85	-0.02	-0.51	-0.17	-0.53



Changes from clean OGF. Missing cells stay missing. Different colour scale from the native-score plate.

Figure 22: Visual and spatial interventions relative to clean OGF.

Table 33: Omni-7B 96K visual, spatial, and video benches, two-seed mean accuracy (%). 3DSR and VSI are the benchmarks' own percentages. Where the table has no separate audio route, the V-route equals full OGF and $\mathrm{Syn}_{\mathrm{route}}$ is undefined; these rows test non-regression and whether the sidecar reads visual evidence. $S^{\dagger}$ pairing shuffle, $T^{\dagger}$ time shift, $F^{\dagger}$ wrong frame, $-$T no-transport retrain, Unif. uniform posterior, $-$C no correspondence. – : not applicable.

Bench	n	Frozen	LoRA	OGF	$S^{\dagger}$	$T^{\dagger}$	$F^{\dagger}$	$-$T	Unif.	$-$C
Video-Holmes	1837	16.44	17.28	17.77	16.36	–	17.72	17.34	17.61	17.09
RefCOCO val	10834	90.50	90.29	90.24	89.44	–	89.47	90.05	90.29	90.23
RefCOCO+ val	10758	85.40	85.54	85.69	84.60	–	85.28	85.45	85.65	85.50
RefCOCOg val	4896	87.40	87.60	87.78	86.54	–	87.10	87.49	87.72	87.53
3DSRBench	–	45.50	45.95	46.16	45.31	–	45.48	44.67	46.19	45.95
VSI-Bench	–	29.40	29.81	29.66	28.62	–	28.98	28.58	29.80	29.44
All-Angles	2132	40.85	41.42	41.37	40.24	–	40.41	39.56	41.44	41.28
ViewSpatial	5712	33.19	33.78	34.13	33.18	–	32.86	32.26	34.00	33.72
SPBench-SI	1009	39.54	40.04	39.99	39.00	–	39.30	38.01	39.99	39.74
CrossView	17000	36.16	36.54	36.75	36.10	–	35.98	35.23	36.68	36.50
NExT-QA val	4996	65.73	66.69	66.98	65.39	–	66.87	66.50	66.95	66.48
Perception Test val	19140	59.71	60.72	61.19	59.74	–	61.03	60.68	61.09	60.76
Charades-STA R@1	3720	23.04	24.60	25.40	23.00	23.55	25.38	24.89	25.23	24.87

Pairing shuffle still lowers every row (RefCOCO val 89.44 versus 90.24): the sidecar reads the current item's visual evidence. The wrong-frame intervention costs most on ViewSpatial (-1.27pp) and about 0.77pp on CrossView and RefCOCO val. OGF is 0.26pp below the frozen trunk on RefCOCO val and 0.15pp below LoRA on VSI-Bench. The no-transport retrain stays above LoRA on Video-Holmes ($+0.06$pp) and Charades-STA ($+0.29$pp). It falls below LoRA by 1.23 to 2.03pp on 3DSR, VSI, All-Angles, ViewSpatial, SPBench-SI, and CrossView, and by only 0.04 to 0.24pp on the RefCOCO splits, NExT-QA, and Perception. table 34 is a different switch: transport is turned off at inference on the checkpoint trained with it, and the model is not retrained. Relative to the retrain, that switch costs a further 0.18 to 0.75pp on the audio–visual, referring, and video benches, and 0.85 to 1.31pp on the six spatial benches. RefCOCO+ and RefCOCOg stay above the frozen trunk after the retrain and fall below it under the switch; RefCOCO val and the six spatial benches are already below the frozen trunk after the retrain.

Table 34: Transport disabled only at inference on the Omni-7B checkpoint trained with it. Two-seed mean accuracy (%). This is not the no-transport retrain in tables 29 and 33.

Bench	Accuracy	Bench	Accuracy
OmniBench	59.06	RefCOCOg val	87.10
WorldSense	47.12	3DSR Overall	43.47
Daily-Omni	64.54	VSI Avg	27.73
XMod-Lite	61.95	All-Angles	38.25
AVE	69.42	ViewSpatial	31.03
AV-Odyssey	47.48	SPBench-SI	36.72
XMod six-direction pool	60.39	CrossView	33.99
Video-Holmes	17.09	NExT-QA val	66.07
RefCOCO val	89.72	Perception val	60.18
RefCOCO+ val	85.27	Charades-STA	24.52

H.5 THE OTHER FOUR TRUNKS



O3: Omni-3B; IV: InternVL2-8B; LV: LLaVA-OV-7B; QV: Qwen3-VL-4B.
Clean gain uses LoRA; shuffle/frame changes use clean OGF.

Figure 23: The other four trunks on shared benches.

d All available XMod directions

	Frozen	LoRA	OGF	G − L
Omni-7B · A→V	48.00	50.55	51.74	+1.19
Omni-7B · V→A	50.50	52.94	54.39	+1.45
Omni-7B · T→A	55.40	56.79	57.84	+1.05
Omni-7B · A→T	62.01	63.54	64.29	+0.75
Omni-7B · T→V	59.61	60.91	61.06	+0.15
Omni-7B · V→T	76.30	77.04	77.19	+0.15
Omni-3B · A→V	45.87	48.18	49.10	+0.92
Omni-3B · V→A	48.16	50.33	51.29	+0.96
Omni-3B · T→A	50.33	52.25	53.25	+1.00
Omni-3B · A→T	57.15	58.81	59.69	+0.88
Omni-3B · T→V	57.69	58.98	59.30	+0.31
Omni-3B · V→T	73.52	74.76	74.95	+0.20
InternVL2-8B · T→V	46.89	48.41	48.92	+0.51
InternVL2-8B · V→T	72.04	73.00	73.26	+0.25
LLaVA-OV-7B · T→V	44.21	45.64	46.26	+0.63
LLaVA-OV-7B · V→T	68.92	70.14	70.47	+0.33
Qwen3-VL-4B · T→V	61.36	62.66	62.86	+0.20
Qwen3-VL-4B · V→T	75.48	76.42	76.58	+0.16

XMod gains are source-reported. Acoustic directions are absent for visual-only trunks.

Figure 24: Available XMod directions with OGF−LoRA gains.

Table 35: Omni-3B 96K audio–visual benches: accuracy (%) and $\mathrm{Syn}_{\mathrm{route}}$ (pp). Interventions are on the same checkpoint. $\mathrm{S}^{\dagger}$ pairing shuffle, $\mathrm{T}^{\dagger}$ time shift, $\mathrm{F}^{\dagger}$ wrong frame.

Bench	n	Frozen	LoRA	V-route	A-route	OGF	$\mathrm{S}^{\dagger}$	$\mathrm{T}^{\dagger}$	$\mathrm{F}^{\dagger}$	$\mathrm{Syn}_{\mathrm{route}}$
OmniBench	1142	52.19	55.08	54.64	55.43	56.48	52.45	–	56.30	+1.05
WorldSense	3172	41.30	43.13	43.25	43.22	44.10	40.54	42.81	43.79	+0.85
Daily-Omni	1197	60.23	62.49	62.32	62.16	63.74	58.81	61.82	63.58	+1.42
XMod-Lite	6000	57.13	58.92	58.63	58.97	59.70	56.87	–	59.62	+0.73
AVE	4020	66.59	67.49	66.77	67.44	67.86	64.83	66.09	67.76	+0.42
AV-Odyssey	4555	47.99	49.53	48.45	49.79	50.25	47.90	49.99	50.19	+0.46

Table 36: The other four trunks at 96K, one seed, in pp: OGF − LoRA / pairing shuffle† − OGF / wrong frame† − OGF. Negative OGF − LoRA entries are in bold. Referring shuffle is reported for Qwen3-VL; for InternVL2 and LLaVA-OV it is not in this table. The Omni-7B referring shuffle is in table 33. These trunks have no retrained ablations; the three vision–language trunks have no audio.

Bench	Omni-3B	InternVL2	LLaVA-OV	Qwen3-VL
WorldSense	+0.98 / −3.56 / −0.32	+0.35 / −2.30 / −0.19	+0.41 / −2.49 / −0.19	+0.32 / −1.73 / −0.22
Video-Holmes	+0.27 / −1.14 / −0.16	+0.27 / −1.74 / −0.16	+0.38 / −1.63 / −0.22	+0.27 / −1.36 / −0.11
RefCOCO	+0.08 / – / −0.89	+0.06 / – / −1.00	+0.15 / – / −0.67	+0.04 / −0.72 / −0.17
RefCOCO+	+0.15 / – / −0.35	+0.20 / – / −0.38	+0.20 / – / −0.41	+0.06 / −0.77 / −0.17
RefCOCOg	+0.18 / – / −0.67	+0.16 / – / −0.78	+0.27 / – / −0.63	+0.10 / −1.16 / −0.49
3DSR	+0.15 / −0.80 / −0.70	**−0.05** / −0.75 / −0.60	+0.08 / −0.80 / −0.65	+0.07 / −0.60 / −0.48
VSI	**−0.10** / −1.10 / −0.80	**−0.05** / −1.40 / −0.70	+0.05 / −1.25 / −0.80	+0.10 / −1.17 / −0.69
All-Angles	+0.23 / −1.13 / −1.03	**−0.05** / −1.08 / −0.89	+0.38 / −1.45 / −1.03	+0.28 / −1.22 / −1.03
ViewSpatial	+0.40 / −1.02 / −1.21	+0.11 / −1.03 / −1.16	+0.32 / −1.17 / −1.38	+0.35 / −1.02 / −1.24
SPBench-SI	+0.20 / −1.09 / −0.79	**−0.10** / −0.99 / −0.59	+0.30 / −1.19 / −0.79	+0.40 / −1.19 / −0.79
CrossView	+0.11 / −0.65 / −0.85	+0.07 / −0.69 / −0.84	+0.22 / −0.93 / −0.98	+0.24 / −0.61 / −0.88
NExT-QA	+0.56 / −1.68 / −0.12	+0.48 / −1.84 / −0.14	+0.16 / −1.22 / −0.08	+0.16 / −1.14 / −0.16
Perception	+0.20 / −1.72 / −0.15	+0.32 / −1.88 / −0.13	+0.37 / −1.57 / −0.14	+0.14 / −1.17 / −0.10
Charades-STA	+0.46 / −1.85 / −0.05	+0.54 / −2.23 / −0.08	+0.54 / −2.26 / −0.05	+0.48 / −2.02 / −0.05

Omni-3B repeats the Omni-7B pattern with interventions: routed synergy is positive on all six audio–visual benches, shuffle lowers every one, and time shift lowers the four benches that carry a clock pair. Pairing shuffle lowers WorldSense, Video-Holmes, and Charades-STA on every vision–language trunk as well. Portability of the gain is not uniform. InternVL2 trails LoRA on 3DSR, VSI, All-Angles, and SPBench-SI. Qwen3-VL-4B leads LoRA on all sixteen benches at 96K (mean +0.21pp), including every RefCOCO split and every spatial bench. The three vision–language trunks have no audio, so this comparison does not measure audio–visual synergy.

H.6 DATA SCALE

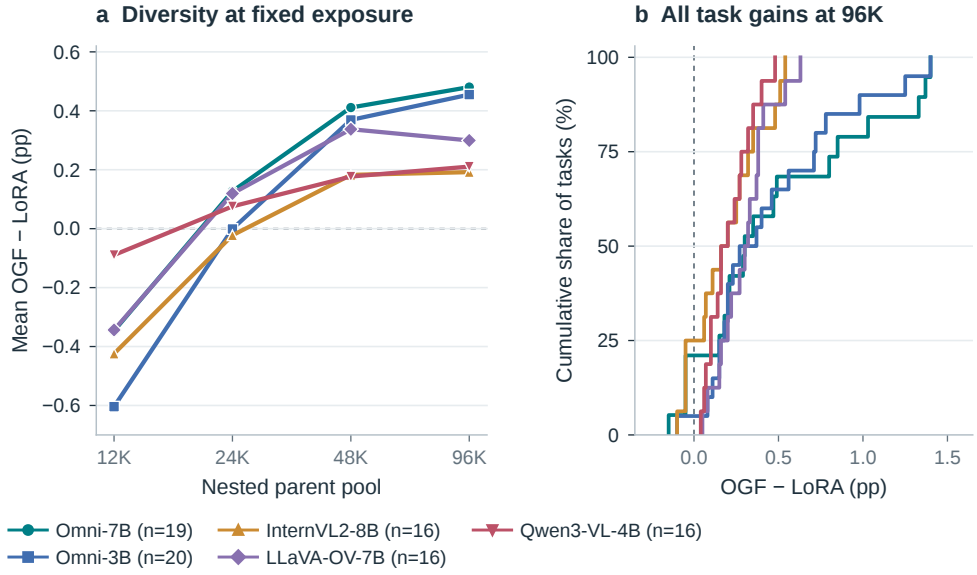


Figure 25: Nested Core96K mean gains and 96K gain CDFs at fixed exposure.

Table 37: Nested data scale on all five trunks (every scale: 6000 steps, 384K views). For each trunk: benches with OGF $>$ LoRA / benches measured, and the mean OGF $-$ LoRA (pp). Omni-7B at 96K is the mean of two seeds; every other entry is one seed. Per-bench values are in table 38.

Trunk	12K	24K	48K	96K
Omni-7B	1/19 (-0.35)	12/19 ($+0.13$)	18/19 ($+0.41$)	15/19 ($+0.48$)
Omni-3B	0/20 (-0.60)	8/20 (0.00)	19/20 ($+0.37$)	19/20 ($+0.46$)
InternVL2-8B	0/16 (-0.42)	6/16 (-0.02)	13/16 ($+0.18$)	12/16 ($+0.19$)
LLaVA-OV-7B	0/16 (-0.34)	13/16 ($+0.12$)	16/16 ($+0.34$)	16/16 ($+0.30$)
Qwen3-VL-4B	0/16 (-0.09)	13/16 ($+0.08$)	16/16 ($+0.18$)	16/16 ($+0.21$)

The correspondence path needs data. At 12K, OGF is below LoRA on 85 of the 87 trunk–benchmark pairs; OmniBench on Omni-7B is higher, and Qwen3-VL SPBench-SI ties. At equal steps and views, the mean gap turns positive by 24K on Omni-7B, LLaVA-OV, and Qwen3-VL, and by 48K on Omni-3B and InternVL2. On Qwen3-VL the mean is -0.09 / $+0.08$ / $+0.18$ / $+0.21$pp at $12/24/48/96$K. Omni-7B's LoRA plateaus after 48K on the audio–visual benches while OGF keeps gaining.

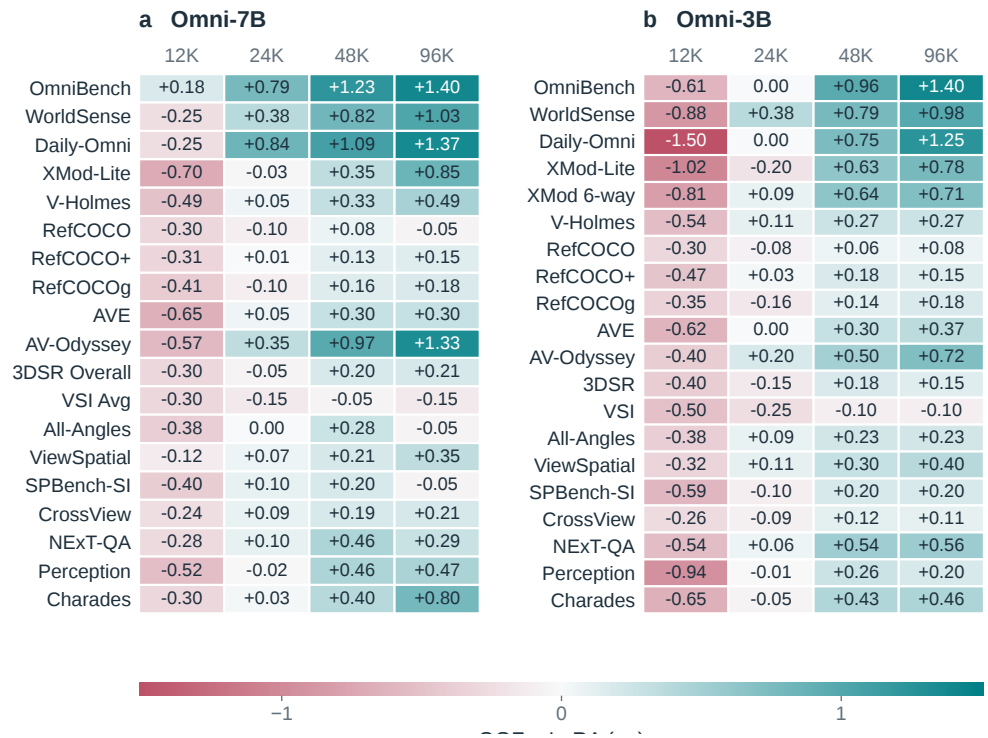


Figure 26: Omni-7B and Omni-3B scale cells, one zero-centered scale.

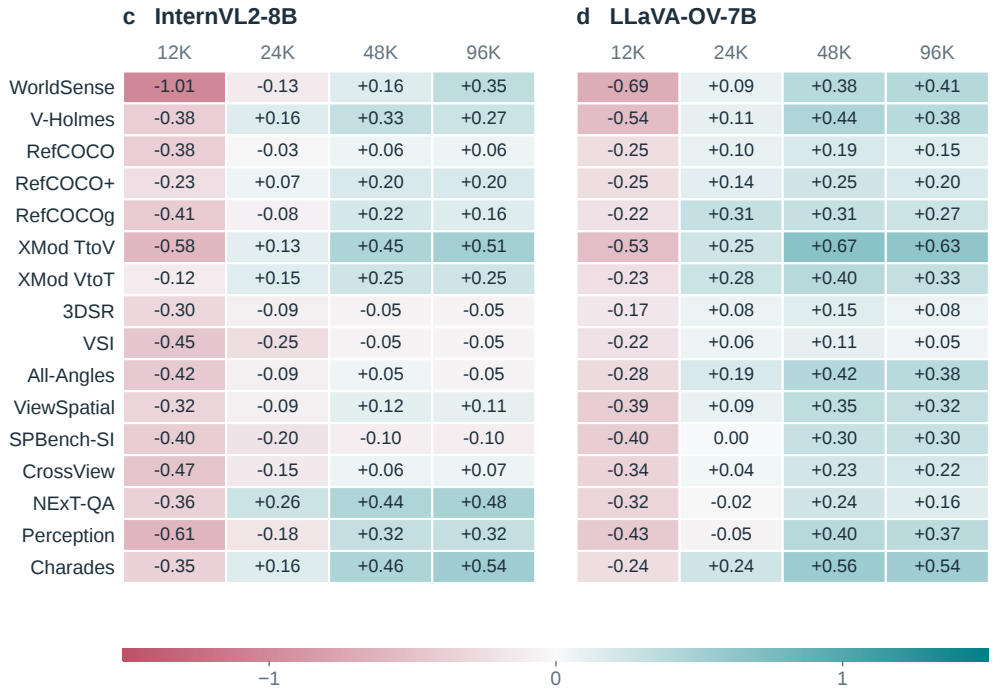


Figure 27: InternVL2 and LLaVA-OV scale cells, same scale.

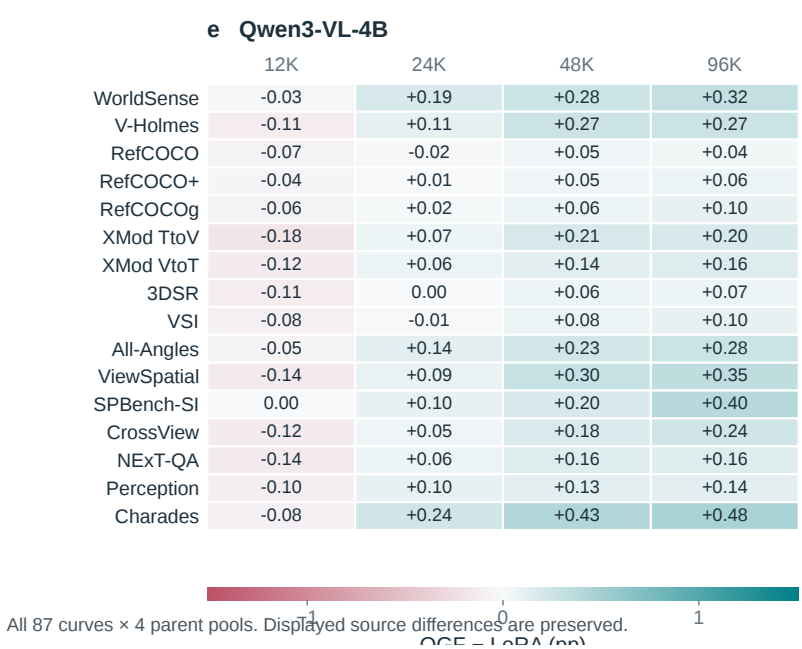


Figure 28: Qwen3-VL-4B scale cells, on the same -1 to $+1$ color scale as the other trunks. The 348 trunk–bench gaps are this figure together with figs. 26 and 27 and table 38.

Table 38: Nested data scale, every trunk and bench: OGF − LoRA (pp). Omni-7B at 96K is the mean of two seeds; every other entry is one seed.

Trunk	Bench	12K	24K	48K	96K
Omni-7B	OmniBench	+0.18	+0.79	+1.23	+1.40
	WorldSense	−0.25	+0.38	+0.82	+1.03
	Daily-Omni	−0.25	+0.84	+1.09	+1.37
	XMod-Lite	−0.70	−0.03	+0.35	+0.85
	Video-Holmes	−0.49	+0.05	+0.33	+0.49
	RefCOCO	−0.30	−0.10	+0.08	−0.05
	RefCOCO+	−0.31	+0.01	+0.13	+0.15
	RefCOCOg	−0.41	−0.10	+0.16	+0.18
	AVE	−0.65	+0.05	+0.30	+0.30
	AV-Odyssey	−0.57	+0.35	+0.97	+1.33
	3DSR Overall	−0.30	−0.05	+0.20	+0.21
	VSI Avg	−0.30	−0.15	−0.05	−0.15
	All-Angles	−0.38	0.00	+0.28	−0.05
	ViewSpatial	−0.12	+0.07	+0.21	+0.35
	SPBench-SI	−0.40	+0.10	+0.20	−0.05
	CrossView	−0.24	+0.09	+0.19	+0.21
	NExT-QA	−0.28	+0.10	+0.46	+0.29
	Perception	−0.52	−0.02	+0.46	+0.47
	Charades-STA	−0.30	+0.03	+0.40	+0.80
Omni-3B	OmniBench	−0.61	0.00	+0.96	+1.40
	WorldSense	−0.88	+0.38	+0.79	+0.98
	Daily-Omni	−1.50	0.00	+0.75	+1.25
	XMod-Lite	−1.02	−0.20	+0.63	+0.78
	XMod six-way mean	−0.81	+0.09	+0.64	+0.71
	Video-Holmes	−0.54	+0.11	+0.27	+0.27
	RefCOCO	−0.30	−0.08	+0.06	+0.08
	RefCOCO+	−0.47	+0.03	+0.18	+0.15
	RefCOCOg	−0.35	−0.16	+0.14	+0.18
	AVE	−0.62	0.00	+0.30	+0.37
	AV-Odyssey	−0.40	+0.20	+0.50	+0.72
	3DSR	−0.40	−0.15	+0.18	+0.15
	VSI	−0.50	−0.25	−0.10	−0.10
	All-Angles	−0.38	+0.09	+0.23	+0.23
	ViewSpatial	−0.32	+0.11	+0.30	+0.40
	SPBench-SI	−0.59	−0.10	+0.20	+0.20
	CrossView	−0.26	−0.09	+0.12	+0.11
	NExT-QA	−0.54	+0.06	+0.54	+0.56
	Perception	−0.94	−0.01	+0.26	+0.20
	Charades-STA	−0.65	−0.05	+0.43	+0.46
InternVL2-8B	WorldSense	−1.01	−0.13	+0.16	+0.35
	Video-Holmes	−0.38	+0.16	+0.33	+0.27
	RefCOCO	−0.38	−0.03	+0.06	+0.06
	RefCOCO+	−0.23	+0.07	+0.20	+0.20
	RefCOCOg	−0.41	−0.08	+0.22	+0.16
	XMod T→V	−0.58	+0.13	+0.45	+0.51
	XMod V→T	−0.12	+0.15	+0.25	+0.25
	3DSR	−0.30	−0.09	−0.05	−0.05

Table 38: Nested data scale, every trunk and bench (continued).

Trunk	Bench	12K	24K	48K	96K
	VSI	−0.45	−0.25	−0.05	−0.05
	All-Angles	−0.42	−0.09	+0.05	−0.05
	ViewSpatial	−0.32	−0.09	+0.12	+0.11
	SPBench-SI	−0.40	−0.20	−0.10	−0.10
	CrossView	−0.47	−0.15	+0.06	+0.07
	NExT-QA	−0.36	+0.26	+0.44	+0.48
	Perception	−0.61	−0.18	+0.32	+0.32
	Charades-STA	−0.35	+0.16	+0.46	+0.54
LLaVA-OV-7B	WorldSense	−0.69	+0.09	+0.38	+0.41
	Video-Holmes	−0.54	+0.11	+0.44	+0.38
	RefCOCO	−0.25	+0.10	+0.19	+0.15
	RefCOCO+	−0.25	+0.14	+0.25	+0.20
	RefCOCOg	−0.22	+0.31	+0.31	+0.27
	XMod T→V	−0.53	+0.25	+0.67	+0.63
	XMod V→T	−0.23	+0.28	+0.40	+0.33
	3DSR	−0.17	+0.08	+0.15	+0.08
	VSI	−0.22	+0.06	+0.11	+0.05
	All-Angles	−0.28	+0.19	+0.42	+0.38
	ViewSpatial	−0.39	+0.09	+0.35	+0.32
	SPBench-SI	−0.40	0.00	+0.30	+0.30
	CrossView	−0.34	+0.04	+0.23	+0.22
	NExT-QA	−0.32	−0.02	+0.24	+0.16
	Perception	−0.43	−0.05	+0.40	+0.37
	Charades-STA	−0.24	+0.24	+0.56	+0.54
Qwen3-VL-4B	WorldSense	−0.03	+0.19	+0.28	+0.32
	Video-Holmes	−0.11	+0.11	+0.27	+0.27
	RefCOCO	−0.07	−0.02	+0.05	+0.04
	RefCOCO+	−0.04	+0.01	+0.05	+0.06
	RefCOCOg	−0.06	+0.02	+0.06	+0.10
	XMod T→V	−0.18	+0.07	+0.21	+0.20
	XMod V→T	−0.12	+0.06	+0.14	+0.16
	3DSR	−0.11	0.00	+0.06	+0.07
	VSI	−0.08	−0.01	+0.08	+0.10
	All-Angles	−0.05	+0.14	+0.23	+0.28
	ViewSpatial	−0.14	+0.09	+0.30	+0.35
	SPBench-SI	0.00	+0.10	+0.20	+0.40
	CrossView	−0.12	+0.05	+0.18	+0.24
	NExT-QA	−0.14	+0.06	+0.16	+0.16
	Perception	−0.10	+0.10	+0.13	+0.14
	Charades-STA	−0.08	+0.24	+0.43	+0.48

H.7 Efficiency and seeds

Trainable parameters on Omni-7B are $9{,}204{,}869$ for OGF (sidecar $1{,}537{,}157$, LoRA $r = 16$) and $9{,}584{,}640$ for LoRA-only ($r = 20$, 4.13% more). The no-transport retrain has the same count as OGF. Memory counted from parameters and activations is 16.745 versus 16.744 GiB per GPU. InternVL2 applies LoRA to a fused query-key-value projection (`wqkv`) rather than separate query and value projections, so its trainable count is not comparable to the *q*/*v* counts above. Omni-7B at 96K trains OGF, LoRA-only, and the three ablations twice. The two seeds differ by a few items, and the tables report their mean. No confidence interval or significance test is attached to that two-seed mean.

I Limitations, Negative Results, and Boundaries

Negative results fall into four kinds.

Expected ties. On determinate frames OGF matches hard canonicalization (0.915 vs. 0.927, $p = 1.0$). On a four-element C_4 orbit, augmentation can tile the group and the interval against a C_4-augmented Q-Former includes zero. When AVE streams are already aligned, naive fusion is the right model (0.656 vs. 0.568).

Expected losses of accuracy that still leave a different win. Under frame ambiguity, OGF is less accurate than MAP (0.254 vs. 0.284) and is the calibrated model. The group MLP beats full OGF on the inner-product label (0.903 vs. 0.773) after transport has already done the work.

Mechanisms that were predicted and did not appear. Frame-poisoning entropy stays at 0.34 instead of $\ln 4$. The AVE posterior's top offset is correct in at most 15% of clips and never at $+5$ s. CREMA SNR at -10 dB has an interval that includes 0.5. Soft C_4 averaging on frozen CLIP tokens loses to four-view TTA.

Comparisons that missed a pre-specified threshold. AVE family recall@1 $0.829 < 0.90$. SigLIP2 uncompiled runtime 1.22–$1.70\times$. LongVALE F1/IoU below the matched baseline. QVHighlights

Hit@1 interval includes zero. Core96K: Omni-7B 96K has two seeds and no significance test, RefCOCO val 0.26pp below the frozen trunk, and at 12K OGF is below LoRA on 85 of 87 trunk–benchmark pairs.

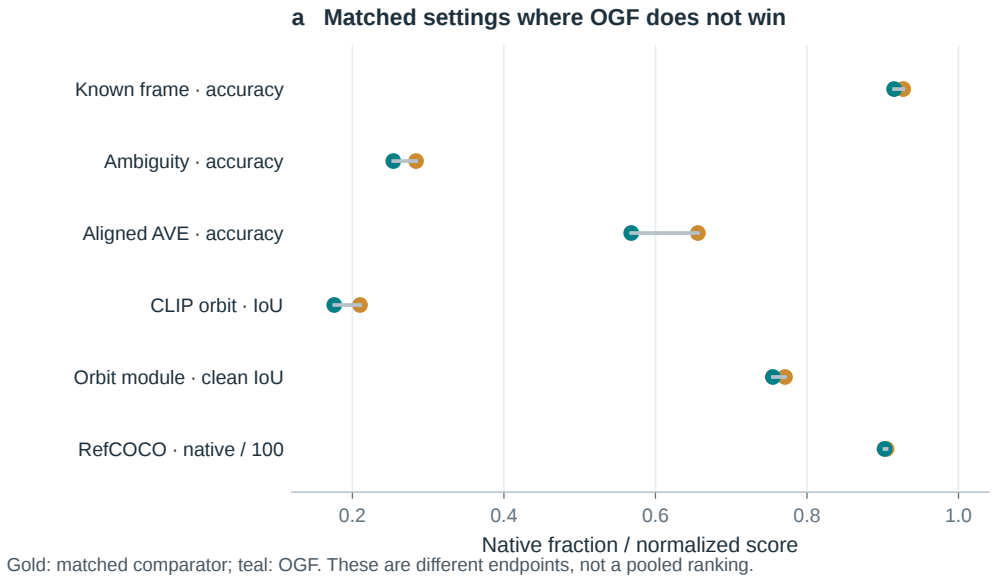


Figure 29: Matched settings where OGF does not win.



Figure 30: SNR, runtime, p-values, and quantified boundaries.

Table 39: Results that narrow the claims.

Result	Measurement	What it narrows
Determinate frame	OOD 0.915 vs hard canon. 0.927, $p = 1.0$	OGF is not more accurate everywhere
Pure ambiguity	acc 0.254 vs MAP 0.284	calibration, not accuracy
Aligned AVE	naive clean acc 0.656 vs 0.568	no posterior needed without a nuisance
AVE offset	top offset correct $\leq 15\%$	flat AP $\neq$ recovered offset
Factorial	group MLP 0.903 vs OGF 0.773; OGF seed-bimodal	fast fit on an inner-product label
Frame poisoning	entropy flat at 0.34	posterior does not flag broken frames
Audio+Point pair	0.332, interval includes chance	not a stable second bit

Table 39: Results that narrow the claims (continued).

Result	Measurement	What it narrows
CREMA SNR sweep	-10 dB interval includes 0.5	weak; does not replace circular sync
AVE family recall@1	$0.829 < 0.90$	pre-specified threshold missed
LongVALE	downstream F1/IoU below matched	moment mAP is not open-vocabulary localization
SigLIP2 uncompiled runtime	1.22–$1.70\times$	slower than a compiled forward
Discrete C_4	ties a C_4-augmented Q-Former	exactness, not score
Frozen Qwen2-VL C_4	interval against C_4-aug crosses zero ($p = 0.29$)	discrete orbits are tileable by augmentation
Frozen CLIP geometry	0.176 vs TTA 0.210	post-hoc geometry is insufficient
Orbit-module clean IoU	clean IoU 0.755 vs LoRA 0.771	orbit module is not an unconditional clean gain
Audio reversal	speech flip 0.344	not a meaning-preserving nuisance
QVHighlights Hit@1	relative to dense $+0.31$ $[-0.57, 1.13]$ ($p = 0.625$)	moment mAP does not imply highlight hit rank
Six frozen-encoder tasks	mean drop 0.321pp (max 1.42pp on CREMA audio)	frozen-encoder heads; Core96K reports the trunks
Coupling spectrum	Dense rank higher than OGF; Late/GMU balanced	gradient balance is not unique to group fusion
Core96K seeds	two seeds differ by a few items; spread is not a CI	no significance test
Core96K visual	RefCOCO val 90.24 vs frozen 90.50	slight regression
Core96K 12K	OGF $<$ LoRA on 85 of 87 trunk–benchmark pairs	the path needs data
Core96K Qwen3-VL-4B	leads LoRA on $16/16$ at 96K (mean $+0.21$pp); 12K is below or tied	the margin appears with scale

J PUBLIC ABSOLUTE-REFERENCE BASELINES

The following table reproduces publicly reported RefCOCO-family val scores for absolute positioning only. Matched experimental baselines remain the sole basis for causal attribution.

Table 40: Public RefCOCO-family val scores (Yu et al., 2016; Mao et al., 2016). Absolute context only; not used for attribution.

Method	Venue	Rec val	Rec+ val	Recg val
MDETR (RN101)	ICCV'21	86.75	79.52	81.64
Qwen-VL	arXiv'23	89.36	83.12	85.58
Grounding-DINO-L	arXiv'23	90.56	82.75	86.13
OneRef-L	NeurIPS'24	92.87	87.98	88.11

K MATCHED BASELINES AND ABLATIONS

Every matched method in a given table receives the same frontends, the same training split, the same update count, and the same optimizer family, unless the table caption states a deliberate capacity or budget advantage (capacity-fair TUM baselines, ScanRefer regressor). Reported comparisons are therefore of two kinds: parameter-matched or same-order (Stage I fusion table; Core96K OGF 9.20M vs. LoRA-only 9.58M) and capacity-favored against OGF (TUM three-arm protocol, ScanRefer regressor). Those two kinds are never averaged.

Ablations that appear in the paper, with the section that records them: no typed stream / $\ell_{\max} = 0$ (App. C); identity frames and Anchor-OT (App. D); point versus posterior frame (ambiguity/conflict regimes and ScanRefer); oracle / no / wrong frame (sections C, F and H); no temporal posterior (AVE no-group); no correspondence (Core96K no-correspondence retrain, ScanRefer uniform); pairing shuffle (Stage I and Core96K); missing-modality and conflict (sections C and D and TUM RGB–depth conflict). The construction uses $L = 2$, so $\ell_{\max} = 3, 4$ are outside the claim.

L SYSTEM AND REPRODUCIBILITY PROTOCOL

Stage I and Stage II tables are means over the seeds named in each caption. A pre-specified threshold that is missed is reported as a miss (table 39). A seed that did not fit is kept in the mean: OGF seed 1 on the 10-seed world, and two of five independent-frame factorial seeds near 0.47. Omni-7B at 96K is the mean of two training seeds. Those seeds differ by a few items. No significance test is attached. Every other Core96K trunk and scale is one training seed.

M Additional Limitations and Societal Considerations

The semantic/typed decomposition is itself a modeling choice: some concepts mix geometry, convention, and culture. A wrong typing can hard-code an inappropriate invariance. Sensor metadata may be unavailable or malicious. Confidence heads can be miscalibrated under distribution shift. Multimodal provenance is imperfect, and explanations derived from anchors are not causal guarantees. Evaluation must include demographic and device diversity where people, speech, or embodied interaction are involved. The release should support opt-out and deletion linkage between training lists and model versions.

N Sources of Reported Numbers

Controlled-world and frozen-encoder tables are means over the seeds named in each caption. Omni-7B at 96K, including its column of table 38 and its row of table 32, is the mean of two training seeds. Omni-3B, LLaVA-OV, InternVL2, Qwen3-VL, and Omni-7B at 12K/24K/48K are one training seed. Published Omni-7B scores are used only to check the frozen trunk, not to attribute gains.

O Notation and Formal Construction

Let $\mathcal{G} \rightrightarrows \mathrm{Ob}(\mathcal{G})$ be a small groupoid of local frames. For each object i and order ℓ, attach a vector space $V_i^{(\ell)}$. A typed feature functor maps an arrow $g_{j\leftarrow i}$ to an invertible linear map $\mathrm{D}^{(\ell)}(g) : V_i^{(\ell)} \to V_j^{(\ell)}$ and preserves identities and composition. Semantic features use the trivial representation only with respect to geometric coordinate changes; they are not assumed invariant to content changes, paraphrase, or interventions.

A world anchor is another groupoid object chosen as a computational gauge. A posterior $q_{a\leftarrow i}$ is a probability measure over arrows from i to a. Under a gauge change k_i at i and k_a at a, the posterior transforms by pushforward under $g \mapsto k_a g k_i^{-1}$. This is the key condition behind gauge independence.

P Proofs

P.1 Global equivariance

For a common world action u, equivariant frontends give $h_i'^{(\ell)} = \mathrm{D}^{(\ell)}(u_i)h_i^{(\ell)}$. Invariant descriptors leave α_{ia} unchanged. Posterior covariance maps g to $u_a g u_i^{-1}$. Therefore

$$\widetilde{h}_{ia}'^{(\ell)} = \int \mathrm{D}^{(\ell)}(u_a g u_i^{-1})\mathrm{D}^{(\ell)}(u_i)h_i^{(\ell)}\, dq(g) \tag{12}$$

$$= \mathrm{D}^{(\ell)}(u_a) \int \mathrm{D}^{(\ell)}(g)h_i^{(\ell)}\, dq(g), \tag{13}$$

so each anchor transforms in the anchor frame. Normalized weighted sums preserve this law, and compositions of intertwiners, invariant attention weights, CG products, and invariant gates remain equivariant. Invariant contractions remove the final representation action.

P.2 Gauge independence

Let k_i change the local convention and k_a the anchor convention. Features become $\mathrm{D}(k_i)h_i$, while the posterior is pushed forward by $g \mapsto k_a g k_i^{-1}$. Substitution in eq. (7) cancels k_i and yields $\mathrm{D}(k_a)\widetilde{h}_{ia}$. Thus an invariant head is unchanged; an equivariant head changes only with the declared output gauge.

P.3 Stabilizers

Suppose x is unchanged under $H \leq G$. Any representative gH describes the same observable pose. Averaging transport over the posterior on gH applies the projector $|H|^{-1}\sum_{h\in H}\mathrm{D}(gh)$ in the finite

case (or the Haar integral for compact H). Replacing g by gh_0 permutes the integral, so the result is representative-independent. Components outside the H-fixed subspace vanish. Non-compact or continuous cases require an explicit normalized quotient model and are approximated numerically.

P.4 Composition

Functoriality gives $\mathrm{D}(g_{i\leftarrow j})\mathrm{D}(g_{j\leftarrow k}) = \mathrm{D}(g_{i\leftarrow k})$. For independent posterior arrows, sequential transport uses the convolution measure $q_{i\leftarrow j} * q_{j\leftarrow k}$. If $q_{i\leftarrow k}$ differs from this convolution, bounded-Lipschitz duality bounds output discrepancy by the corresponding Wasserstein distance times the representation Lipschitz constant and feature norm.

P.5 Stability

Write a normalized anchor as $H = N/M$ with $M \geq \mu$. Add and subtract perturbed numerator and denominator terms and use $\|N\| \leq BM$. Assignment perturbations contribute at most $O(B\|\widehat{\alpha} - \alpha\|_1/\mu)$. Coupling the two frame posteriors and applying Lipschitz continuity of $\mathrm{D}^{(\ell)}$ contributes $O(BL_\ell W_1(\widehat{q}, q)/\mu)$. Feature perturbations contribute $O(\|\widehat{h} - h\|/\mu)$. Summing and applying the Lipschitz constants of subsequent blocks proves the stated bound. The lower-mass condition is why occupancy and empty-anchor diagnostics are part of the method rather than implementation details.

P.6 Complexity and streaming

Scoring all token–anchor pairs costs $O(NR)$ and aggregating R anchors costs $O(NR)$; anchor self-attention costs $O(R^2)$. Sparse retrieval changes the first term to index lookup plus $O(Nk)$. The recurrent state contains only semantic/typed sums, masses, time statistics, and per-modality summaries for each anchor, hence $O(Rd)$ state independent of processed sequence length. Dense decoders are materialized only for requested outputs.

Q Non-identifiability and Boundary Cases

Proposition 1 (Text gauge)**.** *Losses constructed only from inverse relations, cycle closure, orthogonality, and relative spatial constraints are invariant under one common $Q \in O(3)$ applied to all learned direction vectors. Absolute axes are therefore unidentifiable without visual, gravity, map, body, or other external anchors.*

Proposition 2 (Pose quotient)**.** *If an observation has stabilizer H, no deterministic estimator can identify a unique element of G from that observation alone; only an equivalence class in G/H is identifiable.*

Proposition 3 (Finite raster)**.** *A generic continuous rotation does not act as a permutation of a finite square pixel lattice. Exact arbitrary-angle equivariance therefore requires a declared continuous/band-limited signal model and resampling operator; unconditional exactness is restricted to lattice-preserving subgroups such as C_4.*

R Exact Raster Reference Path and Text Gauge Grounding

For the lattice-preserving subgroup C_4, the raw-RGB reference path forms an orbit lift

$$L_r(x) = \phi(R_{-r}x), \qquad r \in \{0, 1, 2, 3\}, \tag{14}$$

with shared ϕ. Regular-action channels are projected into invariant and typed Fourier components; every subsequent operation follows the same transport rule. Because 90° rotations are exact pixel permutations, the complete RGB-to-output path admits machine-precision checks before and after an optimizer update. For arbitrary angles, the raw-RGB reference path reports error versus angular/radial bandwidth under the declared continuous-field interpolation model rather than claiming unconditional exactness.

For text, relation vectors d_r are learned from inverse relations, cycle closure, orthogonality, and anti-collapse losses. These objectives determine the relational geometry only up to one global $Q \in O(3)$. A visual, gravity, map, or body anchor is introduced in a separate split to test gauge fixing; the unanchored split is never scored against an arbitrary absolute axis.

S Optional causal extension

OGF may expose object and event variables that are more stable across sensor frames, after which conventional causal discovery can be applied under its own assumptions. We do not run causal discovery on the fused variables. Continuous acyclic-graph search, or additive-noise methods, on synthetic structural causal models would be a separate study. Observation-only data leave hidden confounding, selection, measurement error, Markov equivalence, and model misspecification unresolved. No causal result enters the reported evaluations.

T Architecture Details

T.1 Inputs passed into fusion

Each frontend returns semantic tokens, typed tensors indexed by ℓ, validity and availability masks, local coordinates, timestamps and clock source, modality/sensor identifiers, entity-event distributions, frame and temporal posteriors, separate semantic/structural confidence, and provenance. Missing modalities are absent, not represented by fake zero observations. Schema, tokenizer, codec, and architecture identifiers are stored in checkpoints.

T.2 Training schedule

The reported Core96K runs follow the three-stage schedule in App. H: the sidecar alone, then the sidecar with LoRA on query and value projections, then both while up-weighting shuffle, time shift, conflict, and missing modalities. Checkpoints are selected on a held-out development split.

T.3 Numerical precision

Transformer GEMMs use bfloat16. Sinkhorn log normalization, posterior logits, Gram accumulation, Kabsch/SVD, and loss reductions use float32. Float64 randomized checks are the reference for lower precisions. Parity, finite gradients, an optimizer step, and post-update equivariance hold at each precision, including empty masks, zero confidence, repeated singular values, and symmetric inputs.

U Complete Evaluation Protocol

U.1 Held-out multi-frame set

This set is separate from Core96K. Each example contains two to four modalities, an entity/event correspondence, and at least one relative frame or temporal variable. Train/test frame distributions differ. Symmetric cases carry quotient labels. Every item receives swap, shuffle, wrong-frame, delay, and missing-modality counterfactuals. Strong single-modality and proper-subset oracles are trained first; samples they solve above a frozen confidence threshold are removed. A manually reviewed subset audits label validity and residual shortcuts. Construction code, manifests, and a held-out test evaluator are released.

U.2 Metrics

Stage I–II report Syn; Stage III reports $\mathrm{Syn}_{\mathrm{route}}$, together with shuffle gaps, geodesic residuals, and posterior NLL/ECE/Brier.