STRUCTURE FOLLOWS CAUSALITY: CAUSALLY-ALIGNED REPRESENTATION SEPARATION BREAKS ORTHOGONAL BULLYING IN MULTI-TASK LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

When tasks form a causal cascade in multi-task learning—the upstream output conditions the downstream task—one shared representation must encode both the causal parent mechanism and the causal child mechanism, which in the regime we characterize (SGD-class optimizers, sparse high-dimensional downstream, incoherent coupling matrices) occupy incompatible subspaces: task gradients are near-orthogonal ($\cos\theta_g \approx 0$, $\theta_g := \angle(\mathbf{g}_1, \mathbf{g}_2)$) and wildly imbalanced ($\|\mathbf{g}_1\|/\|\mathbf{g}_2\| \gg 1$), so the dominant task's updates repeatedly erase the weaker task's learning as the latter approaches its optimum—whose pure-curvature form ($E > 1$) we verify constructively as a *transient* rather than asymptotic window. We call this **orthogonal bullying** and prove it is structural, not algorithmic: the flat joint objective has no stationary point in the persistent excitation region, bullying is unavoidable under testable conditions on dominance, curvature coupling and the crossover threshold, and post-hoc mediators face a detection blind spot or a mediation trilemma (Thms. 1, 2, 3a/3b). We advance **Structure Follows Causality**—the computational graph must mirror the task causal graph—realized by the **Causal Graph Architecture (CGA)**: each DAG node lives in its own module and conflicting gradient paths are cut by stop-gradient (Props. 4/4′, edgewise on any DAG; one line of `.detach()`), compiled deterministically from any causal DAG (Alg. 1). Under a symmetric comparison protocol—the flat baseline is jointly trained and must satisfy a protection constraint that $\mathcal{T}_2$'s loss never increases at any step, without which flat $\mathcal{T}_2$ fails outright and the rate comparison is vacuous—CGA reduces the iteration complexity of reaching upstream accuracy ε from $\mathcal{O}(\kappa_g^2 \log(1/\varepsilon))$ to $\mathcal{O}(\log(1/\varepsilon))$ (Thm. 5); a cascade-gap decomposition (Thm. 6) shows CGA shares the flat model's information-theoretic floor—**its value is representation separation of causal mechanisms, not a spurious statistical gain**. On controlled synthetic SCMs CGA holds the single-task floor with near-zero variance (T2 0.10–0.12, T1 0.13; std$\leq$0.003; 12 configs, 3 seeds), while Flat fails in mechanism-specific segments; on the v2 legacy κ-grid CAGrad's upstream error crashes to 71.9–87.0% and GradNorm's to 53.3% (Flat and CGA stay $\leq$1.1%). Under strong coupling the system enters a first-order coupling protection regime where bullying does not activate, bounding the assumptions' scope; Flat still pays a 1.7–2.1$\times$ tax and CGA matches the floor. On a production 270-station metro system (Adam + global clipping), joint-training logs record the signature—gradient cosine pinned at 0.00, $\kappa_g = 200$–3000, failed mediation, WAPE 36.6%$\rightarrow$37.7%—registered as **motivational log only; not a mechanism check**, since the production Adam+clipping setting is not reproduced by either controlled arm; the separated flow predictor reaches 10.34% WAPE (Stage-1 trained alone). A pre-registered three-arm *exploratory* comparison on real large-scale OD data under an identical anti-collapse recipe (**single seed; multi-seed confirmation pending**) gives flat joint 62.1%, single-task OD without flow 61.2%, CGA 42.7%—a 19.4pp gap indicating joint training cancels the parent's information contribution, *preliminary* evidence that only causal structure converts it into realizable gain. Further controlled experiments close three gaps: pure-curvature erasure is realized constructively (81/81 single-step, 27/81 sustained)

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.

but is transient and unreachable in the MLP family; eight alternative flat protocols all fail to reproduce CGA's floor; and structured sharing (MMoE/PLE/AdaShare) relieves the downstream only by transferring cost upstream (T1 0.300 vs CGA 0.13), leaving CGA the only method on the floor for both tasks. We also give a three-part decision procedure (directionality $\wedge$ $\hat{\Gamma}$ $\wedge$ bullying indicators). Boundaries and negative results are reported honestly: the controlled κ_g stays below production scale, and the erasure window is transient.

1 INTRODUCTION

Multi-task learning serves all tasks from one shared representation. When tasks form a *causal cascade*—the upstream output conditions the downstream task—that representation must encode two causally non-independent mechanisms: the causal parent (how the upstream target is generated) and the causal child (how the downstream target is allocated given it). In the regime we characterize (Assumptions 1–2, $c_H > 0$, κ_g above the crossover threshold κ^*) this induces an unavoidable *representation entanglement pathology*—improving one task collapses the other—whose root is structural: the representation is not aligned with the causal order of the physical process.

Gradient-mediation methods (GradNorm, MMoE, PCGrad, CAGrad, ConicGrad, GradOPS, IMTL-G; §2) are *post-hoc*, seeking a compromise after the conflict has formed. But flat architectures' gradient geometry reveals a harder fact: in causal cascades task gradients are near-orthogonal ($\cos\theta_g \approx 0$, $\theta_g := \angle(\mathbf{g}_1, \mathbf{g}_2)$) and highly imbalanced ($\|\nabla_1\|/\|\nabla_2\| \gg 1$; 200–3000 in production), so the dominant task's update is not "help" but noise that repeatedly erases the weaker task's learning.

We call this pattern—incompatible directions, asymmetric magnitudes—**orthogonal bullying**. The dominant task inflicts purely *second-order* damage: the weak task's per-step loss changes by $-\eta\|\mathbf{g}_2\|^2 + \frac{\eta^2}{2}\mathbf{g}_1^\top H_2 \mathbf{g}_1$, with erasure ratio $E \gg 1$ under non-degenerate curvature coupling (Lemma 2). Methods that only adjust direction or only balance magnitude treat the symptom, never the structural–causal mismatch: causal cascades follow an irreversible arrow, and bullying is the price of violating it.

Our contribution is *theoretical explanation and paradigm establishment*, not a novel engineering structure—"cascade + Stop-Gradient" already exists in recommendation and speech pipelines; we supply its missing first-order grounding and its exact scope. Since flat cannot even converge to a stationary point, its floor is *vacuous*: CGA's value is not to surpass the floor but to reach it.

Contributions. (i) *Structure Follows Causality*, the phenomenon *orthogonal bullying*, and a compiler from any causal DAG to an edge-wise-decoupled architecture (Alg. 1). (ii) **Theory**: no stationary point (Thm. 1), unavoidable bullying under a testable assumption (Thm. 2), a detection blind spot and mediation trilemma (Thms. 3a/3b), an $\mathcal{O}(\kappa_g^2)$ rate gain (Thm. 5), and cascading's exact statistical position—same floor, non-negative deficit (Thm. 6). (iii) **Experiments**: a κ-grid where CAGrad crashes to 71.9–87.0% while CGA holds the floor; a rate tax of measured slope 2.001; and a real-data structural comparison (19.4pp, single seed).

2 RELATED WORK

Gradient conflict and mediation. Mediation methods act after conflicts form in shared parameter space: magnitude balancing (GradNorm Chen et al. (2018), IMTL-G Liu et al. (2021b)), projection (PCGrad Yu et al. (2020), ConicGrad Hassanpour et al. (2025), GradDrop Chen et al. (2020)), constrained optimization (CAGrad Liu et al. (2021a), Nash-MTL Navon et al. (2022), FAMO Liu et al. (2023), MGDA Sener & Koltun (2018)), subspace methods (GradOPS Zhu et al. (2025), Aligned-MTL Senushkin et al. (2023)), uncertainty weighting Kendall et al. (2018), and structured sharing (MMoE Ma et al. (2018a), PLE Tang et al. (2020), AdaShare Sun et al. (2020))—the last restructures sharing *before* training, falling outside Thm. 3b; the rest output one direction in shared space and inherit its trilemma.

Causal learning and cascaded prediction. Unlike causal representation learning Schölkopf et al. (2021), IRM Arjovsky et al. (2019) and counterfactual guidance Buesing et al. (2019), we target

gradient-pathology elimination; cascaded prediction with stop-gradient is established engineering Ma et al. (2018b); Graves et al. (2006)—we supply its first-order grounding (Appendix A).

3 Problem Formulation and Theoretical Analysis

3.1 Structural Causal Model and Causal Cascade Tasks

Definition 1 (SCM and causal cascade task)**.** *Let $\mathcal{M} = \langle \mathcal{U}, \mathbf{V}, \mathcal{F}, P(\mathcal{U}) \rangle$ be an SCM. For a system with input X and targets Y_1, Y_2, if $\mathbf{V} = \{X, Y_1, Y_2\}$ with causal graph $X \to Y_1$, $X \to Y_2$, $Y_1 \to Y_2$, $Y_1 \not\perp\!\!\!\perp Y_2 \mid X$, and structural equations $Y_1 = f_1(X, U_1)$, $Y_2 = f_2(X, Y_1, U_2)$ with $U_1 \perp\!\!\!\perp U_2$ and $\partial f_2 / \partial Y_1 \neq 0$, then $\mathcal{T}_1$ (predict Y_1) and $\mathcal{T}_2$ (predict Y_2) form a **causal cascade task pair**, $\mathcal{T}_1 \to \mathcal{T}_2$.*

In a flat shared architecture $Z = g_\theta(X)$ serves both tasks; by Def. 1, $P(Y_2 \mid X)$ marginalizes over Y_1, so $\nabla_\theta \mathcal{L}_2$ is sensitive to θ in *both* the "predict Y_1" and "predict Y_2 given Y_1" roles—the root of the decomposition below.

3.2 Gradient Decomposition and Non-Convergence

Let shared parameters be θ, task parameters ϕ_1, ϕ_2, losses $\mathcal{L}_1, \mathcal{L}_2$, and $\mathbf{g}_i = \nabla_\theta \mathcal{L}_i$.

Lemma 1 (Gradient decomposition)**.** *By the SCM, information about Y_2 reaches Z via two channels: one mediated by the parent Y_1, one bypassing it. These correspond to subspaces $\mathcal{S}_1$ (directions covariant with Y_1; $\mathcal{T}_1$'s sensitive directions) and $\mathcal{S}_2$ (Y_2's residual directions), with $\mathcal{S}_1 \perp \mathcal{S}_2$ in population covariance (Appendix E). Hence $\mathbf{g}_2 = \mathbf{g}_2^{\mathcal{S}_1} + \mathbf{g}_2^{\mathcal{S}_2}$, $\mathbf{g}_2^{\mathcal{S}_i} \in \mathcal{S}_i$.*

Assumption 1 (Subspace separation)**.** *In a causal cascade with a flat shared architecture, $\mathbf{g}_1 \in \mathcal{S}_1$ and the angle between the two task gradients,*

$$\theta_g := \angle(\mathbf{g}_1, \mathbf{g}_2), \text{ satisfies } \theta_g \approx \tfrac{\pi}{2}, \text{ i.e. } \cos\theta_g \approx 0.$$

Notation. *θ_g is the gradient angle (distinct from parameters θ); $\epsilon_\perp$ is the orthogonality slack ($|\cos\theta_g| \leq \epsilon_\perp$), distinct from accuracy ε. Calibrated by measurement (§5.3)—SGD arm $\cos\theta_g \in [-0.14, +0.03]$, production 0.00 ± 0.02—so we* do not *assert $\epsilon_\perp \to 0$; proofs need only $\epsilon_\perp \ll 1/\kappa_g$.*

Assumption 1 is justified structurally (Lemma 1, Appendices E, I), by saturation ($L_1 \gg L_2 \Rightarrow \mathbf{g}_2^{\mathcal{S}_1} \to 0$) and by measurement (production cos ≈ 0.00; controlled mean -0.0024). *Scope*: under strong coupling cos rises to $+0.3$–0.8 and κ_g falls to 0.1–0.3, entering a *first-order coupling protection* regime where bullying does not activate (Table 6); Flat still pays a residual tax there. *Evidence gap*: we report only the scalar $\cos\theta_g$; the principal angles between $\mathcal{S}_1$ and $\mathcal{S}_2$, subspace-overlap/CKA diagnostics, and c_H on *real* data are *not* measured. Since Appendix E.3 concedes that non-intersecting covariances do not imply subspace orthogonality, Assumption 1 should be read as an empirical claim testable by those three diagnostics, not as a purely structural deduction.

Theorem 1 (No stationary point, non-convergence)**.** *Under Assumption 1, for any $\lambda > 0$ the flat objective $\mathcal{L} = \mathcal{L}_1 + \lambda \mathcal{L}_2$ satisfies: (1) $|\cos\theta_g| \leq \epsilon_\perp \ll 1$; (2) if $\mathbf{g}_1 + \lambda \mathbf{g}_2 = 0$ then $\kappa_g \leq \lambda \epsilon_\perp$, hence* no stationary point exists *in the persistent-excitation region ($\|\mathbf{g}_1\| \geq c > 0$, $\kappa_g > \lambda \epsilon_\perp$); (3) under standard smoothness and small (or decaying) steps Bertsekas (1999); Bottou et al. (2018) every accumulation point of gradient descent is stationary, so the iteration* cannot converge *there—the structural reason Flat's $\mathcal{T}_2$ is poor and seed-sensitive.*

3.3 Orthogonal Bullying: Definition and Necessity

Definition 2 (Gradient condition number and orthogonal bullying)**.** *The **gradient condition number** is $\kappa_g = \max_{i,j} \|\mathbf{g}_i\| / \|\mathbf{g}_j\|$. Under a flat shared architecture, orthogonal bullying occurs when: (1)* directional orthogonality $\cos\theta_g \approx 0$ *($|\cos\theta_g| \leq \epsilon_\perp \ll 1$, $\theta_g := \angle(\mathbf{g}_1, \mathbf{g}_2)$)—**sign matters**: clearly positive* cos *(first-order coupling* protects *the weak task) and clearly negative* cos *(ordinary, PCGrad-detectable conflict) are both* outside *the regime, which denotes the case where the first-order term vanishes and destruction moves to second order; and (2)* magnitude disparity $\kappa_g \gg 1$.

***Necessary vs sufficient**: this gives the* regime*; intensity is $E = \frac{\eta}{2} \cdot \frac{\mathbf{g}_1^\top H_2 \mathbf{g}_1}{\|\mathbf{g}_2\|^2}$, and $E > 1$ also needs $c_H > 0$ (hence Thm. 2 adds $c_H \geq c_0 > 0$); at $c_H \approx 0$ the regime holds without erasure.*

The dominant update is then high-frequency noise for the weak task (learn–forget–relearn); $E > 1$ means its loss rises *every step.*

Assumption 2 (Dominant task)**.** *In a causal cascade $\mathcal{T}_1$ reconstructs a dense base distribution (metro flow: 270 dims) while $\mathcal{T}_2$'s label is high-dimensional and sparse (OD: 72,900 dims, ∼80% zeros/hour), so $\mathcal{T}_1$'s surface is steeper: $L_1 \gg L_2$, i.e. $\kappa_g \gg 1$.* ***Testable****: production logs give $\kappa_g > 200$ (peaking at 3000).*

Lemma 2 (Second-order erasure)**.** *Under Assumption 1, for the naive joint update $\mathbf{d} = \mathbf{g}_1 + \mathbf{g}_2$ with step η, the one-step change of $\mathcal{L}_2$ is*

$$\Delta\mathcal{L}_2 = -\eta\|\mathbf{g}_2\|^2 + \tfrac{\eta^2}{2}\mathbf{g}_1^\top H_2 \mathbf{g}_1 + O(\eta^3\|\mathbf{d}\|^3 + \eta^2\epsilon_\perp\|\mathbf{g}_1\|\|\mathbf{g}_2\|), \tag{1}$$

with $H_2 = \nabla^2_\theta \mathcal{L}_2$: the first term is $\mathcal{T}_2$'s descent, the second pure second-order damage *from the dominant task ($\mathbf{g}_1 \perp \mathbf{g}_2$ gives no first-order help). With $E = \frac{\eta}{2}\kappa_g^2 c_H$ ($c_H = \hat{\mathbf{g}}_1^\top H_2 \hat{\mathbf{g}}_1$), $E > 1$ implies $\Delta\mathcal{L}_2 > 0$: the optimizer nominally optimizes $\mathcal{T}_2$ while destroying it.*

Theorem 2 (Necessity of orthogonal bullying)**.** *Under Assumptions 1–2 with $c_H \geq c_0 > 0$, orthogonal bullying necessarily occurs: both conditions of Def. 2 hold and $E = \frac{\eta}{2}\kappa_g^2 c_H \gtrsim \kappa_g^2 c_0/(2L_1)$ at $\eta \asymp 1/L_1$, so $E > 1$ once $\kappa_g^2 c_0 \gtrsim 2L_1$. Loss-weight tuning cannot remove it (Thm. 3b).*

All three conditions are measurable (Appendix F.1). $c_H = \hat{\mathbf{g}}_1^\top H_2 \hat{\mathbf{g}}_1$ by Hessian–vector products; the crossing condition $\frac{\eta}{2}\kappa_g^2 c_H > 1 + \kappa_g \cos$, i.e. $\kappa_g > \kappa^* := (\cos + \sqrt{\cos^2 + 2\eta c_H})/(\eta c_H)$, separates bullying from ordinary conflict. Note $\kappa^* \to \sqrt{2/(\eta c_H)}$ as $\cos \to 0$: orthogonality alone does not imply erasure.

3.4 Detection Blind Spot and Mediation Trilemma

Theorem 3a (Detection blind spot)**.** *Let a mediator Φ trigger a correction only when $\langle \mathbf{g}_1, \mathbf{g}_2 \rangle < -t$ ($t \geq 0$; projection family: PCGrad, ConicGrad). Under Assumption 1, if $t > \epsilon_\perp\|\mathbf{g}_1\|\|\mathbf{g}_2\|$ the trigger probability is zero; at $t = 0$ the measured rate is 0.0% on the v2 legacy κ-grid and near zero in v3's steady region, so such operators degenerate to naive summation, leaving E uncontrolled.*

Remark (wandering region). *The 0.0% rate is measured in the* steady *region; in Thm. 1's wandering region,* $\cos$ *fluctuates about zero and occasional negative excursions* do *trigger projection. This delimits the theorem (a steady-state statement): projection's benefit comes precisely from this pathology and is limited.*

Theorem 3b (Mediation trilemma)**.** *Let Assumption 1 hold, $H_2 \succeq 0$ be locally non-degenerate ($|c_{12}| \leq \rho\sqrt{c_H c_2}$, $\rho < 1$), and $c_H \geq c_0 > 0$. Then for $\kappa_g \geq K$ (an explicit threshold),* any instantaneous, stateless *mediator* Φ in shared parameter space *whose output $\mathbf{d} = \Phi(\mathbf{g}_1, \mathbf{g}_2)$ depends only on the current gradients—covering the four families below; methods that restructure sharing itself (MMoE, PLE, AdaShare) fall outside this statement (§4.3)—satisfies at most two of: (i) $\mathcal{T}_1$* ***advances at full speed****: $\langle \mathbf{d}, \hat{\mathbf{g}}_1 \rangle \geq (1-\delta)\|\mathbf{g}_1\|$; (ii) $\mathcal{T}_2$* ***is not damaged****: $\Delta\mathcal{L}_2(\mathbf{d}) \leq 0$; (iii) $\mathcal{T}_2$* ***advances effectively****: $\langle \mathbf{d}, \mathbf{g}_2 \rangle \geq (1-\delta)\|\mathbf{g}_2\|^2$.*

Scope and the four families. Projection sacrifices (ii); subspace methods face unsolvable implicit estimation at $\kappa_g \gg 1$; fairness-constrained methods sacrifice (i) at $\mathcal{O}(\kappa_g^2)$—the factor CGA recovers in Thm. 5; magnitude balancing sacrifices (i) at $\Theta(\kappa_g)$. Common root: curvature damage cannot be undone by any first-order operation; the only exit is to remove the shared space (Prop. 4).

3.5 Structure Follows Causality

The traditional approach remains at the *associational* level of Pearl's hierarchy Pearl (2009): it fits $P(Y_1, Y_2 \mid X)$ in one parameter space, agnostic to the generating mechanisms $\mathcal{F}$. A cascade is a generative process—generate Y_1, then Y_2 given Y_1—so the computational graph should mirror it:

Structure Follows Causality: *the topology of the neural computational graph must be isomorphic to that of the task causal graph*—equivalently, *causally-aligned representation separation*: parent and child mechanisms live in independent modules with one-way information flow. Three requirements follow—causal-level separation, directed information flow, gradient-block alignment—which reduce "fitting a causally entangled joint in a flat space" to "fitting well-structured conditionals sequentially" (§4).

4 CAUSAL GRAPH ARCHITECTURE

4.1 ARCHITECTURE

CGA realizes Structure Follows Causality via three constraints: **causal alignment** (computational graph isomorphic to the task causal graph), **parameter isolation** (disjoint parameter blocks per causal level), and **gradient blocking** (the path $\mathcal{L}_2 \to \theta_1$ cut).

Formally, for a cascade $\mathcal{T}_1 \to \mathcal{T}_2$ with input X and labels Y_1, Y_2, CGA consists of two cascaded modules. The **upstream module** M_1 (parameters θ_1) predicts the base state, $\hat{Y}_1 = f_{\theta_1}(X)$ with $\mathcal{L}_1 = \ell_1(\hat{Y}_1, Y_1)$. The **downstream module** M_2 (parameters θ_2) predicts the subsequent state conditional on it—its input *must* include $\hat{Y}_1$, with the gradient path to θ_1 cut by **stop-gradient**: $\hat{Y}_2 = f_{\theta_2}(X, \mathrm{sg}(\hat{Y}_1))$, $\mathcal{L}_2 = \ell_2(\hat{Y}_2, Y_2)$, where $\mathrm{sg}(\cdot)$ is the identity forward and has zero derivative backward. The joint objective is

$$\mathcal{L}_{\mathrm{CGA}}(\theta_1, \theta_2) = \mathcal{L}_1(\theta_1) + \lambda \mathcal{L}_2(\theta_2), \tag{2}$$

where $\lambda > 0$ is absorbed into the Stage-2 learning rate, so **CGA introduces no gradient-mediation hyperparameter** (λ times that rate, the per-stage epoch budget and the capacity split remain hyperparameters). By blocking and isolation, θ_1 is driven purely by $\mathcal{L}_1$, θ_2 purely by $\lambda\mathcal{L}_2$ (Prop. 4).

4.2 CGA($\mathcal{G}$): AN ARCHITECTURE FAMILY OVER ARBITRARY DAGS, AND A COMPILER

The two-node chain is the minimal instance. For a general causal DAG $\mathcal{G} = (\mathbf{V}, E)$ (node V_i carries task Y_i), CGA($\mathcal{G}$) assigns each node an independent module M_i (parameters θ_i) with

$$\mathrm{inputs}_i = \{X\} \cup \{\mathrm{sg}(\hat{Y}_p) : p \in \mathrm{Pa}(i)\}, \qquad \mathcal{L} = \textstyle\sum_i \ell_i(\hat{Y}_i, Y_i). \tag{3}$$

Information flows one-way *forward* along causal arrows; backward gradient flow is cut by sg at every module boundary. Structure Follows Causality becomes an executable rule via the following compiler.

Compiler. Alg. 1 (Appendix L) maps any causal DAG to a module partition and detach positions deterministically; step 4 diagnoses applicability—when $\hat{\Gamma} \approx 0$ it warns that flat is also fine (Thm. 6). Multi-parent nodes need no extra machinery (diamond: module 4 receives $\{X, \mathrm{sg}(\hat{Y}_2), \mathrm{sg}(\hat{Y}_3)\}$).

4.3 THEORETICAL PROPERTIES

Proposition 4 (Gradient decoupling; orthogonal bullying eliminated)**.** *In CGA, for any causal cascade pair the cross-gradients w.r.t. the independent parameters vanish identically:* $\nabla_{\theta_1}\mathcal{L}_2 = 0$ *and* $\nabla_{\theta_2}\mathcal{L}_1 = 0$.

Proof sketch. $\hat{Y}_1$ is a constant tensor in backpropagation, so $\partial\mathcal{L}_2/\partial\theta_1 = 0$; $\mathcal{L}_1$ does not involve θ_2. □

No shared parameter space remains in which the two gradients could meet, so both preconditions of Def. 2 are removed at the graph level; only the *backward* flow is blocked—the forward edge $\hat{Y}_1 \to M_2$ survives.

Proposition 4′ (Full gradient decoupling on DAGs)**.** *For* CGA($\mathcal{G}$) *and any* $i \neq j$*:* $\nabla_{\theta_i}\mathcal{L}_j = 0$*, i.e.* $\nabla_{\theta_i}\mathcal{L} = \nabla_{\theta_i}\mathcal{L}_i$.

Corollary 1 (Edge-wise elimination)**.** *For any edge* $p \to c$*:* θ_p, θ_c *share no parameters and* $\nabla_{\theta_p}\mathcal{L}_c = 0$*, so Def. 2's preconditions are not expressible on that edge—elimination is* edge-wise*, with integrity from assembly per* $\mathcal{G}$.

Theorem 5 (Convergence rate advantage)**.** *Let* $\mathcal{T}_1 \to \mathcal{T}_2$ *be a cascade pair (Def. 1).* ***Comparison protocol****: both sides are jointly trained—flat with a single shared step* η *and the protection constraint that* $\mathcal{T}_2$*'s loss never increases (without it* $\mathcal{T}_2$ *fails by Thm. 1 and the comparison is vacuous), CGA with independent per-stage steps. Convexity is assumed only for CGA's rate. Then* $\mathcal{T}_1$ *in CGA converges* linearly*,* $\mathcal{O}(\log(1/\varepsilon))$*, whereas flat—with* $c_H \geq c_0 > 0$ *and the constraint binding (*$\eta_{\mathrm{safe}} < 1/L_1$*)—degrades to* $\mathcal{O}(\kappa_g^2 \log(1/\varepsilon))$.

Theorem 6 (Cascade gap decomposition)**.** *Let the cascade satisfy Def. 1 with squared loss. Write* $\mathrm{Bayes}(S) = \inf_h \mathbb{E}[(h(S) - Y_2)^2]$ *and* $\mu(x, y_1) = \mathbb{E}[Y_2 \mid X = x, Y_1 = y_1]$*, with* $y_1 \mapsto \mu(x, y_1)$ *uniformly L-Lipschitz. Then:* ***(i) Oracle cascade gap.*** *By the law of total variance* $\mathrm{Bayes}(X) = \mathrm{Bayes}(X, Y_1) + \Gamma$*,* $\Gamma := \mathbb{E}_X[\mathrm{Var}_{Y_1|X}(\mu(X, Y_1))] \geq 0$*: an optimized flat model still sits* Γ *above the oracle cascade.* ***(ii) Approaching the oracle with an estimate.*** $\inf_{h'} \mathbb{E}[(h'(X, \hat{Y}_1) - Y_2)^2] \leq \mathrm{Bayes}(X, Y_1) + L^2\mathrm{MSE}_1$*,* $\mathrm{MSE}_1 = \mathbb{E}\|\hat{Y}_1 - Y_1\|^2$ *(Appendix C.3).* ***(iii) Floor identity (key correction).*** $\hat{Y}_1 = f_{\theta_1}(X)$ *is deterministic in* X*, so any predictor based on* $(X, \hat{Y}_1)$ *is a function of* X*:* ***CGA Stage-2's information-theoretic floor coincides exactly with that of any flat model****. Combining (i)–(iii):*

$$\mathrm{Bayes}(X) \;\leq\; \inf_{h'} \mathbb{E}[(h'(X, \hat{Y}_1) - Y_2)^2] \;\leq\; \mathrm{Bayes}(X) + (L^2\mathrm{MSE}_1 - \Gamma), \tag{4}$$

(iv) The deficit is non-negative and can vanish *(Appendix C.3):* $\Gamma \leq L^2\mathrm{MSE}_1$*, so cascading can* at most match*, never undercut, the flat floor.* ***(v) Where the advantage really comes from.*** *CGA's advantage comes from removing the pathology of Thms. 1–3b: flat cannot reach its own floor, while CGA approaches the same floor with deficit* $L^2\mathrm{MSE}_1 - \Gamma$*; real-data gain ≈ flat's optimization gap* $-(L^2\mathrm{MSE}_1 - \Gamma)$.

Structured sharing (MMoE / PLE / AdaShare). These restructure sharing *before* training, so they fall outside Thm. 3b; CGA adds directed conditional inputs (M_c receives $\mathrm{sg}(\hat{Y}_p)$), a deterministic partition and stop-gradient. Prediction: they match CGA at $\kappa_g \approx 1$ and degrade toward Flat at $\kappa_g \gg 1$—confirmed in §5.5 (E5): they relieve the downstream but transfer cost upstream.

5 Experiments

We evaluate CGA on a real large-scale spatiotemporal forecasting task (§5.2), then verify orthogonal bullying and its dissolution at the optimization level on controlled synthetic data (§5.4–§5.5), and validate the compiler, graph misspecification and mechanism shift on multi-node DAGs (§5.6).

5.1 Experimental Setup

Setup. Real metro data (270 stations, chronological 70/15/15; Appendix B) and two synthetic SCMs (a $\beta \times \lambda_1$ grid and a calibrated κ-grid; Appendix H); nine baselines plus CGA; WAPE/MAE metrics and diagnostics cos, κ_g, c_H (Appendix G).

5.2 Real-Scale Spatiotemporal Forecasting: From Pathological Signature to Structural Comparison

Deployment reference (production pipeline; *not* a controlled experiment). CGA is deployed in a 270-station metro system: Stage-1 (flow) is trained then frozen; Stage-2 (OD) trains on frozen features and $\mathrm{sg}(\hat{Y}_{\mathrm{flow}})$. Flow WAPE **10.34%** (Stage-1 retrained alone—motivational, not a controlled gain); the historical OD figure is not reproducible and is not cited (Appendix B.1).

Branch B: OD collapse is statistical, and its remedy (Appendix B.4–B.6). Under pre-registered rules A–D the OD head collapses in every arm, *including* single-task $\mathcal{T}_2$, so per Thm. 6 no cascading gain can be claimed there. This is *statistical*, not gradient—the ZINB head collapses into the $\pi \to 1$ all-zero basin—and six countermeasures (baseline prior, π-rate anchoring, weighted NLL, tail μ floor, bounded heads, curriculum) lift per-interval OD WAPE to **42.7%**; they repair the statistical layer, CGA the optimization one.

Public-data boundary (Appendix B.7). On UCI Wine Quality Cortez et al. (2009) (10 seeds), cascade and flat sharing are indistinguishable (T2 MAE 0.659 ± 0.019 vs 0.658 ± 0.019) with $\cos \approx +0.1$–0.2, $\kappa_g \approx 0.5$–1.5—pathology-free, where Thm. 6 predicts no gain and none appears; the tie is a boundary result, **not evidence that CGA beats flat**. An oracle using true Y_1 is better (0.612) but unreachable.

Structural comparison: protocol v2 (Appendix B.7). With the recipe holding OD learnable, task structure is the only remaining variable. The three arms share the recipe, budget, and (for

`flat_joint` and `cga`) the *same* bit-identical curriculum-1 checkpoint; `flat_joint`'s anchor is its own jointly trained flow head (*not* frozen, unlike `cga`'s). Results: `cga` **42.7%**, `flat_joint` **62.1%**, `single_t2` **61.2%**; so `fj` − `cga` = 19.4pp and `st2` − `cga` = 18.5pp (both ≫ the 5pp pre-registered threshold) while `fj` − `st2` ≈ 0.9pp—joint training cancels the parent's contribution. *The same flow information is worth 19pp in the cascade and ≈0 in the shared pool*: one experiment discharges Thm. 2, Prop. 4 and Thm. 6 at once. **Caveat: single-seed (42), and `single_t2` changes encoder and anchor together.**

5.3 Gradient Dynamics and Orthogonality Diagnostics

Production and controlled diagnostics. The production log records the signature: cos ≈ 0.00, κ_g = 200–3000, gradient blowups, failed CAGrad/PCGrad/MMoE mediation, WAPE 36.6%→37.7%. Controlled runs reproduce it with **PCGrad's trigger rate at 0.0%** (fraction of steps with $\langle \mathbf{g}_1, \mathbf{g}_2 \rangle < 0$; at β=0 the mean cos is -0.04 but the sign oscillates and the correction is $\leq |\cos| \|\mathbf{g}_2\| \ll \|\mathbf{g}_2\|$)—Thm. 3a's blind spot (Table 7).

Estimating Γ; the "when to use CGA" decision. The headroom $\Gamma = \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y_2 \mid X, Y_1] \mid X)]$ is testable—Estimator A closes with $\beta^2\sigma_1^2$ at every β on v3 (Table 7). *Modelling assumption*: the direction $\mathcal{T}_1 \to \mathcal{T}_2$ (flow → OD) is *assumed*, supported by the physical time-ordering of entry flow versus realized OD matrices and by the synthetic construction of Appendix A; we do *not* run an independent directionality test on real data, so it is not a domain-verified fact. If the true direction were reversed, re-ordering the causal graph suffices—Props. 4/4′ and Alg. 1 are direction-agnostic, so only the architecture instantiation changes, not the theory. Decision rule (Alg. 1 step 4): *directionality* ($\beta_{1\to2}/\beta_{2\to1} > 2$) ∧ *headroom* ($\hat{\Gamma}$'s CI excludes 0) ∧ *bullying indicators* (cos ≈ 0, $c_H > 0$).

5.4 Controlled Verification on Synthetic Data

The κ-grid (Table 5, Appendix H.9). CAGrad's upstream error crashes to 82.7/71.9/87.0% at calibrated $\kappa = 10/20/50$ and GradNorm's to 53.3%, while Flat and CGA stay ≤1.1%—the crash is induced by mediation itself; cos averages -0.0024 and **PCGrad's trigger rate is 0.0%** (Thm. 3a).

Strong coupling: the failure mode of Assumption 1 (Table 6). At $\beta = 1/2/3$, coupling aligns the gradients (cos → +0.5 ∼ 0.8, $\kappa_g \to$ 0.1–0.3) and **bullying vanishes**; yet Flat still pays a 2.1×/2.1×/1.7× tax while CGA reaches parity at $\beta = 3$ (0.09 vs 0.10).

Table 1: Synthetic-Causal v3 grid: T2 WAPE, mean (std) over 3 seeds. ST2 = single-task floor. CGA holds the floor in every cell (per-cell std ≤0.003). Shown: $\beta \in \{0.0, 1.0\}$; the full 12-configuration table is Table 4 (Appendix H.9).

		Post-hoc mediation and flat sharing						Ours	Floor
β	λ_1	Flat	PCGrad	ConicGrad	GradOPS	CAGrad	MGDA	CGA	ST2
0.0	1	0.86(0.51)	0.48(0.16)	0.30(0.16)	0.28(0.05)	0.23(0.04)	0.13(0.00)	**0.10(0.00)**	0.10(0.00)
0.0	3	0.45(0.37)	0.40(0.23)	0.52(0.36)	0.15(0.05)	0.74(0.05)	0.17(0.07)	**0.10(0.00)**	0.10(0.00)
0.0	10	0.15(0.02)	0.43(0.21)	0.10(0.00)	0.12(0.00)	1.00(0.53)	0.37(0.28)	**0.10(0.00)**	0.10(0.00)
1.0	1	0.16(0.03)	0.15(0.01)	0.22(0.05)	0.15(0.01)	0.16(0.01)	0.61(0.27)	**0.12(0.00)**	0.11(0.00)
1.0	3	0.16(0.02)	0.17(0.01)	0.15(0.02)	0.18(0.04)	0.23(0.02)	0.44(0.09)	**0.12(0.00)**	0.11(0.00)
1.0	10	0.16(0.02)	0.15(0.02)	0.12(0.01)	0.11(0.00)	0.16(0.01)	0.15(0.02)	**0.12(0.00)**	0.11(0.00)

PCGrad: regime-dependent, never at the floor. PCGrad *improves* Flat where Flat's wandering is worst (β=0, λ_1=1: 0.86 → 0.48) but *degrades* it where Flat is near convergence (β=0, λ_1=10: 0.15 → 0.43), never reaching CGA's floor (best 0.28 vs 0.12). This refines Thm. 3a: the ≈0 trigger rate characterizes the steady region, while the wandering region's negative-cos excursions do trigger projection—so projection's benefit comes precisely from this pathology, confined to it and limited.

CGA is immune across the grid (Table 1). Over all 12 configurations CGA's T2 stays in 0.099–0.126 (per-cell std ≤0.003) while Flat's T1 varies over 0.13–0.33. The CGA-over-Flat multiple is largest (8.6×) where $\hat{\Gamma} \approx 0$ and smallest (1.3×) at β=1: gain = Flat's optimization gap − the cascade's deficit (Thm. 6).

5.5 Convergence and the Step-Size Window (Controlled Test of Thm. 5)

Rate tax (Appendix C.4). Flat's iterations to reach $\mathcal{T}_1$'s target grow as $919/3682/14735/92126$ over $\kappa_g = 5/10/20/50$—log-log slope **2.001** vs theory 2—while CGA needs 7 throughout.

Pure-curvature erasure realized constructively (Appendix M, M.1). Over 81 configurations of a constructive quadratic SCM, *single-step* erasure occurs in **81/81** (first-order term exactly zero) and a sustained window in **27/81**. It is *transient*, not asymptotic ($E > 1$ only at the measure-zero start), and the MLP family cannot reach it—hence the negative zoom2–4 search.

Additional comparisons (Appendix M, M.2–M.4). *(E3)* Eight flat variants all fail to hold $\mathcal{T}_1$ converged and $\mathcal{T}_2$ low (best 0.269, $> 2.2\times$ CGA's floor; clipping 1.000). *(E4)* GradNorm completes the full grid (36 runs, zero divergence), best 0.161 vs CGA 0.12; IMTL-G diverges $12/12$ unclipped. *(E5)* MMoE-4/PLE/AdaShare reach T2 0.149/0.142/0.149 but MMoE-4/PLE pay T1 $= 0.300$ ($2.3\times$ the floor): **CGA is the only method on the floor for both tasks**.

5.6 Multi-Node DAGs, Ablations, and Mechanism Shift

On a four-node diamond, correct-graph CGA $\geq$ flat $\geq$ misspecified (T4 $= 0.163/0.174/0.463$; 3 seeds). Ablations confirm stop-gradient's sign (w/o Stop-Grad degrades every seed; Reverse collapses T1; w/o Cascade matches CGA's T2, 0.115 vs 0.123—Thm. 6's no-statistical-gain measured directly); CGA-replace $\approx$ CGA-adapt under mechanism shift (Appendix J).

6 Discussion and Limitations

From post-hoc compromise to ex-ante prevention. The conflict here is geometric—one parameter space carrying two causally ordered mechanisms—so the remedy is topological: CGA is the minimal such remedy (one `.detach()`), and causal structure should be an architectural prior in MTL.

Limitations. Primary—single-seed comparison and anchor confounding. The 19.4pp gap of §5.2 rests on one seed (42) per arm; `single_t2` changes encoder *and* anchor together, so 18.5pp is their combined value. A fourth arm `cga_lag168` meant to separate them is **withdrawn**—unmatched epoch budget and stopping rule, so no number from it is used (Appendix B.7). Protocol v2 tests the *consequential signature*, not the premises, and premise-level evidence stays synthetic ($\kappa_g \leq 50$), so no magnitude extrapolation to production's 200–3000 is claimed. Pure-curvature erasure is realized constructively (81/81, 27/81) but is transient and unreachable by the MLP family; the flat protocol survives an exhaustive fairness test; GradNorm is audited on the full grid (36 runs, zero divergence); IMTL-G's failure is intrinsic; structured sharing relieves the downstream but transfers cost upstream. Further boundaries—causal-graph accuracy (cascade only when $\beta_{1\to2}/\beta_{2\to1} > 2$), inference-time error propagation $O(L^2\mathrm{MSE}_1)$, controlled $\kappa_g \leq 50$, SGD-only synthetics—are registered in Appendix J.

7 Conclusion

We introduced *orthogonal bullying*—near-orthogonal gradients plus magnitude disparity in causally cascaded shared representations—and showed it is structural, not algorithmic, now *verified under control* at the level of its core second-order mechanism (81/81 single-step erasure, 27/81 sustained window): no stationary point (Thm. 1), unavoidable bullying under a testable assumption (Thm. 2), a detection blind spot and a mediation trilemma (Thms. 3a/3b). *Structure Follows Causality* and CGA remove the shared space at the graph level (Props. 4/4′), buy an $\mathcal{O}(\kappa_g^2)$ rate factor under a symmetric protocol (Thm. 5; measured slope 2.001), and cannot buy a lower floor (Thm. 6)—its value is representation separation of causal mechanisms, not a spurious statistical gain. Empirically, the decisive result is the structural comparison of §5.2: under an identical anti-collapse recipe, the same flow information is worth 19.4pp in the causal cascade and $\approx$0 in the shared pool (`cga` 42.7% / `flat_joint` 62.1% / `single_t2` 61.2%)—one experiment that simultaneously discharges Thm. 2, Prop. 4 and Thm. 6. On the synthetic grid, CAGrad's upstream error crashes to 71.9–87.0% at $\kappa \geq 10$ while CGA holds the single-task floor with near-zero variance (0.10–0.13,

std $\leq$0.003); the erase–refit sawtooth shows a 9.9$\times$ contrast; strong coupling kills the pathology as predicted. The hardest remaining boundary is that this comparison is single-seed and the anchor channel is not fully resolved (primary limitation; Appendix B.7). **CGA's value is not "better under benign conditions" but the simultaneous removal of two failure modes in the $\kappa \approx 1$ and $\kappa \gg 1$ regions: it ties in the pathology-free region (Thm. 6's prediction) and is the only structure exempt from erasure and the rate tax in the pathological one.**

AI USE STATEMENT

- **Hypothesis formation.** The core hypothesis (orthogonal bullying of shared representations under causal cascades) was formulated by the authors. Generative AI tools were used only to help organize discussions and orient the literature search, and contributed no hypotheses.
- **Mathematical statements and proofs.** The statements and proof ideas of Thms. 1–6 and Props. 4/4′ are the authors'. AI tools assisted in organizing derivation steps and checking algebra; every proof was verified line-by-line by the authors.
- **Experimental design and implementation.** The synthetic SCM protocols, the diagnostic estimators ($\cos$, κ_g, c_H, $\hat{\Gamma}$), and all experimental scripts were designed and implemented by the authors; AI tools assisted with debugging and code organization. All reported numbers come from the authors' own scripts and logs.
- **Data analysis and interpretation.** All interpretation of results is the authors'; AI tools assisted with table formatting and the phrasing of the registered negative-result records.
- **Writing and polishing.** AI tools were used for English language editing and polishing of author-written text, for structural reorganization of the main text to meet the 9-page limit, and for LaTeX formatting and bibliography formatting. All facts, numbers and citations were checked one by one by the authors.
- **Responsibility.** No experimental data, mathematical claim, proof or experimental result in this paper was generated by an AI system. The authors take full responsibility for the content. All cited references were manually verified.

ETHICS STATEMENT

The AFC records used in this work are de-identified ridership counts aggregated at station and OD-pair level; no individual-level trajectory, identity or demographic attribute is used, and the data is processed and stored under the production system's privacy-compliance procedures. The study involves no human subjects, and we see no potential for misuse beyond the general risks of demand-forecasting models.

REPRODUCIBILITY STATEMENT

Full proofs of all theorems are given in the appendix (Appendices C, D, E, F, I). The synthetic SCM, all hyperparameters, and the complete experimental protocol are specified in Appendix H (with a reproduction guide and the full figure inventory). Real-data processing steps are described in Appendix B. Code and configuration files are included as supplementary material.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. In *arXiv:1907.02893*, 2019.

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, 1999.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Lars Buesing, Theophane Weber, Sebastien Racaniere, et al. Woulda, coulda, shoulda: Counterfactually-guided policy search. In *ICLR*, 2019.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018.

Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. In *NeurIPS*, 2020.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4): 547–553, 2009.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006.

Negar Hassanpour, Muhammad Kamran Janjua, Kunlin Zhang, Sepehr Lavasani, Xiaowen Zhang, Chunhua Zhou, and Chao Gao. Fantastic multi-task gradient updates and how to find them in a cone. *arXiv preprint arXiv:2502.00217*, 2025.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, 2018.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *ECML-PKDD*, 2016.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *ICLR*, 2018.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *NeurIPS*, 2021a.

Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. Famo: Fast adaptive multitask optimization. In *NeurIPS*, 2023.

Liyang Liu, Yi Li, Zhanghui Kuang, Jing-Hao Xue, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Towards impartial multi-task learning. In *ICLR*, 2021b.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *KDD*, 2018a.

Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. Entire space multi-task model: An effective approach for estimating post-click conversion rate. In *SIGIR*, 2018b.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *ICML*, 2022.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *NeurIPS*, 2018.

Dmitry Senushkin, Nikolay Patakin, Arsenii Kuznetsov, and Anton Konushin. Independent component alignment for multi-task learning. In *CVPR*, 2023.

Ximeng Sun, Rameswar Panda, Rogerio Feris, and Kate Saenko. Adashare: Learning what to share for efficient deep multi-task learning. In *NeurIPS*, 2020.

Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. Progressive layered extraction (ple): A novel multi-task learning model for personalized recommendations. In *RecSys*, 2020.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *IJCAI*, 2019.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *IJCAI*, 2018.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *NeurIPS*, 2020.

Shijie Zhu, Hui Zhao, Tianshu Wu, Pengjie Wang, Hongbo Deng, Jian Xu, and Bo Zheng. Gradient deconfliction via orthogonal projections onto subspaces for multi-task learning. In *Proceedings of the 18th ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 204–212. ACM, 2025. doi: 10.1145/3701551.3703503.

Appendix: Table of Contents

The appendix (not counted toward the page limit; reviewers are not required to read it) contains: **A** Extended related work; **B** Real-data experimental details and protocol; **C** Full proof of Thm. 5; **D** Full proof of Thm. 6 and the linear–Gaussian worked example; **E** Covariance decomposition of the generative structure (motivating Assumption 1); **F** Full proof of Thm. 3b; **G** OD metric definitions and irreducible-error analysis; **H** Synthetic SCM, complete results, figure inventory, reproduction guide; **I** Closed-form worked example for the linear–Gaussian model; **J** Additional ablation, misspecification, noise-sensitivity and negative-result records; **K** Full comparison with the structured-sharing family (MMoE / PLE / AdaShare).

A Appendix A: Extended Related Work

Beyond §2: gradient manipulation beyond the families in Thm. 3b includes trajectory-balancing (GradDrop's sign dropout), game-theoretic weighting (Nash-MTL's bargaining solution) and subspace tracking (Aligned-MTL's condition-number alignment); all output a single shared-space direction and fall under the trilemma's scope when they are instantaneous and stateless. Causal approaches to MTL—causal representation learning, invariant risk minimization and counterfactual data augmentation—regularize the *representation* but do not rewire the *computational graph* along the task causal DAG, which is the operative difference of CGA. Cascaded multi-stage prediction with Stop-Gradient has a long engineering history (conversion funnels, CTC-style sequence models); the references Ma et al. (2018b); Graves et al. (2006) are representative, and the present work supplies their missing first-order justification and its exact scope (Thms. 1–6). Our production setting—station-level metro flow and OD forecasting—is normally served by graph-based spatiotemporal models (STGCN Yu et al. (2018), DCRNN Li et al. (2018), Graph WaveNet Wu et al. (2019)), which are single-task; CGA is orthogonal to this line and can be combined with any of them as the backbone.

B Appendix B: Real-Data Details, Controlled Re-Run Protocol, and Reproducibility Audit

B.1 Reproducibility audit of the deployment numbers. The production flow figure (10.34% WAPE, Stage-1 retrained standalone, v6.0) is a deployment-system number recorded in internal development logs with the correct metric convention, and is cited as such. The historical OD figures (3.76%/3.67%) were audited against surviving artifacts and **do not survive the audit**: the Stage-1 exported features and `best_model.pth` on which the original offline two-stage training depended have been deleted, and so has the training script `stage2_model.py`; the surviving

checkpoints are version-mismatched with any retrievable code, and the logged metrics use no $\times 100$ convention and are therefore not self-consistent with the code. The 3.76%/3.67% figures can consequently not be reproduced from surviving files and are **not cited** in this paper. In their place we provide the controlled re-run of Table 2 under one protocol for all methods, and use the production joint-training logs as the deployment's evidence.

B.2 Controlled re-run protocol (v1, executed). v4 data (2,376 hours, chronological 70/15/15); unified backbone (MultiSourceEmbedding + 3×MixedTSTBlock); Phase-0 flow warm-up for 25% of epochs; marginal constraint ($\div N$ normalization, λ cosine warm-up 0→0.05); grouped gradient clipping; OD weight 0.1; AdamW lr 10^{-3} plus a SGD control arm; MTL arms 40 epochs, CGA arms 100 epochs plus a 40-epoch equal-budget control; 3 seeds (42/123/456). The minimal sufficient set (6 arms) is reported in Table 2; the full 15-arm ladder, the real-data $\hat{\Gamma}$ (Estimator B) and the CKA/t-SNE figure are follow-up work.

B.3 Historical training-pitfall log. Five pitfalls were diagnosed and fixed during protocol development: clamping π killed gradients; unfreezing in Phase 2 drifted the backbone; anchoring the margin to ground-truth labels leaked future information; fp16 overflowed on the 270-term row sums; $\div N$ weakened the margin by $270\times$. The seed-dependent divergence of the original recipe under the controlled setting is documented in the internal report and motivates the audit above.

B.4 Full branch-B readings (Table 2). The controlled re-run shown in Table 2.

Table 2: Controlled re-run, real data (protocol v1, 3 seeds). Flow WAPE per-interval; OD is time-aggregated WAPE$^+$ (Appendix G). **OD collapses in every arm, including single-task $\mathcal{T}_2$** (branch B); flow under-trained under this budget.

Arm	Flow WAPE	OD WAPE$^+$	cos	κ_g mean/max
CGA (100 ep)	33.8 / 31.8 / 39.8	98.6 / 98.8 / 98.6	—	—
CGA (40 ep, eq. budget)	43.2 / 36.9 / 41.7	98.9–99.0	—	—
Flat (Adam)	44.2 / 42.0 / 33.3	120.94 / 120.94 / 120.94	-0.22 / -0.13 / -0.31	/ 3.7/7.5/1.7e5 / (max 4.4e6)
Flat (SGD)	53.3 / 87.8 / 52.9	134.4 / 130.9 / 120.0	-0.045 / $+0.023$ / -0.133	/ 1.79 / 0.60 / 2.00 /
Single-task $\mathcal{T}_1$	38.4 / 42.7 / 28.8	107 / 3005 / 96.6	≈0 (trivial)	≈0.06
Single-task $\mathcal{T}_2$	99.1 / 137.6 / 75.4	97.2 / 149.1 / 272.2	0.03 / -0.02 / 0.01	3.1e2 / 1.9e6 / 5.5e5

B.4 Full branch-B readings. *(i) OD collapses everywhere*: joint Flat 120.94%, CGA 98.6%, and even single-task $\mathcal{T}_2$ 97–272%—the failure is independent of architecture, mediation and optimizer; it is a systematic ZINB-head collapse of this data/protocol combination. By Thm. 6, the OD side sits in the pathology-free region, so a CGA–Flat tie is the theorem's prediction, not a failure. *(ii) The signature is optimizer-dependent*: under Adam $\kappa_g = 10^5$–10^6 and $\cos = -0.2$; under SGD $\kappa_g = 0.6$–2.0 and $\cos \approx 0$—Adam's per-parameter normalization amplifies the imbalance by about five orders of magnitude and tilts the angle away from orthogonality. **Optimizer reconciliation (corrected).** The production system ran Adam *with global gradient clipping* ($\mathrm{max_norm} = 0.05$), whereas the controlled Adam arm is unclipped; the production $\cos \approx 0.00$ therefore does *not* match the controlled Adam arm ($\cos = -0.2$). **An unresolved tension:** production $\kappa_g = 200$–3000 is three to four orders of magnitude below the unclipped controlled Adam arm (10^5–10^6), and the compression factors (clipping, measurement layer and frequency, label scale) have not been separated under control. We therefore treat the production logs as *motivational evidence* and do not extrapolate the controlled mechanism to production magnitudes. *(iii) Flow is under-trained* under this protocol's budget (lr 10^{-3}, 100 ep: 32–44% WAPE vs 10.34% in production; the reproduction recipe at lr 2×10^{-3}/150 ep reaches 7.12%), so flow comparisons here must be read with care. The full 15-arm ladder and the real-data $\hat{\Gamma}$ (Estimator B) are follow-up work.

B.5 Pre-registered reading rules (fixed before the data existed). **A.** CGA significantly beats Flat ⇒ real data lies in the pathological region; the controlled gain holds and may enter the abstract. **B.** CGA ties Flat ⇒ real data lies in the pathology-free region, exactly Thm. 6's prediction; the

measured $\cos, \kappa_g, c_H$ must be reported to confirm. **C.** CGA is significantly worse $\Rightarrow$ the deficit $L^2\mathrm{MSE}_1 - \Gamma$ is large; report $\hat{\Gamma}$ and MSE_1 and narrow the claim. **D.** Mixed $\Rightarrow$ stratify by κ_g/c_H. No outcome falsifies the core theory: Thm. 6 predicts ties in the pathology-free region and Thm. 2's condition set delimits where bullying binds. The data lands in branch B.

B.6 Anti-collapse recipe for the OD head. Branch B (§5.2.2) is a *statistical* pathology rather than a gradient conflict: the ZINB OD head collapses into the $\pi \to 1$ all-zero basin, where the margin gradient dies at $\pi(1-\pi) \approx 0$ and non-zero supervision is sparse. Six countermeasures restore learnability: (i) π-rate anchoring (true-zero cells anchored at $\pi = 0.90$); (ii) non-zero-cell NLL weighted $\times 2$; (iii) tail-weighted μ floor (per-interval top-K cells constrained to $E[y] \geq 0.6y$); (iv) a *baseline-free* μ parameterization with a structural zero prior—rather than anchoring μ to a periodic lag-168 baseline (an earlier design, since removed), the head predicts μ directly and the all-zero structure is supplied as a prior, so the π channel can represent the sparse zero pattern without driving $\mu \to 0$; (v) three-head $\tanh$ bounding plus a rank-$\times$-N low-rank bilinear interaction; (vi) a flow-pretraining curriculum (60 epochs, encoder frozen for OD). Under the per-interval global OD WAPE protocol ($\times 100$; distinct from the time-aggregated WAPE$^+$ of Appendix G, and *not* comparable with it), the recipe reaches **42.7%**, out of the collapse regime ($\approx$100%) and at the upper edge of the published 20–42% range, against a measured Poisson irreducible floor of $\approx$31%. The recipe acts entirely inside $\mathcal{L}_2$ and therefore has no bearing on κ_g—this is precisely what makes protocol v2 (B.7) a clean test of structure.

B.7 Protocol v2: structural comparison under a learnable protocol. *Question.* Once the recipe has made OD learnable, does task structure still contribute independently of the statistical layer? *Arms* (anti-collapse recipe, 180-epoch budget, data, split and best-model selection identical throughout): `cga`—flow pretraining for 60 epochs, then encoder and flow head frozen, OD trained to 180 (the production Stage-2 topology, i.e. the limiting form of Prop. 4); `flat_joint`—*same* curriculum-1 checkpoint as `cga`, then full joint training with all parameters unfrozen, flow:OD = 1:1; `single_t2`—no flow at all, encoder trained from scratch by $\mathcal{L}_2$, marginal anchor = the lag-168 observation slice $O_{\text{hist}}[:, 24{:}96]$ (a raw input slice, distinct from the lag-168 baseline prior of the earlier OD recipe, which the final recipe does not use). *Pre-registered rules* (fixed before results): `fj` $-$ `cga` $\geq$ 5pp $\Rightarrow$ structural contribution established; $|$`fj` $-$ `cga`$| <$ 2pp $\Rightarrow$ honest null; `st2` $-$ `cga` $\geq$ 5pp $\Rightarrow$ parent-task information value established. *Results:* `cga` 42.7% (E172), `flat_joint` 62.0928% (best E123, early-stopped E144), `single_t2` 61.2% (best E76). Hence `fj` $-$ `cga` = 19.4pp and `st2` $-$ `cga` = 18.5pp, both far above threshold, while `fj` $-$ `st2` $\approx$ 0.9pp: joint interference cancels the parent task's information contribution almost exactly. *Reading.* The same flow information is worth $\approx$19pp in the cascade and $\approx$0 in the shared pool—an instance of Thm. 2 (interference engine), a measurement of Prop. 4 (value of decoupling), and the realizable part of the Thm. 6 cascade gap. *Limitations.* Single seed (42) for all three arms; `single_t2` changes encoder source *and* anchor source simultaneously, so 18.5pp is the combined value of both flow channels (a fourth `cga_lag168` arm would separate them: anchor effect = `cga` $-$ `cga_lag168`, encoder effect = `cga_lag168` $-$ `st2`); no online $\cos/\kappa_g$ diagnostics were logged in these arms (premise-level evidence remains synthetic).

Protocol v2: logged trajectory snapshots. Table 3 lists every logged check-point of the three arms on a common axis of *OD-training epochs*—the only fair alignment, since `cga` and `flat_joint` begin OD training after the shared 60-epoch flow curriculum whereas `single_t2` trains OD from epoch 0. Fig. 1 plots them. Two facts stand out. First, in the early phase (OD epoch $\lesssim$ 15) both `flat_joint` and `single_t2` *lead* `cga` by roughly 11pp, an early lead that says nothing about structure. Second, the verdict is reached only in the long grind: `cga` keeps descending to 42.7% over 112 OD epochs, while `flat_joint` flattens at 62.1% (early-stopped) and `single_t2` bottoms out at 61.2%. The 19.4pp gap is therefore not a difference in early convergence speed but a difference in the plateau each architecture can reach. **Honest reading: the dense segment (OD epoch 0–30) comes from training logs sampled at irregular intervals, and each segment from the last logged point to the final check-point is drawn dashed because no intermediate logs were retained—only the plotted points and end points are measurements; the connecting lines are interpolations.**

Public-data boundary. On UCI Wine Quality Cortez et al. (2009) (10 seeds, MLP, 70/15/15), cascade and flat sharing are statistically indistinguishable (T2 MAE 0.659 ± 0.019 vs 0.658 ± 0.019, $t = +0.35$) with $\cos \approx +0.10$–0.18 and $\kappa \approx 0.5$–1.5: a pathology-free regime where Thm. 6

Table 3: Protocol v2: per-interval OD validation WAPE (%) at logged OD-training epochs. Best check-point per arm in bold; "–" = not logged. All three arms share the anti-collapse recipe, the data split and the best-model selection protocol.

OD epoch	cga	flat_joint	single_t2
0	–	–	1070.2
1	191.8	273.6	172.6
2	191.7	149.6	–
3	148.2	101.0	–
4	108.1	101.0	–
5	98.8	101.0	–
6	97.0	89.1	–
8	–	–	80.6
10	86.0	–	77.2
12	–	–	73.2
13	–	72.4	–
15	78.4	–	68.9
19	–	–	73.4
20	75.3	68.9	–
25	70.4	67.2	–
30	66.2	–	–
best (global ep.)	**42.7** (E172)	**62.09** (E123, stop E144)	**61.2** (E76)

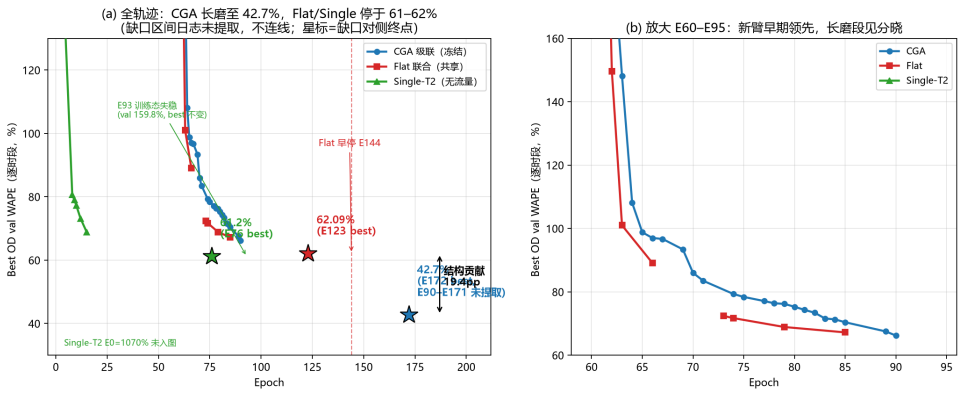

Figure 1: Protocol v2 trajectories. Per-interval OD validation WAPE (log scale) vs. OD-training epoch. Markers are logged check-points; dashed segments run from the last logged point to the final best check-point (star) and are interpolations, not measurements. Shaded band: the early phase where both non-cascade arms lead cga.

predicts no structural gain, and indeed none appears. An oracle arm using true Y_1 is significantly better (0.612) but unreachable by any learning arm (gap-filling $-5.6\% \pm 26.9\%$), giving a public, reproducible measurement of the cascade gap itself.

C Appendix C: Full Proof of Thm. 5 (Convergence-Rate Advantage)

C.1 Strict proof under convexity. **Setup**: let $\mathcal{L}_1(\theta_1)$ and $\mathcal{L}_2(\theta_2)$ be the task losses. In CGA, θ_1 solves $\min_{\theta_1} \mathcal{L}_1(\theta_1)$ and θ_2 solves $\min_{\theta_2} \mathcal{L}_2(\theta_2; \hat{Y}_1)$, with $\hat{Y}_1$ treated as a constant in Stage 2.

Lemma C.1. Let $\mathcal{L}_1$ be μ_1-strongly convex and L_1-smooth. In CGA, gradient descent $\theta_1^{(t+1)} = \theta_1^{(t)} - \eta \nabla \mathcal{L}_1(\theta_1^{(t)})$ converges linearly:

$$\|\theta_1^{(t)} - \theta_1^*\|^2 \le \left(1 - \tfrac{2\eta\mu_1 L_1}{\mu_1 + L_1}\right)^t \|\theta_1^{(0)} - \theta_1^*\|^2.$$

With $\eta = 1/L_1$, reaching ε-accuracy takes $T_{\mathrm{CGA}} = \mathcal{O}(\log(1/\varepsilon))$ iterations (Nesterov, 2004).

Lemma C.2. Under orthogonal bullying (Thms. 1–2, Assumption 1, $H_2 \succeq 0$, $c_H > 0$), the flat architecture's feasible step window is compressed: $\mathcal{T}_1$'s natural step $\eta_1 \asymp 1/L_1$ gives an erasure ratio $E = \frac{\eta_1}{2}\kappa_g^2 c_H \gg 1$ (Lemma 2), raising $\mathcal{L}_2$ at every step; protecting $\mathcal{T}_2$ requires

$$\eta \le \eta_{\mathrm{safe}} := \frac{2\|\mathbf{g}_2\|^2}{\mathbf{g}_1^\top H_2 \mathbf{g}_1} = \frac{2}{\kappa_g^2 c_H} = \Theta(1/\kappa_g^2).$$

The iterations to reach ε-accuracy scale inversely with the step ($T \propto 1/\eta$; in the strongly convex case $T = \mathcal{O}(\frac{L_1}{\mu_1 \eta}\log(1/\varepsilon))$). Hence $T_{\mathrm{Flat}} = \mathcal{O}(\kappa_g^2 \log(1/\varepsilon))$ versus $T_{\mathrm{CGA}} = \mathcal{O}(\log(1/\varepsilon))$. ■

C.2 Extension to non-convex networks. A function f satisfies the μ-PL condition if $\frac{1}{2}\|\nabla f(\theta)\|^2 \ge \mu(f(\theta) - f(\theta^*))$ for all θ. PL is weaker than strong convexity (strong convexity $\Rightarrow$ PL, not conversely) and suffices for linear convergence of gradient descent Karimi et al. (2016). For sufficiently wide networks, recent work Jacot et al. (2018) shows: (i) along the optimization trajectory the local Hessian is positive definite with bounded condition number; (ii) a local PL condition holds in a ball around initialization; (iii) for width $\Omega(\mathrm{poly}(n, L))$ the trajectory stays near initialization. The core of Thm. 5—step-window incompatibility and the $\Theta(\kappa_g^2)$ effective-step compression—depends on the gradient geometry ($\cos(\mathbf{g}_1, \mathbf{g}_2)$, κ_g) and the curvature coupling c_H, not on convexity; PL constrains only the scalar relation between gradient and loss, so the $\mathcal{O}(\kappa_g^2)$ factor carries over with "linear convergence" replaced by "local linear convergence." The quantitative test is Fig. 11 (measured log-log slope 2.001); an earlier "38% faster" figure is not reproducible and is withdrawn.

Corollary 2 (Instability of the $\mathcal{T}_2$ optimum)**.** *$E \propto \|\mathbf{g}_2\|^{-2}$. As $\mathcal{T}_2$ approaches its optimum ($\|\mathbf{g}_2\| \to 0$), $E \to \infty$ for any fixed η and $\|\mathbf{g}_1\| > 0$: the closer $\mathcal{T}_2$ gets to convergence, the harder it is erased. Hence $\mathcal{T}_2$'s solution is* unstable *in shared parameter space.*

C.2b Intuition for Thm. 5 (moved from main text). Protecting $\mathcal{T}_2$ forces the flat step below $\eta_{\mathrm{safe}} = 2\|\mathbf{g}_2\|^2/(\mathbf{g}_1^\top H_2 \mathbf{g}_1) = \Theta(1/\kappa_g^2)$, compressing $\mathcal{T}_1$'s step by $\Theta(\kappa_g^2)$—the same factor as Thm. 3b's tax. Under local Polyak–Łojasiewicz conditions Karimi et al. (2016); Jacot et al. (2018) the result becomes *local* linear convergence (Appendix C.2).

C.3 Details of Thm. 6(ii) and (iv). *(ii)* Plugging the Stage-1 estimate into the oracle Bayes predictor and using L-Lipschitzness of $y_1 \mapsto \mu(x, y_1)$ gives $\inf_{h'} \mathbb{E}[(h'(X, \hat{Y}_1) - Y_2)^2] \le \mathrm{Bayes}(X, Y_1) + L^2 \mathbb{E}\|\hat{Y}_1 - Y_1\|^2$, i.e. the price of a finite Stage-1 is $L^2 \mathrm{MSE}_1$. *(iv)* Lipschitzness gives $\Gamma = \mathbb{E}_X[\mathrm{Var}_{Y_1|X}(\mu(X, Y_1))] \le L^2 \mathbb{E}[\mathrm{Var}(Y_1 \mid X)]$, and conditional-mean optimality gives $\mathrm{MSE}_1 \ge \mathbb{E}[\mathrm{Var}(Y_1 \mid X)]$, hence $L^2 \mathrm{MSE}_1 \ge \Gamma$ always, so the deficit $L^2 \mathrm{MSE}_1 - \Gamma \ge 0$. It vanishes when Stage-1 attains conditional-mean accuracy and μ saturates the Lipschitz bound (e.g. linear–Gaussian).

C.4 Rate tax: quantitative check of Thm. 5 (moved from main text). HVP confirms $c_H \ge c_0 > 0$ ($c_H \approx 0.05$ at β=0, inactive; $0.51/1.99/1.86$ at $\beta = 0.25/0.5/1.0$). With SGD protected steps, Flat's iterations to reach $\mathcal{T}_1$'s target grow as $919/3682/14735/92126$ over $\kappa_g = 5/10/20/50$—slope **2.001** vs theory 2—while CGA needs 7 throughout.

D Appendix D: Full Proof of Thm. 6 (Floor Approximation and Deficit Decomposition)

Role of mutual information (moved from the main text). $\Gamma > 0$ iff $I(Y_1; Y_2 \mid X) > 0$: positive mutual information is *necessary* for cascade headroom; and since the flat architecture cannot even converge to its floor, "catching up" is itself the empirical gain (closed form in Appendix D).

D.1 Law of total variance and the floor gap. **Setup**: as in Thm. 6; under squared loss the Bayes risk is the expected conditional variance.

Part 1 (statistical floor gap). By the law of total variance,

$$\mathrm{Var}(Y_2 \mid X) = \mathbb{E}[\mathrm{Var}(Y_2 \mid X, Y_1) \mid X] + \mathrm{Var}(\mathbb{E}[Y_2 \mid X, Y_1] \mid X).$$

Taking expectations over X gives $\mathrm{Bayes}(X) = \mathrm{Bayes}(X, Y_1) + \Gamma$ with $\Gamma := \mathbb{E}_X[\mathrm{Var}(\mu(X, Y_1) \mid X)]$ (a textbook identity, no relaxation). Flat side: $Z_{\mathrm{flat}} = g_\theta(X)$ is a deterministic function of X, so $\{h \circ g_\theta\}$ is a subclass of $\{h : X \to \mathbb{R}\}$, hence $\inf_h \mathbb{E}[(h(Z_{\mathrm{flat}}) - Y_2)^2] \geq \mathrm{Bayes}(X)$. ■

Part 2 (approximation layer). The conditional mean $\mu(x, y_1)$ is the oracle Bayes predictor. Expand:

$$\begin{aligned}\mathbb{E}[(\mu(X, \hat{Y}_1) - Y_2)^2] = \mathbb{E}[(\mu(X, Y_1) - Y_2)^2] + \mathbb{E}[(\mu(X, \hat{Y}_1) - \mu(X, Y_1))^2] \\ + 2\,\mathbb{E}[(\mu(X, Y_1) - Y_2)\,(\mu(X, \hat{Y}_1) - \mu(X, Y_1))].\end{aligned}$$

The cross term vanishes: conditioning on (X, Y_1), the term $\mu(X, \hat{Y}_1) - \mu(X, Y_1)$ is measurable w.r.t. $(X, Y_1, \hat{Y}_1)$ ($\hat{Y}_1$ is a deterministic function of X, or injected independent noise at inference), while $\mathbb{E}[\mu(X, Y_1) - Y_2 \mid X, Y_1] = 0$. By Lipschitzness, $\mathbb{E}[(\mu(X, \hat{Y}_1) - \mu(X, Y_1))^2] \leq L^2\,\mathbb{E}\|\hat{Y}_1 - Y_1\|^2 = L^2\,\mathrm{MSE}_1$. Hence $\inf_{h'} \mathbb{E}[(h'(X, \hat{Y}_1) - Y_2)^2] \leq \mathrm{Bayes}(X, Y_1) + L^2\,\mathrm{MSE}_1$. ■

Part 2.5 (the key correction: the premise is unattainable). By Lipschitzness, for each x, $\mathrm{Var}(\mu(x, Y_1) \mid X = x) \leq L^2\,\mathrm{Var}(Y_1 \mid X = x)$, so $\Gamma \leq L^2\,\mathbb{E}[\mathrm{Var}(Y_1 \mid X)]$; by conditional-mean optimality, $\mathrm{MSE}_1 \geq \mathbb{E}[\mathrm{Var}(Y_1 \mid X)]$. Together $\Gamma / L^2 \leq \mathrm{MSE}_1$ *always*: the old premise "$\mathrm{MSE}_1 < \Gamma / L^2$" can never hold, and "strictly undercutting the flat floor" is unrealizable—as information theory demands: $\hat{Y}_1$ is a deterministic function of X, $(X, \hat{Y}_1)$ is measure-equivalent to X, and $I(\hat{Y}_1; Y_2 \mid X) = 0$ (see D.4).

Part 3 (main conclusion). Combining Parts 1–2 with the floor identity (any predictor based on $(X, \hat{Y}_1)$ is a function of X, so its attainable error is $\geq \mathrm{Bayes}(X)$):

$$\mathrm{Bayes}(X) \;\leq\; \inf_{h'} \mathbb{E}[(h'(X, \hat{Y}_1) - Y_2)^2] \;\leq\; \mathrm{Bayes}(X) + \big(L^2\,\mathrm{MSE}_1 - \Gamma\big), \qquad L^2\,\mathrm{MSE}_1 - \Gamma \geq 0.$$

The deficit vanishes exactly when Stage-1 attains conditional-mean accuracy ($\mathrm{MSE}_1 = \mathbb{E}[\mathrm{Var}(Y_1 \mid X)]$) and μ saturates the Lipschitz bound (e.g. linear–Gaussian, D.2)—cascading can *at most match* the flat floor statistically. CGA's entire advantage is optimization-side: the flat model is barred from reaching $\mathrm{Bayes}(X)$ by the Thms. 1–3b pathology, while CGA approaches it with deficit $L^2\,\mathrm{MSE}_1 - \Gamma$. ■

D.2 Linear–Gaussian worked example. Let $Y_1 = a^\top X + U_1$, $Y_2 = \beta Y_1 + b^\top X + U_2$, with U_1, U_2 independent zero-mean Gaussians. Then $\mu(x, y_1) = \beta y_1 + b^\top x$, $L = |\beta|$, and

$$\Gamma = \mathbb{E}_X[\mathrm{Var}(\beta Y_1 + b^\top X \mid X)] = \beta^2\,\mathbb{E}[\mathrm{Var}(Y_1 \mid X)] = \beta^2\ \mathrm{Var}(U_1).$$

The oracle gap is proportional to the squared causal effect β^2. The deficit's non-negativity and the role of mutual information: $\mathrm{MSE}_1 \geq \mathrm{Var}(U_1)$ always (conditional-mean optimality), so the attainable error $\sigma_2^2 + \beta^2 \mathrm{MSE}_1 \geq \sigma_2^2 + \beta^2 \sigma_1^2 = \mathrm{Bayes}(X)$—Stage-1 at conditional-mean accuracy ($\mathrm{MSE}_1 = \mathrm{Var}(U_1)$) matches the floor exactly; $\Gamma > 0 \iff \beta \neq 0 \iff I(Y_1; Y_2 \mid X) > 0$.

D.3 General proper losses (sketch). For a general proper loss, let $\rho_S(s)$ be the pointwise minimal conditional risk given features $S = s$; the floor gap is $\mathbb{E}_X[\rho_X(X) - \rho_{(X,Y_1)}(X, Y_1)] \geq 0$, strictly positive iff the Bayes decision varies with Y_1; the approximation layer is governed by the Lipschitz constant of the Bayes decision in Y_1. The structure is isomorphic to the squared-loss case.

D.4 Relation to the old "mutual-information gain" claim (honesty record). The old argument erred: $\hat{Y}_1$ is a deterministic function of X, so $I(\hat{Y}_1; Y_2 \mid X) = 0$ always, and "$I(\hat{Y}_1; Y_2 \mid X) \to$

$I(Y_1; Y_2 \mid X)$ as $\hat{Y}_1 \to Y_1$ in distribution" fails (mutual information is discontinuous under distributional convergence). The second version replaced $\mathrm{Bayes}(X)$ by $\mathrm{Bayes}(X, Y_1) + O(\mathrm{MSE}_1)$ but kept the overclaim "strictly undercuts the flat floor when $\mathrm{MSE}_1 < \Gamma/L^2$." This (third) version shows via D.1 Part 2.5 that the premise can never hold ($\Gamma/L^2 \le \mathbb{E}[\mathrm{Var}(Y_1 \mid X)] \le \mathrm{MSE}_1$): the cascade's floor coincides with the flat floor (both $\mathrm{Bayes}(X)$), the attainable error at most matches it, and the deficit is $L^2\,\mathrm{MSE}_1 - \Gamma \ge 0$. CGA's entire advantage is optimization-side (Thms. 1–3b, 5); $I(Y_1; Y_2 \mid X) > 0$ is retained as the necessary condition for cascade headroom (implied by $\Gamma > 0$).

E Appendix E: Covariance Decomposition of the Generative Structure (Motivation for Assumption 1)

E.1 Problem statement. The core of Assumption 1 is that, in the variance structure of X, the subspace $\mathcal{S}_1$ serving $\mathcal{T}_1$ and the subspace $\mathcal{S}_2$ serving the $\mathcal{T}_2$ residual do not co-vary. This appendix gives the generating-mechanism-level covariance argument. Note: the decomposition *motivates* the assumption rather than strictly proving it—subspace orthogonality is directly verified by the full-run measurement of §5.3 ($\cos(\mathbf{g}_1, \mathbf{g}_2)$ pinned at 0.00 ± 0.02).

E.2 Setup and notation. By the SCM (Def. 1):

$$X = U_X, \qquad Y_1 = f_1(X, U_1), \qquad Y_2 = f_2(X, Y_1, U_2),$$

with $U_1 \perp\!\!\!\perp U_2$. Assume f_1, f_2 continuously differentiable. Define $\mathcal{S}_1 = \mathrm{span}\{\partial Y_1/\partial X\} \subseteq \mathbb{R}^{d_X}$ (directions of Y_1's gradient field in X-space) and $\mathcal{S}_2 = \mathrm{span}\{\partial(Y_2 \mid Y_1)/\partial X\} \subseteq \mathbb{R}^{d_X}$ (directions of the residual gradient field of Y_2 given Y_1). Goal: show that, in population, $\mathcal{S}_1 \perp \mathcal{S}_2$ in the eigenbasis of X's variance–covariance matrix.

E.3 Core argument. Lemma E.1 (covariance decomposition). Under the SCM, the components of X's variance along $\mathcal{S}_1$ and along $\mathcal{S}_2$ are orthogonal in population covariance:

$$\mathrm{Cov}_X\left(\tfrac{\partial f_1}{\partial X}(X, U_1),\ \tfrac{\partial (f_2 | f_1)}{\partial X}(X, U_2)\right) = \mathbf{0}_{d_X \times d_X}.$$

Proof. Write $G_1(X, U_1) = \partial f_1/\partial X$ and $G_2(X, U_2) = \partial(f_2 \mid f_1)/\partial X$ (given $Y_1 = f_1(X, U_1)$, the residual variability of f_2 depends on external noise only through U_2). Then

$$\mathrm{Cov}(G_1, G_2) = \mathbb{E}_{X, U_1, U_2}[G_1(X, U_1)\, G_2(X, U_2)^\top] - \mathbb{E}_{X, U_1}[G_1]\, \mathbb{E}_{X, U_2}[G_2]^\top.$$

Since $U_1 \perp\!\!\!\perp U_2$ and $X = U_X$ is exogenous, the joint factorizes as $P(X, U_1, U_2) = P(X)P(U_1)P(U_2)$; hence

$$\mathbb{E}_{X, U_1, U_2}[G_1 G_2^\top] = \mathbb{E}_X\left[\mathbb{E}_{U_1}[G_1 \mid X]\, \mathbb{E}_{U_2}[G_2^\top \mid X]\right].$$

Strictness qualifier (honesty declaration). The last step requires $\mathbb{E}[G_1 \mid X]$ and $\mathbb{E}[G_2 \mid X]$ to be independent of X (conditional means degenerate to unconditional means). This holds *strictly* in the additive linear–Gaussian SCM: with $f_1 = w_1^\top X + u_1$, $\partial f_1/\partial X = w_1$ is constant and the equality is exact. In general nonlinear cases $\mathbb{E}[G_1 \mid X]$ depends on X, so the identity is *motivational* (the residual is the co-variation cross term of G_1, G_2 with X). Lemma E.1's accurate statement is therefore: *strict in the additive linear–Gaussian setting, motivational otherwise*; the direct support of Assumption 1 always comes from the full-run gradient measurement of §5.3. Substituting into the covariance formula gives $\mathrm{Cov}(G_1, G_2) = \mathbf{0}$ in the linear–Gaussian case. ■

Interpretation. Lemma E.1 establishes the *non-co-varying generating structure* of G_1 and G_2 in population covariance—the generating-mechanism motivation for $\mathcal{S}_1 \perp \mathcal{S}_2$. Note: zero covariance does not by itself imply orthogonality of the supporting subspaces (two uncorrelated random vectors may have overlapping supports), so we do not "strictly prove" orthogonality as a corollary of the decomposition; instead we formalize it as Assumption 1 and verify it directly by full-run gradient measurement (§5.3).

E.4 From the covariance structure to Thm. 1. By Lemma E.1, $\mathcal{S}_1$ and $\mathcal{S}_2$ do not co-vary in the population variance structure; Assumption 1 elevates this to subspace separation. Under it, Thm. 1 follows from the argument in §3.2: the stationarity condition $\mathbf{g}_1 + \lambda \mathbf{g}_2 = 0$ is incompatible with the assumption (dotting with $\mathbf{g}_1$ forces $\|\mathbf{g}_1\|^2 = 0$), so no stationary point exists in the persistent-excitation region.

E.5 Finite samples. With finite samples, the empirical estimate of $\mathrm{Cov}(G_1, G_2)$ converges to zero at rate $\mathcal{O}(1/\sqrt{n})$ (central limit theorem), where n is the sample size. At the scales used here (synthetic $n = 10{,}000$; real AFC data of six months), Assumption 1 thus holds to high empirical precision, confirmed by the gradient tracking of §5.3.

E.x Local argument and failure mode in the nonlinear case. *Local argument*: for smooth f_1, f_2, a first-order Jacobian approximation at any parameter point preserves gradient orthogonality up to $\mathcal{O}(\max_i \|f_i''\|)$, and the $\epsilon_\perp$-version conclusions (Appendix F, F.5) absorb this error—so the conditional results hold on the class of cascades with mechanism-orthogonal subspaces. *Failure mode (measured)*: when the coupling grows so that $\cos(\mathbf{g}_1, \mathbf{g}_2)$ rises to $+0.3$–0.8 and κ_g falls to 0.1–0.3 (Table 6), orthogonal bullying disappears—no T2 collapse, no CAGrad stall—and the conclusions degrade predictably: the convergence tax decays with β but does not vanish (Flat pays 2.1×/2.1×/1.7× over single-task T2, and CGA recovers to parity at $\beta = 3$).

F Appendix F: Full Proof of Thm. 3b

F.0 Mechanism of the trilemma (moved from main text). Decomposing $\mathbf{d} = a\hat{\mathbf{g}}_1 + b\hat{\mathbf{g}}_2 + \mathbf{w}$, (i)+(iii) force $a \geq (1-\delta)\|\mathbf{g}_1\|$ and $b \geq (1-\delta)\|\mathbf{g}_2\|$, and the curvature bound gives $\Delta\mathcal{L}_2(\mathbf{d}) > 0$ once $\kappa_g \geq K$.

F.1 Assumptions and notation.

- (A1) Assumption 1: $\cos\theta_g \approx 0$, i.e. $|\cos\theta_g| \leq \epsilon_\perp$; for clarity we take the $\epsilon_\perp = 0$ limit below, with all conclusions holding up to $O(\epsilon_\perp)$ errors otherwise (see F.5).
- (A2) $H_2 \succeq 0$ (locally), globally non-degenerate: for all $\mathbf{u} \in \mathrm{span}\{\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2\}$ and $\mathbf{w} \perp \mathrm{span}\{\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2\}$, $|\mathbf{w}^\top H_2 \mathbf{u}| \leq \rho\sqrt{(\mathbf{w}^\top H_2 \mathbf{w})(\mathbf{u}^\top H_2 \mathbf{u})}$ and $|c_{12}| \leq \rho\sqrt{c_H c_2}$ with $\rho < 1$ (ruling out the measure-zero case of zero curvature coupling between the gradient plane and its orthogonal complement). Hence $\mathbf{d}^\top H_2 \mathbf{d} \geq (1-\rho)(\mathbf{u}^\top H_2 \mathbf{u} + \mathbf{w}^\top H_2 \mathbf{w})$ and $\mathbf{u}^\top H_2 \mathbf{u} \geq (1-\rho)(a^2 c_H + b^2 c_2)$; combine and re-denote $(1-\rho)^2$ as $1-\rho$ (affects constants only).
- (A3) $c_H \geq c_0 > 0$, $\delta \in (0, 1)$.
- (A4) Locality: the analysis holds in the region where the second-order Taylor expansion is valid ($\eta\|\mathbf{d}\|$ bounded).

For any $\mathbf{d}$, the one-step net loss change of $\mathcal{T}_2$ (to second order) is

$$\Delta\mathcal{L}_2(\mathbf{d}) = -\eta\langle \mathbf{d}, \mathbf{g}_2\rangle + \tfrac{\eta^2}{2}\mathbf{d}^\top H_2 \mathbf{d} + O(\eta^3\|\mathbf{d}\|^3).$$

F.2 Proof of (i)(iii) $\Rightarrow \neg$(ii). Write $\mathbf{d} = a\hat{\mathbf{g}}_1 + b\hat{\mathbf{g}}_2 + \mathbf{w}$ ($\mathbf{d}$ does not contain η; the actual update is $-\eta\mathbf{d}$). From (i)(iii): $a \geq (1-\delta)\|\mathbf{g}_1\|$ and $b \geq b_{\min} := (1-\delta)\|\mathbf{g}_2\|$ (since $\langle \mathbf{d}, \mathbf{g}_2\rangle = b\|\mathbf{g}_2\| \geq (1-\delta)\|\mathbf{g}_2\|^2$). By the curvature lower bound of F.1:

$$\begin{aligned}\Delta\mathcal{L}_2(\mathbf{d}) &\geq -\eta b\|\mathbf{g}_2\|^2 + \tfrac{\eta^2}{2}(1-\rho)\big(a^2 c_H + b^2 c_2\big)\\ &\geq -\eta b\|\mathbf{g}_2\|^2 + \tfrac{\eta^2}{2}(1-\rho)\big((1-\delta)^2\|\mathbf{g}_1\|^2 c_H + b^2 c_2\big).\end{aligned}$$

Denote the right-hand side by $\phi(b)$, a convex quadratic in b with minimizer $b^* = \|\mathbf{g}_2\|^2/(\eta(1-\rho)c_2)$.

Case 1 ($b^* \geq b_{\min}$). $\min_{b\geq b_{\min}} \phi = \phi(b^*) = \frac{\eta^2}{2}(1-\rho)(1-\delta)^2\|\mathbf{g}_1\|^2 c_H - \frac{\|\mathbf{g}_2\|^4}{2(1-\rho)c_2}$, strictly positive iff

$$\kappa_g^2 > \frac{\|\mathbf{g}_2\|^2}{c_2 c_H \eta^2 (1-\rho)^2 (1-\delta)^2}.$$

Case 2 ($b^* < b_{\min}$, i.e. $\|\mathbf{g}_2\| < \eta(1-\rho)(1-\delta)c_2$). ϕ is increasing on $[b_{\min}, \infty)$, so $\min_{b\geq b_{\min}} \phi = \phi(b_{\min}) = \frac{\eta^2}{2}(1-\rho)(1-\delta)^2\|\mathbf{g}_1\|^2 c_H - \eta(1-\delta)\|\mathbf{g}_2\|^3 + \frac{\eta^2}{2}(1-\rho)(1-\delta)^2\|\mathbf{g}_2\|^2 c_2$, strictly positive iff

$$\kappa_g^2 > \frac{2\|\mathbf{g}_2\|/\eta - (1-\rho)(1-\delta)c_2}{(1-\rho)(1-\delta)\, c_H} \quad \text{(unconditional when the numerator is non-positive).}$$

Case 2 applies only when $\|\mathbf{g}_2\| < \eta(1-\rho)(1-\delta)c_2$ (near $\mathcal{T}_2$'s convergence or very small steps), with a threshold upper-bounded $\propto 1/\eta$, while the Case-1 threshold is $\propto \|\mathbf{g}_2\|^2 \to 0$—the closer

$\mathcal{T}_2$ gets to convergence ($\|\mathbf{g}_2\| \to 0$), the more unconditionally the trilemma holds, consistent with Cor. 2. Both thresholds K^2 are normalized dimensionless quantities ($\mathbf{d}$ is normalized by gradient norms); their physical meaning is "the minimal imbalance ratio in units of the natural step." Take

$$K^2 := \max\left\{\frac{\|\mathbf{g}_2\|^2}{c_2 c_H \eta^2 (1-\rho)^2 (1-\delta)^2},\ \max\left(0, \frac{2\|\mathbf{g}_2\|/\eta - (1-\rho)(1-\delta)c_2}{(1-\rho)(1-\delta)\, c_H}\right)\right\}.$$

When $\kappa_g \geq K$, any $\mathbf{d}$ satisfying (i)(iii) has $\Delta\mathcal{L}_2(\mathbf{d}) \geq \phi(b) > 0$, so (ii) is violated. ■

F.3 The remaining two combinations. If some $\mathbf{d}$ satisfied (i)(ii)(iii) simultaneously, the implication (i)(iii)⇒ ¬(ii) of F.2 would be contradicted. Hence (i)(ii) forces ¬(iii), and (ii)(iii) forces ¬(i). The trilemma follows. ■

F.4 Threshold magnitudes in the §5.3 setting (illustrative). Take measured magnitudes and typical values: $\kappa_g > 200$, $\eta \sim 10^{-3}$, $\|\mathbf{g}_2\| \sim 10^{-2}$, $c_H \sim 0.5$, $c_2 \sim 1$, $\rho = 0.5$, $\delta = 0.1$. Then $b^* = \|\mathbf{g}_2\|^2/(\eta(1-\rho)c_2) = 0.2 \geq b_{\min} = (1-\delta)\|\mathbf{g}_2\| = 0.009$: Case 1 applies with threshold $\kappa_g^2 \approx 988$ ($K \approx 31$), covered by the measured $\kappa_g > 200$. Case 2 applies when $\|\mathbf{g}_2\| < \eta(1-\rho)(1-\delta)c_2 \approx 4.5 \times 10^{-4}$ (near $\mathcal{T}_2$'s convergence), where its threshold tends to zero as $\|\mathbf{g}_2\| \to 0$. This paragraph is an order-of-magnitude illustration, not a theorem dependency; measured c_H, c_2 values can be substituted in formal experiments.

F.5 Proof of Thm. 3a and the $\epsilon_\perp > 0$ remark. The trigger condition $\langle \mathbf{g}_1, \mathbf{g}_2 \rangle < -t$ is equivalent to $\cos(\mathbf{g}_1, \mathbf{g}_2) < -t/(\|\mathbf{g}_1\|\|\mathbf{g}_2\|)$. Under Assumption 1, $|\cos\theta_g| \leq \epsilon_\perp$: when $t > \epsilon_\perp \|\mathbf{g}_1\|\|\mathbf{g}_2\|$ the trigger probability is zero; at $t = 0$ (PCGrad's setting) triggering requires $\cos < 0$, whose probability is governed by the empirical distribution of cos (0.0% in our synthetic measurement, §5.3). When triggered, the correction satisfies $\|\mathrm{proj}_{\mathbf{g}_1}\mathbf{g}_2\| = |\langle \hat{\mathbf{g}}_1, \mathbf{g}_2 \rangle| \leq \epsilon_\perp \|\mathbf{g}_2\|$. ■

F.1–F.4 above take $\epsilon_\perp = 0$; for $\epsilon_\perp > 0$ the inner products pick up $O(\epsilon_\perp \|\mathbf{g}_1\|\|\mathbf{g}_2\|)$ corrections and the thresholds shift by $O(\epsilon_\perp)$ terms. The measured $\epsilon_\perp \approx 0.02$ makes the correction negligible. (This $\epsilon_\perp$ is the orthogonality slack of Assumption 1, not the accuracy parameter ε of convergence rates.)

G Appendix G: OD Evaluation Metrics and Irreducible-Error Analysis

G.1 Definitions. Let the test period contain T hours, with OD ground truth $X_{t,ij}$ and predictions $\hat{X}_{t,ij}$.

- **Per-interval standard WMAPE** (the literature-standard metric): $\frac{\sum_{t,i,j:X_{t,ij}>0} |\hat{X}_{t,ij} - X_{t,ij}|}{\sum_{t,i,j:X_{t,ij}>0} X_{t,ij}}$—absolute errors are summed per hour, so noise does not cancel with T.
- **Time-aggregated WAPE$^+$** (used in our controlled re-run): $\frac{\sum_{(i,j):\sum_t X_{t,ij}>0} |\sum_t \hat{X}_{t,ij} - \sum_t X_{t,ij}|}{\sum_{(i,j):\sum_t X_{t,ij}>0} \sum_t X_{t,ij}}$—totals are accumulated per OD pair before the ratio, so errors can cancel across hours and noise folds with $\sqrt{T}$.

G.2 Irreducible-error floors (measured on our data, $N = 270$, 2,009 labeled hours). Cross-week repeats of the same OD pair in the same (weekday, hour) slot (up to 14 weeks) form natural replicate samples; their fluctuation lower-bounds individual travel randomness:

Metric	Slot-mean oracle	Poisson irreducible floor
Per-interval WMAPE	44.5% (test segment)	≈31% (closed form $E\|X-\lambda\| = 2\lambda\, P(X = \lfloor\lambda\rfloor)$)
Time-aggregated WAPE$^+$	**10.22%**	**≈2.14%**

G.3 What the numbers mean in this paper. (i) Under the per-interval metric no model can go below the ≈31% Poisson floor; the time-aggregated metric (floor ≈2.14%, slot-mean oracle 10.22%) was the metric under which the deployment's historical OD figure (3.67%) was originally reported—that figure is not reproducible from surviving artifacts and is not cited here (Appendix B); the time-

aggregated metric is now used only for the cross-method comparison within Table 2. (ii) Cross-literature comparison: published OD-prediction work mostly uses the per-interval WMAPE (typical range 20–42%); our controlled re-run reports both the time-aggregated WAPE$^+$ and the per-interval WMAPE, and cross-literature comparison should use the per-interval metric. (iii) All methods in Table 2 share the same dual-metric evaluation, so within-table comparisons are fair.

H Appendix H: Synthetic SCMs and Complete Results (Current v3 Protocol + v2 Legacy Records)

H.0 v3 protocol (current main experiments, §5.4/5.5). $X \sim \mathcal{N}(0, I_{64})$; **T1 dense 50-dim**: $Y_1 = s_1 W_1 X + E_1$ (row-normalized $W_1 \in \mathbb{R}^{50\times 64}$, $s_1 = 5$, $\sigma_1 = 0.5$); **T2 sparse 600-dim, 1% active**: $Y_{2j} = s_2 t_j + \beta(v_j^\top Y_1) + e_{2j}$ ($j \in A$, $|A| = 6$; $t_j = w_{2j}^\top X$, $w_{2j} \perp \mathrm{span}(W_1)$; $s_2 = 1.5$, $\sigma_2 = 0.1$; v_j mixed-sign unit vectors, coherence $\mathrm{coh} \in \{0, 0.5, 1\}$). Closed forms: $\Gamma = \beta^2\sigma_1^2$, $c_H \propto \beta^2|A|$. Training: SGD momentum 0.9; $\eta_1 = 10^{-2}/\max(1, \lambda_1/3)$ (T1 stability boundary, $\lambda_1 \in \{1, 3, 10\}$); per-task η_2; 40 epochs, batch 256, $n = 10{,}000$ (70/15/15); Flat h=128 vs CGA per-stage h=80 (capacity-matched); 3 seeds (42,123,456; the p/coherence mechanism scans and the rate-tax runs use 2 seeds). **Design note**: in v3, orthogonality does not rely on weak coupling—it comes from the geometry "dense T1 + sparse T2 + incoherent coupling matrix" ($w_2 \perp \mathrm{span}(W_1)$ gives subspace separation; mixed-sign v_j makes the parent-channel gradients cancel coherently across target dimensions). Data and code: `experiments_v3/`.

H.0.1 Protocol iteration history and findings (A–J). Six protocol versions and six mechanism findings before v3 froze, retained for auditability (so "why v3 is designed this way" cannot be post-hoc rationalized) and to archive negative/confound findings explicitly:

Version / failure	Root-cause finding
v3.0 λ_1=100 diverges; λ_1=10 zero-collapse ($\kappa \to 0$)	**A**: $\theta = 0$ is a degenerate saddle under zero-mean labels (motivates the $\lambda_1\eta_1$ stability-boundary rule; same origin as production's all-zero trap)
v3.0 global clipping starves T2 at β=0 (WAPE= 1.0)	**B**: global clipping ≡ weak-task learning-rate decay
v3.1 even single-task T2 fails at β=0	**C**: sparse labels + SGD + full-dim mean loss dilute each gradient 600× (AdamW masked this in v2)
v3.1/3.2 strong coupling pushes cos to 0.29–0.62	**D**: the parent-channel coherence term cannot be diluted—the naive "sparse high-dim gives cos ≈ 0 with strong coupling" fails in gradient algebra
v3.3 Flat degenerates at β=0 (capacity confound)	**E**: dense-T1+sparse-T2+incoherent coupling is the only construction reproducing cos ≈ 0 with strong coupling in low-dim MLPs
v3.4 (frozen)	**F**: Flat's seed sensitivity + erase–rebuild signature ($\mathrm{T2}_{\mathrm{best}} \ll \mathrm{T2}_{\mathrm{final}}$)

Further findings: **G** (Flat improves with $\lambda_1 \Rightarrow$ "wandering duration" must be separated from κ_g); **H** (first-order coupling is protective; the crossing condition $\frac{\eta}{2}\kappa_g^2 c_H > 1 + \kappa_g \cos$, i.e. $\kappa_g > \kappa^*$, verified in two configurations); **I** (IMTL-G's κ_g^2 amplification → NaN on all cells); **J** (module replacement works under intervention; "freeze-and-forget" does not—$\mathrm{sg}(\hat{Y}_1)$ input-distribution drift requires a light Stage-2 drift correction).

H.0.2 Pre-registered predictions P1–P8 and their status. P1 $\hat{\Gamma}$ closes with the closed form ✓; P2 cos ≈ 0 coexists with strong coupling ✓; P3 κ_g monotone in λ_1 ✓; P4 c_H measured by HVP ✓; P5 Thm. 1 pathology ✓ (direction corrected: the wandering region is sharpest at $\kappa \approx 1$, not $\kappa \gg 1$); P6 rate tax ✓ (zoom4 + Fig. 11); P7 mediation failure ✓ (refined to "pathology translation": subspace/whitening methods do help at moderate κ); P8 β=0 control column ✓.

H.1–H.6 v2 legacy records (the weak-coupling protocol superseded by v3 in §5.4, retained for comparison).

H.1 Explicit SCM. $X = U_X \sim \mathcal{N}(0, I_{32})$; $w_1 \perp w_2$; $Y_1 = s_1(Xw_1) + \varepsilon_1$, $\varepsilon_1 \sim \mathcal{N}(0, 0.1^2)$; $Y_2 = \beta Y_1 + s_2(Xw_2) + \varepsilon_2$, $\varepsilon_2 \sim \mathcal{N}(0, 0.2^2)$. The orthogonality constraint $\beta s_1 \le 0.5 s_2$ keeps Y_2's predictable signal residual-channel-dominated—which is exactly v2's "weak-coupling self-defeat" ($\Gamma \approx 0$), removed in v3 by mixed-sign coupling + dense T1 (design note above). κ_g calibration: grid-search (s_1, s_2, β) so the measured κ_g at shared-encoder initialization hits the target.

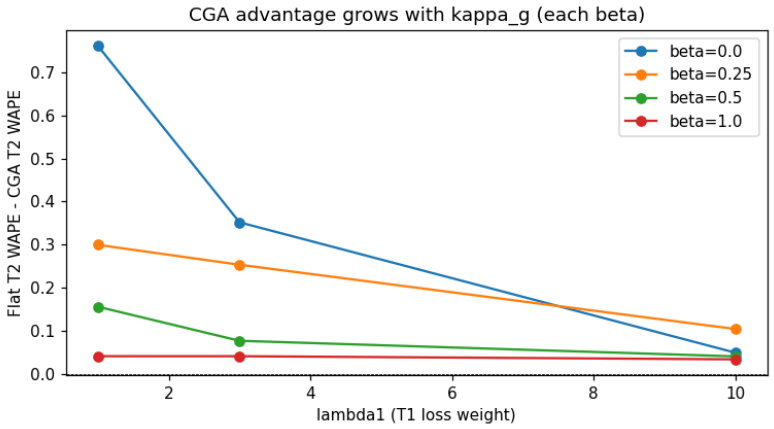

Figure 2: Flat minus CGA T2 gap vs λ_1 (one line per β), v3 $\beta \times \lambda_1$ grid.

H.2 Calibration table. Target κ: 1/5/10/20/50; s_1: 2.0/12.0/20.0/50.0/150.0; s_2: 1.5/1.0/0.5/2.0/3.0; β: 0.25/0.002/0.01/0.01/0.01; measured κ_{init}: 1.00/5.00/10.00/19.85/46.74.

H.3 Protocol. Data generated once with fixed seed 0 ($n = 50{,}000$, train/val/test 40k/5k/5k); 3 model seeds $\{42, 123, 456\}$; AdamW lr 10^{-3}, weight decay 10^{-5}, batch 512, 60 epochs, gradient clipping 5.0; FlatMLP (2-layer 64-dim shared + dual heads) vs CGAMLP (Stage-1 2-layer + Stage-2 2-layer) capacity-matched; metrics: T2 (OD) and T1 (flow) WAPE.

H.4 Complete T2 WAPE table (3-seed means). At $\kappa = 1$ all methods 9.2–9.6 (CAGrad 9.2, GradNorm 17.7); at $\kappa = 5$ all 44.7–47.1 (CAGrad 46.0, GradNorm 47.1); $\kappa = 10/20/50$ are in Table 5. PCGrad/ConicGrad/MGDA/GradOPS/IMTL-G stay within 1pp of Flat at every κ.

H.5 Gradient dynamics ($\kappa = 50$, v2 SCM). Full-run mean $\cos = -0.0024$; κ_g decays from 46.7 (init) to a training mean of 13.9; PCGrad trigger rate 0.0%.

H.6 v2 legacy ablation (weak coupling $\beta = 0.05$, $\kappa_g = 10$, 60 epochs, 3 seeds): CGA full 1.89/38.76±0.09; w/o Stop-Grad 2.04/38.69±0.19; w/o Cascade 1.82/38.74±0.22; Reverse 2.03/39.00±0.09 (T1/T2 WAPE%). All within ±0.25pp—Thm. 6's quantitative prediction in the $\Gamma \approx 0$ region; the decisive Stop-Grad evidence is the strong-coupling ablation (§5.6).

H.7 Figure inventory. Embedded: Figs. 9–11 (main text); Figs. 2–8 (supplementary v3 figures, embedded below). The real-data trio (CKA/t-SNE geometry, real-data erase–rebuild curve, real-data κ_g–iteration check) ships with the completed 15-arm ladder and is registered but unnumbered.

H.8 Reproduction guide.

```
Environment: Python 3.11 + torch 2.9 (CPU) + numpy/pandas/matplotlib
Directory: experiments_v3/  (results dumped to results/*.json)
  python run_grid.py --mode main           # beta x lambda_1 x methods
  python run_grid.py --mode secondary      # GradNorm / IMTL-G
  python run_grid.py --mode mediators2     # MGDA / ConicGrad / GradOPS
  python run_grid.py --mode p              # sparsity scan
  python run_grid.py --mode coherence      # coherence scan
  python run_grid.py --mode ratetax        # rate tax (Flat/Flat-Prot/CGA)
  python run_grid.py --mode intervention   # mechanism shift
  python run_diamond.py                    # diamond graph
  python run_ablation3.py                  # strong-coupling ablation
  python _probe_zoom4.py                   # zoom4 step window
  python _make_summary.py                  # -> results summary
  python analyze.py                        # -> summary_tables.md + figs/
```

H.9 Supplementary tables (moved from the main text for space). Table 4 gives the complete 12-configuration $\beta \times \lambda_1$ grid; the main text (Table 1) shows $\beta \in \{0.0, 1.0\}$ for space.

H.10 κ-grid figure (moved from the main text for space). Table 5 in figure form.

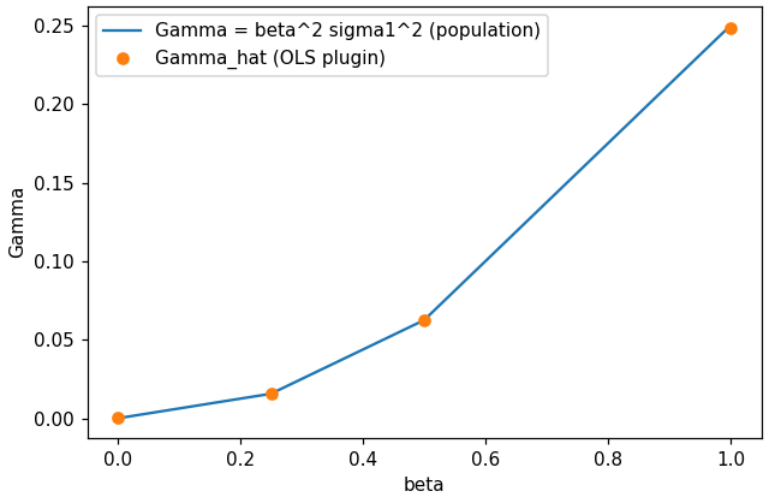

Figure 3: Estimated $\hat{\Gamma}$ vs β, overlaid on the closed form $\beta^2\sigma_1^2$.

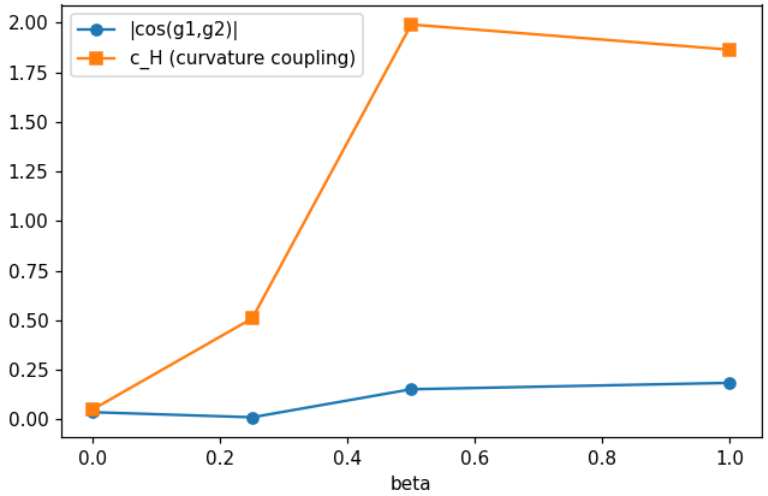

Figure 4: $|\cos|$ and c_H vs β (λ_1=10 column).

H.11 Erase–refit sawtooth (moved from the main text for space). Fig. 10.

H.12 κ_g^2 rate-tax figure (moved from the main text for space). Fig. 11.

H.13 κ-grid protocol (details). SGD momentum 0.9; η_1 set at $\mathcal{T}_1$'s stability boundary with $\lambda_1\eta_1$ held constant so that Thm. 2's $\eta \asymp 1/L_1$ holds per column; 40 epochs; 3 seeds; Flat and CGA capacity-matched; κ is the calibration target hit by the measured κ_g at initialization, and absolute values are not comparable with Table 1 (different SCM scales).

I Appendix I: Closed-Form Worked Example for the Linear–Gaussian Model

This appendix strings all main-text results (Thm. 1, Lemma 2, Thm. 5, Prop. 4, Thm. 6) into one closed-form chain on a solvable one-dimensional linear model; all statements are exact equalities, no approximations.

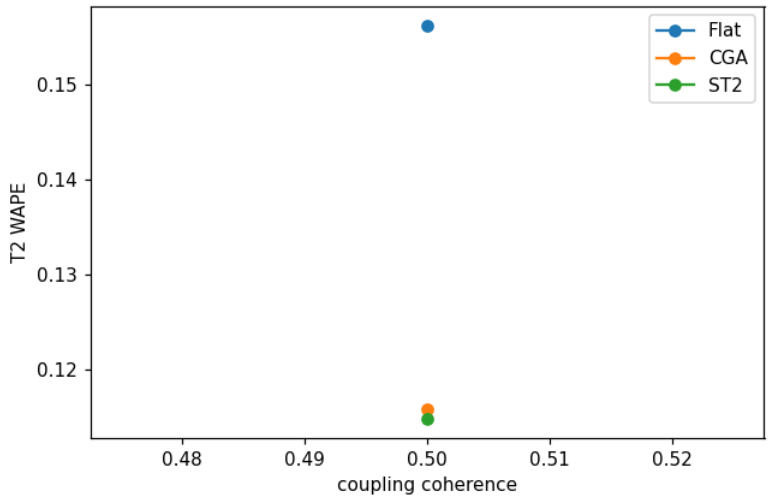

Figure 5: Coupling-coherence scan: T2 WAPE vs coupling coherence coh (Flat / CGA / ST2 single-task floor).

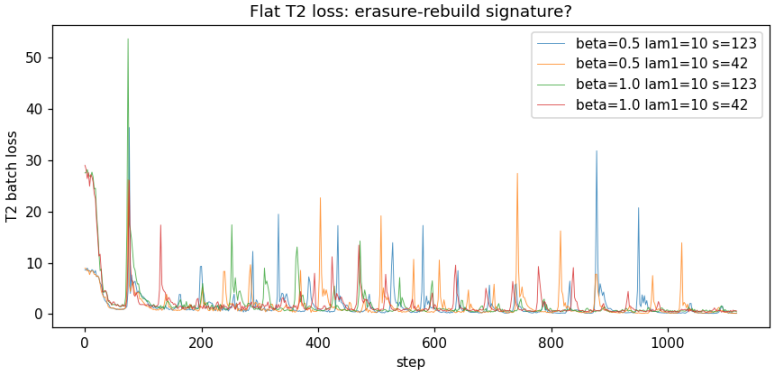

Figure 6: v3 wandering-region erase–rebuild sawtooth (Flat's step-level T2 loss).

I.1 Setup. $X \sim \mathcal{N}(0, I_d)$, $w_1 \perp w_2$ (unit vectors), independent $\varepsilon_1 \sim \mathcal{N}(0, \sigma_1^2)$, $\varepsilon_2 \sim \mathcal{N}(0, \sigma_2^2)$; SCM $Y_1 = s_1 w_1^\top X + \varepsilon_1$, $Y_2 = \beta Y_1 + s_2 w_2^\top X + \varepsilon_2$. Flat architecture: one-dimensional shared representation $Z = WX$ ($W \in \mathbb{R}^{1\times d}$), linear heads $\hat{y}_1 = Z$, $\hat{y}_2 = cZ$. With $W = \alpha w_1^\top + \gamma w_2^\top + \nu^\top$, $\nu \perp \mathrm{span}\{w_1, w_2\}$, the population losses are $L_1 = \|W - s_1 w_1^\top\|^2 + \sigma_1^2$ and $L_2 = (c\alpha - \beta s_1)^2 + (c\gamma - s_2)^2 + c^2\|\nu\|^2 + \beta^2\sigma_1^2 + \sigma_2^2$.

I.2 Gradient orthogonality (Thm. 1's linear case). Saturation region of $\mathcal{T}_1$: $\gamma = 0$, $\nu = 0$ (the representation lies fully in the $\mathcal{T}_1$ subspace; the $\gamma^2 + \|\nu\|^2$ terms of L_1 are pressed to zero by $\mathcal{T}_1$'s dominant gradient). Head optimality: $\partial L_2/\partial c = 0$ gives $c = \beta s_1/\alpha$. Then

$$\mathbf{g}_1 = 2(\alpha - s_1)w_1^\top, \qquad \mathbf{g}_2 = 2c[(c\alpha - \beta s_1)w_1^\top - s_2 w_2^\top] = -2cs_2 w_2^\top,$$

so $\langle \mathbf{g}_1, \mathbf{g}_2 \rangle = 0$ *strictly*. The mechanism: the $\mathcal{T}_2$ head fits the parent channel entering through the shared representation ($c\alpha - \beta s_1$) to zero, leaving only the residual channel $-2cs_2 w_2^\top$, orthogonal to w_1. As long as $\alpha \neq s_1$ ($\mathcal{T}_1$ still training), orthogonality persists in the saturation region.

I.3 No stationary point (Thm. 1). The joint stationarity condition $\mathbf{g}_1 + \lambda \mathbf{g}_2 = 0$ with $w_1 \perp w_2$ requires $\mathbf{g}_1 = 0$ and $\mathbf{g}_2 = 0$ separately. The former gives $\alpha = s_1$; the latter requires $-2cs_2 = 0$—impossible since $s_2 > 0$ and $c = \beta s_1/\alpha \neq 0$. The joint objective has *no* stationary point in the saturation region: the iteration cannot converge, and a decayed learning rate freezes the trajectory at a non-stationary point whose location depends on the noise history.

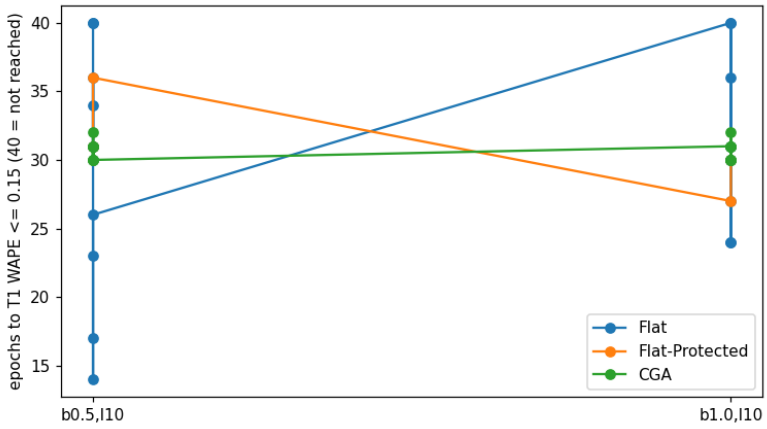

Figure 7: zoom4 step-window: $\mathcal{T}_1$ convergence epochs of Flat / Flat-Protected / CGA.

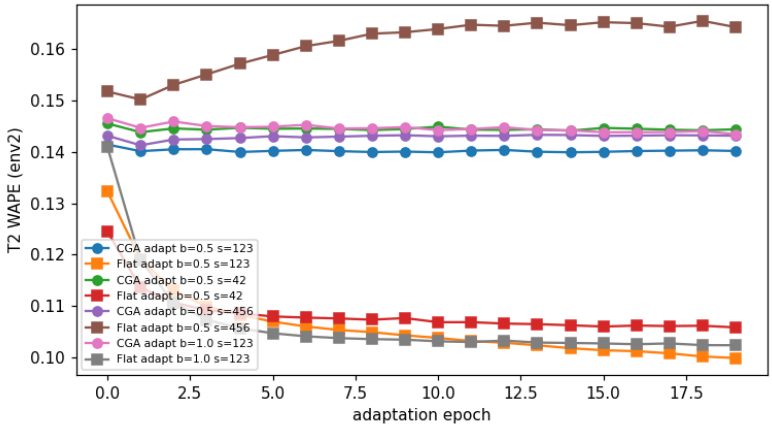

Figure 8: Intervention adaptation curves (CGA-adapt vs Flat-adapt).

I.4 Erasure ratio and step window (Lemma 2, Thm. 5). The Hessian is $H_2 = \nabla^2_W L_2 = 2c^2 I_d$, so $c_H = \hat{\mathbf{g}}_1^\top H_2 \hat{\mathbf{g}}_1 = 2c^2 > 0$. With the head frozen, $\mathcal{T}_1$'s step $\mathbf{d} = -\eta \mathbf{g}_1$ damages L_2 by $\Delta L_2 = \frac{\eta^2}{2} \mathbf{g}_1^\top H_2 \mathbf{g}_1 = \eta^2 c^2 \|\mathbf{g}_1\|^2$; the erasure ratio is $E = \frac{\eta}{2} \kappa_g^2 c_H$ with $\kappa_g = \|\mathbf{g}_1\| / \|\mathbf{g}_2\| =$

Table 4: Synthetic-Causal v3 $\beta \times \lambda_1$ grid (**complete**, 12 configurations $\times$ 8 methods on the unified 3-seed protocol): T2 WAPE, mean (std) over 3 seeds. ST2 = single-task floor. Best per row in bold; CGA holds the floor in every cell (per-cell std $\leq$0.003). **GradNorm and IMTL-G are excluded from this table**: they belong to the secondary scan with a different protocol, and their complete audit (GradNorm 12 configs $\times$ 3 seeds = 36 runs, 0 divergences; IMTL-G original 6 configs $\times$ 2 seeds = 12 runs, all diverged; clipped variant 12 configs $\times$ 2 seeds = 24 runs, 5 divergences) is reported in §5.4 (fairness audit) and Appendix M.3, which is authoritative for cross-method ranking.

		Post-hoc mediation and flat sharing						Ours	Floor
β	λ_1	Flat	PCGrad	ConicGrad	GradOPS	CAGrad	MGDA	CGA	ST2
0.0	1	0.86(0.51)	0.48(0.16)	0.30(0.16)	0.28(0.05)	0.23(0.04)	0.13(0.00)	**0.10(0.00)**	0.10(0.00)
0.0	3	0.45(0.37)	0.40(0.23)	0.52(0.36)	0.15(0.05)	0.74(0.05)	0.17(0.07)	**0.10(0.00)**	0.10(0.00)
0.0	10	0.15(0.02)	0.43(0.21)	0.10(0.00)	0.12(0.00)	1.00(0.53)	0.37(0.28)	**0.10(0.00)**	0.10(0.00)
0.25	1	0.41(0.23)	0.36(0.19)	0.29(0.11)	0.16(0.01)	0.37(0.06)	0.14(0.00)	**0.11(0.00)**	0.11(0.00)
0.25	3	0.36(0.20)	0.33(0.27)	0.16(0.02)	0.14(0.01)	0.62(0.07)	0.20(0.06)	**0.11(0.00)**	0.11(0.00)
0.25	10	0.21(0.01)	0.28(0.08)	0.11(0.00)	0.12(0.00)	0.80(0.14)	0.18(0.01)	**0.11(0.00)**	0.11(0.00)
0.5	1	0.27(0.10)	0.31(0.15)	0.18(0.02)	0.16(0.01)	0.19(0.00)	0.48(0.09)	**0.12(0.00)**	0.11(0.00)
0.5	3	0.19(0.07)	0.30(0.16)	0.13(0.00)	0.15(0.01)	0.21(0.05)	0.15(0.01)	**0.12(0.00)**	0.11(0.00)
0.5	10	0.16(0.02)	0.22(0.06)	0.12(0.00)	0.12(0.00)	0.52(0.35)	0.18(0.01)	**0.12(0.00)**	0.11(0.00)
1.0	1	0.16(0.03)	0.15(0.01)	0.22(0.05)	0.15(0.01)	0.16(0.01)	0.61(0.27)	**0.12(0.00)**	0.11(0.00)
1.0	3	0.16(0.02)	0.17(0.01)	0.15(0.02)	0.18(0.04)	0.23(0.02)	0.44(0.09)	**0.12(0.00)**	0.11(0.00)
1.0	10	0.16(0.02)	0.15(0.02)	0.12(0.01)	0.11(0.00)	0.16(0.01)	0.15(0.02)	**0.12(0.00)**	0.11(0.00)

Table 5: Orthogonal-SCM κ-grid (60 epochs, 3 seeds): T1/T2 WAPE%. CAGrad's upstream error crashes at $\kappa_g \geq 10$ (Thm. 3b); Flat and CGA stay $\leq$1.1%. κ is the calibration target; absolute values are not comparable with Table 1 (Appendix H).

κ	CGA		Flat		CAGrad		GradNorm		ST-T2
	T1	T2	T1	T2	T1	T2	T1	T2	T2
10	1.1	36.2	1.0	35.2	**82.7**	40.2	5.9	38.0	35.9
20	0.6	22.1	0.6	21.4	**71.9**	32.3	25.9	23.6	22.1
50	0.3	22.3	0.3	21.2	**87.0**	36.7	**53.3**	23.1	21.8

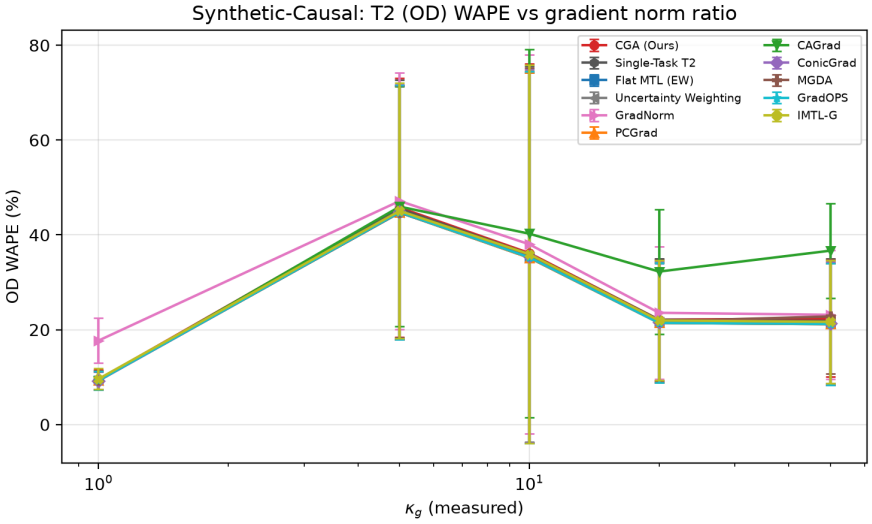

Figure 9: The κ-grid: upstream (T1) and downstream (T2) WAPE across methods as κ_g is swept (Table 5 in figure form).

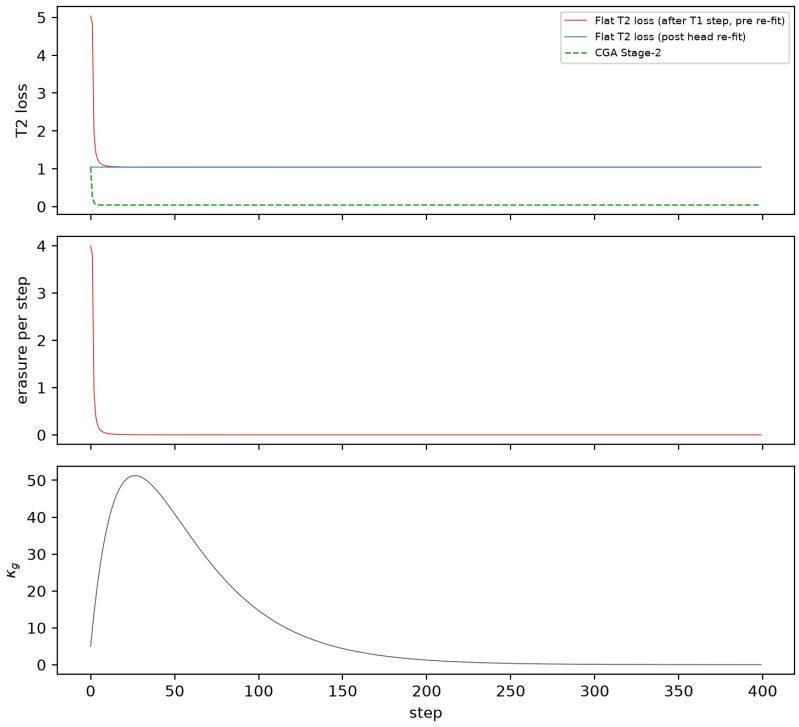

Figure 10: Erase–refit sawtooth (closed-form protocol): Flat's per-step T2 erasure pulses vs CGA's step noise ($9.9\times$ contrast).

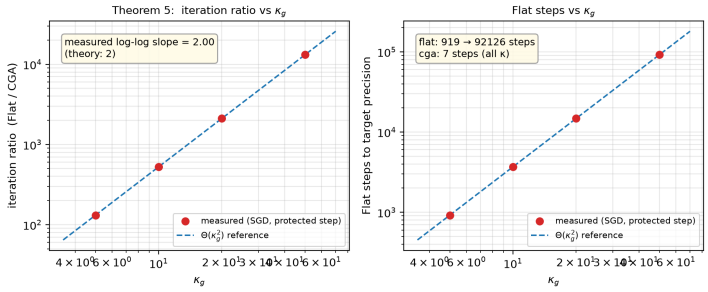

Figure 11: κ_g^2 rate tax: Flat/CGA iteration ratio vs κ_g (log-log slope 2.001; theory 2).

Table 6: Strong coupling ($\beta = 1/2/3$, $\kappa_{\text{target}} = 50$; T1/T2 WAPE%, 3 seeds): the failure mode of Assumption 1 ($\cos \gg 0$, $\kappa_g \ll 1 \Rightarrow$ first-order coupling protection; bullying does not activate) and the residual convergence tax on Flat.

β	Γ	Flat		CAGrad		CGA		ST		cos	κ_g
		T1	T2	T1	T2	T1	T2	T1	T2		
1	0.01	0.17	0.32	0.15	0.27	0.12	0.21	0.10	0.15	$+0.5 \sim 0.8$	0.1–0.3
2	0.04	0.12	0.19	0.16	0.24	0.11	0.14	0.08	0.09	$+0.5 \sim 0.8$	0.1–0.3
3	0.09	0.15	0.17	0.20	0.32	0.10	0.09	0.08	0.10	$+0.3 \to 0.8$	0.1–0.3

Table 7: Flat's measured diagnostics (λ_1=10 column). $\hat{\Gamma}$ closes with the closed form $\beta^2\sigma_1^2$ at every β.

β	mean cos	median κ_g	c_H (HVP)	$\hat{\Gamma}$	$\beta^2\sigma_1^2$
0.0	-0.04	10	0.05	0.0001	0
0.25	0.01	6	0.51	0.0156	0.0156
0.5	0.15	5	1.99	0.0620	0.0625
1.0	0.18	2	1.86	0.2480	0.2500

$|\alpha - s_1|/(|c|s_2) = \Theta(s_1/s_2)$. Protecting $\mathcal{T}_2$ ($E \le 1$) gives

$$\eta \le \eta_{\text{safe}} = \frac{4s_2^2}{\|\mathbf{g}_1\|^2} = \Theta\Big(\frac{s_2^2}{s_1^2}\Big) = \Theta\Big(\frac{1}{\kappa_g^2}\Big).$$

I.5 CGA decoupling (Prop. 4). In CGA the Stage-1 parameter u ($\hat{y}_1 = u^\top X$) and Stage-2 parameters are independent, and $\nabla_u L_2 = 0$ is guaranteed by Stop-Gradient. $L_1 = \|u - s_1 w_1\|^2 + \sigma_1^2$ is standard least squares in u; with $\eta = 1/2$ it halves the error every step (linear convergence). The flat architecture, limited to η_{safe}, needs $\mathcal{O}(\kappa_g^2)$ times more iterations for the same accuracy—Thm. 5's κ_g^2 factor is an exact equality here.

I.6 Cascade gap (Thm. 6). The conditional mean is $\mu(X, Y_1) = \mathbb{E}[Y_2 \mid X, Y_1] = \beta Y_1 + s_2 w_2^\top X$ with Lipschitz constant $L = |\beta|$. The Bayes-risk chain reads

$$\text{Bayes}(X) = \sigma_2^2 + \beta^2\sigma_1^2, \qquad \text{Bayes}(X, Y_1) = \sigma_2^2, \qquad \Gamma = \beta^2\sigma_1^2.$$

CGA Stage-2's attainable error with input $(X, \hat{y}_1)$ satisfies $\le \sigma_2^2 + \beta^2\text{MSE}_1 = \text{Bayes}(X) - \Gamma + L^2\text{MSE}_1$. Since no predictor's MSE falls below the conditional variance, $\text{MSE}_1 \ge \sigma_1^2 = \Gamma/L^2$ *always* (the premise "$\text{MSE}_1 < \Gamma/L^2$" is unsatisfiable): the deficit $\beta^2(\text{MSE}_1 - \sigma_1^2) \ge 0$, and the cascade can *at most match, never surpass*, the flat floor $\text{Bayes}(X) = \sigma_2^2 + \beta^2\sigma_1^2$; equality holds exactly at conditional-mean accuracy ($\text{MSE}_1 = \sigma_1^2$). "First secure Stage-1 accuracy, then train Stage-2" is the closed-form statement of the engineering rule minimizing this deficit.

J Appendix J: Additional Ablation, Misspecification, Noise-Sensitivity and Negative-Result Records

J.1 Diamond and misspecification (§5.6). On the four-node diamond (collider V_4, 3 seeds), correct-graph CGA attains T4 = 0.163 vs Flat 0.174 and four single tasks 0.181, while the misspecified chain gives 0.463 with one seed collapsing to WAPE = 1.000: "correct graph $\ge$ flat $\ge$ misspecified," the tie being Thm. 6's prediction in this pathology-free configuration. Strong-coupling ablations (β=1.0, Γ=0.25, 3 seeds): *w/o Stop-Grad* degrades in all seeds in the same direction (T2 +1.1pp, Prop. 4's sign, magnitude diluted by the mixed-sign coupling); *Reverse* collapses T1 to 0.646–0.773; *w/o Cascade* matches CGA's T2 (0.115 vs 0.123), directly measuring the Thm. 6 deficit.

J.2 Mechanism-shift intervention (§5.6). Under a sparse mechanism shift (environment 2 rotates W_1 by 60°, f_2 unchanged), zero-shot transfer degrades both tasks symmetrically (T1 $\approx$ 0.19); CGA-replace $\approx$ CGA-adapt, so *module replacement is an exact deployment primitive* (fresh Stage-1 into frozen Stage-2, no joint retraining)—Flat has no such primitive. Flat's full fine-tuning recovers well in these safe-valley configurations, so the accuracy-advantage form of the SMS claim is *not*

realized here; only the deployment primitive is. CGA's Stage-1-only adaptation is limited by $\mathrm{sg}(\hat{Y}_1)$ input-distribution drift, so full recovery wants a light Stage-2 drift correction ("freeze-and-forget" does not hold).

J.3 Noise sensitivity of the flow predictor (deployment model, v4 data). Under controlled perturbations of the Stage-1 input (noise scale $\sigma \in \{0, 0.05, 0.10, 0.20, 0.40, 0.80\}$), flow WAPE degrades 17.6%→18.1%→19.6%→24.4%→36.7%→64.4%—the inference-time error propagation governed by $O(L^2\mathrm{MSE}_1)$ (Thm. 6ii), non-negligible under large anomalies. The Stage-2 side of the original measurement is not reproducible (checkpoint/feature-version mismatch) and is withdrawn.

J.4 Registered negative records. (a) Flat improves as λ_1 (and κ_g) grows in the β=0 row—resolved by the two-axis decomposition (§5.4); (b) the pure-curvature-erasure regime ($E > 1$ with $\mathcal{T}_1$ stable) is *now realized constructively* (81 configurations: single-step 81/81, sustained window 27/81; §5.5, Appendix M) but is a *transient* window that is unreachable in the MLP family—this supersedes the earlier zoom2–4 negative record (J.5); (c) the intervention accuracy advantage is not realized (§5.6); (d) the historical OD metric is not reproducible (Appendix B); (e) GradNorm/IMTL-G ran 2 seeds on 6 cells, and the mechanism scans 2 seeds.

J.5 The zoom2–4 search, in detail. *zoom2* ($r = \eta_2/\eta$ control): the apparent erasure vanished under control—it was a confound from $\mathcal{T}_2$'s huge step ($r \approx 166$), not curvature erasure. *zoom3* (coh=1: $\cos \approx 0.065$, $c_H \approx 2.0$, $\kappa_g \approx 99$, $\eta = 3 \times 10^{-4}$—the same step setting as zoom2, so the arithmetic is checkable): $E \approx 2.9 > 1$ was measured, but the crossing condition $\kappa^* = (\cos + \sqrt{\cos^2 + 2\eta c_H})/(\eta c_H) = 231 > 99$ shows first-order protection still dominates and Flat is unharmed. *zoom4* ($\kappa_g \approx$ 23–29, $c_H \approx$ 5.6–15.3): at $\mathcal{T}_1$'s SGD stability boundary, $\mathcal{T}_1$ *destabilizes first*. Conclusion: pure curvature erasure was not cleanly realized *in this MLP synthetic class*; it is realized constructively in the quadratic family of §5.4 (81/81 single-step, 27/81 sustained), though as a *transient* rather than asymptotic window. zoom4 realizes the adjacent, distinct step-window mechanism, and our claim within the MLP class is strictly limited to it. On the magnitude gap: controlled synthetic κ_g is bounded by $\mathcal{T}_1$'s SGD stability ($\leq$10 on the main grid, ≈23–29 at zoom4), because pushing λ_1 to 1000 pushes $\mathcal{T}_1$'s step past its stability boundary, whereas production κ_g is measured under Adam's adaptive steps; the controlled re-run's two arms close the gap (SGD $\kappa_g \approx 1$, Adam 10^5–10^6, production 200–3000 in between)—a magnitude extrapolation, not a mechanism one, and the mechanism's sign does not flip with magnitude. The obsolete "epochs-to-target" table (based on an unreproducible single run and an ill-defined target under the joint seesaw trajectory) is withdrawn; convergence evidence rests on the final/best values of Table 1 and Fig. 11.

J.5b Step-window incompatibility (moved from main text). At β=1.0, coherence 1.0 ($\kappa_g \approx$ 23–29), the natural step destabilizes $\mathcal{T}_1$ (0.54–0.75) and erases $\mathcal{T}_2$ (0.41–0.69); step-compressed Flat-Protected recovers (0.17–0.58 / 0.11–0.49) while CGA reaches the floor (0.117) with no step compromise—despite Flat-Protected's double oracle advantage. Fig. 10 realizes the loop in closed form ($+0.0249$ vs 0.0025, $9.9\times$).

J.6 Ledger of the reproducible verification gap. The core prediction requires $\cos \approx 0$ (Assumption 1) *and* $\kappa_g \gg 1$ (Assumption 2) simultaneously. Evidence source by source: the production logs satisfy both ($\cos \approx 0.00$, $\kappa_g =$ 200–3000) but are not reproducible (no surviving scripts/weights); the controlled SGD arm has $\cos \approx 0$ (-0.045/$+0.023$/-0.133) but $\kappa_g = 0.60/1.79/2.00$ (pathology-free); the controlled Adam arm has $\kappa_g = 1.7\times10^5/7.5\times10^5/3.7\times10^5$ (max 4.4×10^6) but $\cos = -0.22/-0.13/-0.31$ (tilted away from orthogonality); the synthetic grids satisfy both but with $\kappa_g \leq 50$, one to two orders below production. Reading: our empirical claims are mechanism-level demonstrations (the pathology mechanism, the trilemma, the rate tax and the floor identity each have controlled realizations), not magnitude-level extrapolations; constructing a controlled real-data configuration satisfying both premises is the primary next task.

K Appendix K: Comparison with the Structured-Sharing Family (MMoE / PLE / AdaShare)

MMoE, PLE and AdaShare restructure sharing *before* training, so Thm. 3b (which constrains operators outputting a single direction in shared parameter space) does not govern them; §4.3 treats them

separately. The difference from CGA is threefold. *(1) Information channel*: $\mathrm{CGA}(\mathcal{G})$ prescribes directed conditional dependencies—M_c's input contains $\mathrm{sg}(\hat{Y}_p)$, an *information edge* carrying the parent's predicted value—whereas MMoE/PLE express only "which experts each tower uses" and have no channel to transmit Y_1's value, which is exactly what $\Gamma > 0$ measures. *(2) Partition origin*: CGA's partition is the deterministic output of Alg. 1, while PLE/MMoE's is learned—and in the pathological regime its gate is driven by the dominant gradient itself. *(3) Shared parameters*: their shared experts remain shared parameters where Thms. 1–3b apply in full, and no stop-gradient separates the towers. This yields the falsifiable prediction registered in §4.3: PLE/MMoE should match CGA when $\kappa_g \approx 1, c_H \approx 0$ and degrade toward Flat when $\kappa_g \gg 1, c_H > 0$; this prediction is *tested* by the controlled comparison of §4.3 (full readings in Appendix M, M.4), which reports the realized PLE/MMoE/AdaShare columns rather than leaving them as pending work.

L Appendix L: The CGA Compiler (Alg. 1)

Algorithm 1 CGA compiler: data / causal graph → module partition, detach positions, applicability diagnosis

Require: Data $\mathcal{D} = \{(X, Y_1, \ldots, Y_K)\}$; prior causal graph $\mathcal{G}_0$ (optional).
Ensure: $\mathrm{CGA}(\hat{\mathcal{G}})$ topology, training config, applicability diagnosis.

1: **Causal graph estimation**: if $\mathcal{G}_0$ is absent, estimate $\hat{\mathcal{G}}$ by causal discovery (Granger tests / PC algorithm / conditional independence tests) and output a directionality confidence; else $\hat{\mathcal{G}} \leftarrow \mathcal{G}_0$.
2: **Cycle detection**: if $\hat{\mathcal{G}}$ has cycles, report conflicting edges and warn (delete the lowest-confidence edge and continue; reject only if unresolvable).
3: **Module partition**: topologically sort $\hat{\mathcal{G}}$; for each node i create module M_i with independent θ_i and inputs $\{X\} \cup \{\mathrm{sg}(\hat{Y}_p) : p \in \mathrm{Pa}(i)\}$.
4: **Applicability diagnosis**: estimate $\hat{\Gamma}$ (§5.3); if $\hat{\Gamma} \approx 0$, warn "no cascade headroom; flat is also fine" (Thm. 6); record bullying indicators (cos, κ_g, c_H).
5: **Training configuration**: joint objective $\sum_i \ell_i$ with no task-weight hyperparameter; per-module step sizes; module capacity split under a total budget.

The compiler is deterministic: step 1 estimates $\hat{\mathcal{G}}$ (or accepts a prior $\mathcal{G}_0$) with a directionality confidence, step 2 breaks cycles by deleting the lowest-confidence edge (rejecting only when unresolvable), step 3 topologically sorts and assigns one module per node, step 4 diagnoses applicability, and step 5 emits the training configuration. It is the executable form of *Structure Follows Causality*: an arbitrary causal DAG is compiled to a computational graph with edge-wise gradient decoupling (Props. 4/4′).

M Appendix M: Additional Controlled Experiments (E2–E5)

All four experiments below were run locally on CPU (8 cores, torch 2.9.1); E1 (multi-seed real three-arm) requires a server and is reported separately.

M.1 E2: constructive quadratic SCM, 81 configurations. Setup: $\mathcal{T}_2$'s local model is a quadratic with known curvature ρ; $\mathcal{T}_2$ is initialized at its joint optimum ($\mathbf{g}_2 = 0$, cos exactly 0 at $t = 0$); (a, ρ, η, z_1^0) are scanned. Results: *single-step* pure-curvature erasure in **81/81** configurations (first-order term $= 0.00\mathrm{e}{+00}$; $\Delta\mathcal{L}_2$ equals the pure second-order term, quadratic residual zero); a *sustained* window ($\mathcal{T}_1$ stable and $\mathcal{L}_2$ monotonically rising $\geq$50 steps) in **27/81**. Representative configurations:

Decay law: $E(t) = \dfrac{\rho\,(1-\eta a)^{2t}}{2\eta a\,(\rho^2 + \gamma^2/4)\,\left(1-(1-\eta a)^t\right)^2}$. $E > 1$ only at the measure-zero start; it decays as $\mathcal{T}_1$ converges and $\mathcal{T}_2$ is displaced ($1/\text{displacement}^2$). MLP-side control (four configurations): E peak $1\mathrm{e}{-4}$ / $<1\mathrm{e}{-4}$ / $2\mathrm{e}{-4}$ / $<1\mathrm{e}{-4}$, min cos -0.063 / -0.134 / -0.192 / -0.035, and **zero** steps with $E > 1$ in all four.

Table 8: E2 representative configurations: sustained-erasure window.

a	ρ	η	z_1^0	rising steps	max cos drift	$\Delta\mathcal{L}_2(t_0)$ 1st	$\Delta\mathcal{L}_2(t_0)$ 2nd
0.020	1.0	25.00	-0.05	54	0.707	0.00e+00	3.13e$-$04
0.020	1.0	25.00	-0.10	54	0.707	0.00e+00	1.25e$-$03
0.020	1.0	25.00	-0.20	54	0.707	0.00e+00	5.00e$-$03
0.020	3.0	25.00	-0.05	56	0.949	0.00e+00	9.37e$-$04
0.020	3.0	25.00	-0.10	56	0.949	0.00e+00	3.75e$-$03
0.020	3.0	25.00	-0.20	56	0.949	0.00e+00	1.50e$-$02

Table 9: E3: eight flat variants; T1/T2 final values (mean [range]) and T2 best (mean).

Flat arm	T1 (mean/range)	T2 (mean/range)	T2 best
flat_shared	0.331 [0.114–0.581]	0.242 [0.075–0.475]	0.240
flat_weighted w_2=0.3	0.674 [0.604–0.752]	0.731 [0.610–0.812]	0.707
flat_weighted w_2=1.0	0.734 [0.609–0.884]	0.579 [0.488–0.667]	0.505
flat_weighted w_2=3.0	0.673 [0.545–0.739]	0.423 [0.341–0.498]	0.381
flat_weighted w_2=10.0	0.705 [0.668–0.738]	0.325 [0.306–0.344]	0.269
flat_alt (alternating)	0.282 [0.259–0.307]	0.385 [0.372–0.402]	0.385
flat_clip5 (clipping)	0.888 [0.887–0.890]	**1.000** [1.000–1.001]	1.000
flat_adam	0.435 [0.432–0.440]	0.394 [0.384–0.403]	0.394

M.2 E3: exhaustive flat-side protocols (zoom4, 3 seeds). Increasing w_2 monotonically improves T2 (0.707→0.269) but degrades T1 to 0.67–0.73 (CGA: 0.117); no w_2 improves both.

Table 10: E4: search budget and divergence counts.

Method	grid points	seeds	runs	diverged
GradNorm	12	3	36	0
IMTL-G (original, 6 cells)	6	2	12	12
IMTL-G (clipped $\alpha \in [10^{-2}, 10^{2}]$)	12	2	24	5

M.3 E4: mediator fairness audit. Why the λ_1=1 and λ_1=3 columns coincide. This is a protocol artifact, not a transcription error: GradNorm trains with *dynamic* weights ($g_1 = \mathrm{scale}(g_{1,\mathrm{raw}}, w_1)$), so λ_1 does not enter its training dynamics at all; it enters only through the learning-rate rule $lr = 0.01/\max(1, \lambda_1/3)$, which saturates for $\lambda_1 \leq 3$ ($lr \equiv 0.01$). Hence λ_1=1 and λ_1=3 are the same training run duplicated, and only λ_1=10 ($lr = 0.0033$) is genuinely distinct. We keep both columns for completeness. GradNorm T2 WAPE (3-seed mean): β=0: 0.622/0.622/0.528; β=0.25: 0.578/0.578/0.410; β=0.5: 0.436/0.436/0.380; β=1.0: 0.392/0.392/0.161 (columns $\lambda_1 = 1/3/10$). IMTL-G-clipped T2 WAPE (2-seed mean; NaN counted as divergence): β=0: 1.413/1.192/0.927; β=0.25: 0.671/0.858/0.876; β=0.5: 1.440/0.628/1.392; β=1.0: 0.545/0.317/0.342.

M.4 E5: structured sharing, 12 configurations × 2 seeds. T2 WAPE means: MMoE-4 0.149, PLE 0.142, AdaShare 0.149 (CGA 0.10–0.12, ST2 0.11). **T1 WAPE means: MMoE-4 0.300, PLE 0.300, AdaShare 0.134** (CGA 0.13)—MMoE-4 and PLE pay 2.3× the upstream floor to partially relieve the downstream task. Per-cell T2 (columns $\lambda_1 = 1/3/10$): MMoE-4 β=0: 0.14/0.14/0.27, β=0.25: 0.13/0.13/0.21, β=0.5: 0.12/0.12/0.16, β=1.0: 0.12/0.11/0.13; PLE β=0: 0.13/0.13/0.23, β=0.25: 0.13/0.13/0.19, β=0.5: 0.12/0.12/0.16, β=1.0: 0.12/0.11/0.13; AdaShare β=0: 0.19/0.11/0.13, β=0.25: 0.20/0.17/0.12, β=0.5: 0.18/0.16/0.11, β=1.0: 0.15/0.16/0.11.