BRIDGING THE DATA GAP: SOURCE-INFORMED TASK-SPACE TRANSPORT FOR FEW-SHOT DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many deployments require a model to move rapidly from a data-rich source to a related target observed only sparsely. This is especially consequential where target data are difficult to collect, as in healthcare settings with persistent population or acquisition imbalances. Retraining or fine-tuning a high-capacity generator for every target is costly and can be statistically unreliable when target data are scarce. We introduce *Meta-Diff*, which instead learns a recurring source–target relation once in a compact task space. A shared encoder maps each distribution to a low-dimensional coordinate; a learned relation transports a new source coordinate to its predicted target, and a shared diffusion backbone generates from that prediction without target-specific retraining. Controlled Gaussian-mixture experiments identify when this works: task identity must survive the representation, the relation must generalise across tasks, and the source–target shift must leave useful headroom for transport. On held-out CIFAR-100 validation tasks, correct source + transport achieves 13.9% lower mean SW than transport from the wrong source. In the final-test one-target comparison, its mean SW is 14.7% lower than target-only modelling and 32.7% lower than full fine-tuning; the main route itself uses no target image. Office-Home exposes a complementary boundary: transport can align target-domain statistics without reliable object identity. Together, these results recast extreme-few-shot adaptation as reusable transport in task space, while making the conditions for its success and failure directly measurable.

1 INTRODUCTION

A new deployment rarely arrives with exactly the data one would like. A category may be abundant in one visual domain but scarce in another; an institution, device or population may be well represented while a related target setting contains only a handful of observations. The imbalance is especially consequential in healthcare, where uneven sampling across demographic groups, institutions and acquisition settings can produce systematic subgroup performance gaps (Larrazabal et al., 2020; Seyyed-Kalantari et al., 2021; Daneshjou et al., 2022). Representative target-data collection remains the preferred long-term remedy, but it can be slow, costly or privacy-sensitive for rare settings.

The conventional response is target adaptation. Full fine-tuning, restricted parameter updates and learned embeddings reduce the cost of fitting a generator to a new target (Ruiz et al., 2023; Kumari et al., 2023; Hu et al., 2022; Gal et al., 2023; Ruiz et al., 2024), but they still infer each target largely from target-side evidence. We exploit a different regularity: source and target domains often recur across semantic tasks. If their relationship is structured, earlier source–target pairs can teach us how a new source task should move before target examples become plentiful.

We therefore consider relation-level generalisation. During meta-training we observe

$$P_S^{y_1} \to P_T^{y_1}, \ldots, P_S^{y_n} \to P_T^{y_n},$$

for semantic tasks $y_i \in \mathcal{Y}_{\text{train}}$. At test time the semantic identity changes to an unseen task $y^\star \in \mathcal{Y}_{\text{test}}$, but the source–target relation is shared. The comparison protocol provides a source set $D_S^{y^\star}$ and a target budget $D_T^{y^\star}$ for independent baselines. The main route uses the source set and relation only, at every reported target budget.

Meta-Diff turns this question into low-dimensional transport (Figure 1). A set encoder estimates z_S, a linear relation predicts $\widetilde{z}_T$, and a frozen shared diffusion backbone generates from that prediction without target-time optimisation. Figure 2 instead intervenes on own, mean, shuffled and zero coordinates of the model; it visualises coordinate sensitivity, not the deployment transport comparison.

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.

Task-space transport also makes the mechanism inspectable. Task identity must survive the encoder; the relation must generalise; the shift must leave headroom over source reuse; and the generator must express the transported information. Section 3 establishes two supporting properties, source-coordinate concentration and stability to coordinate error, but not task identity or relational generalisation. Section 5.1 isolates those missing conditions, Section 5.2 tests their generative consequence, and Section 5.3 locates a further decoder-side semantic bottleneck.

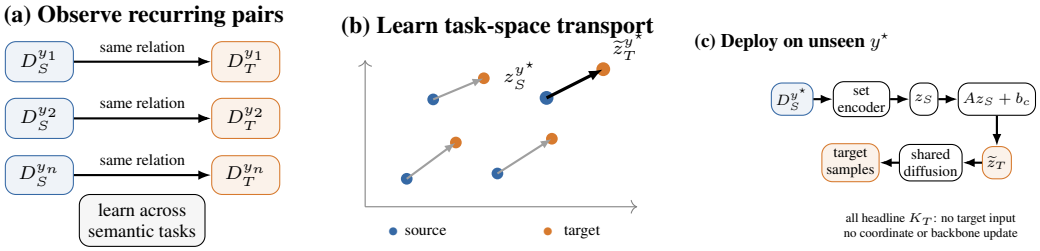

Figure 1: **Meta-Diff learns a relation between task distributions, rather than a separate adaptation for each target.** Meta-training observes the same source→target relation across different semantic tasks; a shared encoder embeds these distributions in task space; and the learned relation is applied to the source coordinate of a held-out task. At every headline K_T, the main route generates directly from the transported coordinate, with no target input, coordinate refinement or backbone update. Target-support budgets apply to independent baselines.

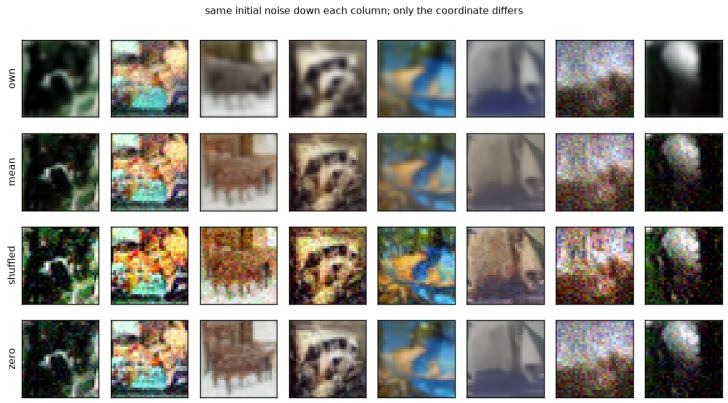

Figure 2: **Qualitative task-coordinate control on held-out CIFAR generations.** Columns share initial diffusion noise; rows differ only in the task coordinate (own, mean, shuffled, or $z = 0$). The samples come from the 50k-step CIFAR checkpoint, with matched tasks and paired noise. Own-task coordinates are encoder coordinates, not transported predictions; deployment claims use the separate paired route evaluations in Section 5.2.

The paper makes four contributions. **First**, we formulate data-asymmetric generation as relation-level meta-learning over a shared source–target transformation. **Second**, a 210-run GMM study maps its operating regime and separates intrinsically recoverable task information from what is represented in z. **Third**, CIFAR-100 tests source-specific generation without target input and the comparison with target-adapted baselines. At one target image for the baselines, SW is 0.1781, 0.2087 and 0.2648 for direct transport, target-only and full fine-tuning, respectively. The paired final-test comparisons support a low-data advantage followed by a crossover to full fine-tuning. FiLM independently reproduces source specificity; the transport-structure ablations test whether nonlinearity is needed. **Finally**, Office-Home extends the analysis to a natural domain shift, where Meta-Diff achieves source-specific target-domain alignment and frozen-checkpoint diagnostics localise the remaining semantic bottleneck to the coordinate-to-generation interface.

2 Related Work

Few-shot diffusion and compact adaptation. Few-Shot Diffusion Models condition on support sets from unseen classes (Giannone et al., 2022); Conditional Distribution Modelling transfers learned conditioning statistics (Gupta et al., 2024); and CRDI obtains sample-wise guidance embeddings

through diffusion inversion (Cao & Gong, 2024). Personalisation methods adapt pretrained text-to-image models through full fine-tuning, compact embeddings, restricted cross-attention updates or amortised hypernetwork prediction (Ruiz et al., 2023; Gal et al., 2023; Kumari et al., 2023; Ruiz et al., 2024). LoRA constrains adaptation to a low-rank parameter space (Hu et al., 2022), while intrinsic-dimension and weight-space studies suggest that useful adaptation can occupy a much smaller space than the full parameter vector (Dravid et al., 2024). These approaches establish that target adaptation can be compressed. Meta-Diff instead amortises the source–target relation across semantic tasks: given a data-rich source for an unseen task, it predicts the corresponding target distribution without requiring target examples or target-specific backbone adaptation.

Meta-learning and task representations. Meta-learning reuses experience across tasks through shared representations or update rules (Finn et al., 2017). Our set encoder follows the permutation-invariant Deep Sets construction (Zaheer et al., 2017); FiLM provides a generic affine mechanism for injecting the resulting coordinate into the denoiser (Perez et al., 2018). Transfer Guided Diffusion Processes likewise combine source diffusion with target-domain guidance (Ouyang et al., 2024). The distinguishing requirement here is *relational generalisation*: semantic identity is held out while the source–target relation has been observed on other tasks. Success therefore depends on the quality of the task representation as well as on adaptation capacity.

Data asymmetry as motivation. Uneven population and institutional coverage can induce subgroup degradation in medical imaging (Larrazabal et al., 2020; Seyyed-Kalantari et al., 2021; Daneshjou et al., 2022), while dataset audits document substantial gaps in demographic coverage and reporting (Daneshjou et al., 2021; Wen et al., 2022). This motivates methods that make use of related abundant data, but synthetic data are not a substitute for representative collection. We therefore treat the clinical studies in Appendix E as stress tests and limitations, not as evidence for clinical deployment.

3 Problem and Method

Data-asymmetric task transfer

Let $y \in \mathcal{Y}$ denote a semantic task and $d \in \{S, T\}$ a source or target domain, with distribution P_d^y. Semantic tasks are partitioned into disjoint $\mathcal{Y}_{\text{train}}, \mathcal{Y}_{\text{val}}, \mathcal{Y}_{\text{test}}$. At deployment, the comparison protocol provides a source support set D_S^y of size M_S and permits independent baselines to use a target support set D_T^y of size K_T; evaluation uses a disjoint target query set. The correct source + transport main route consumes only D_S^y and the relation descriptor for every K_T. We focus on $M_S \gg K_T$. A shared permutation-invariant encoder maps any support set to a task coordinate,

$$z = r_\psi(D) = \rho_\psi\left(\frac{1}{|D|}\sum_{x\in D} h_\psi(x)\right). \tag{1}$$

The same encoder places source and target distributions in a common task space. Transfer therefore becomes a prediction problem in that space rather than a new high-dimensional optimisation problem for every target.

Conditioning and transport

The coordinate must next control a shared generator. We use standard DDPM noise prediction (Ho et al., 2020), with the continuous-time score interpretation of Song et al. (2021). The additive-basis parameterisation writes

$$\widehat{\epsilon}_z(x_t, t) = \widehat{\epsilon}_0(x_t, t) + \sum_{\ell=1}^{k} z_\ell R_\ell(x_t, t), \tag{2}$$

where R_ℓ are learned residual noise-prediction fields. To test whether the task-space effect depends on this construction, we also use FiLM. At U-Net block j, the feature tensor $H^{(j)}$ is modulated as

$$H^{(j)} \leftarrow (1 + \gamma_j(z)) \odot H^{(j)} + \beta_j(z),$$

with learned projections γ_j, β_j from the task coordinate to feature-wise scales and shifts.

For a source set, $z_S = r_\psi(D_S^y)$. We compare identity, relation-only offsets, residual MLPs and linear maps; the experiments show that source dependence matters but non-linearity is not required. The final transport is therefore

$$\widetilde{z}_T = A z_S + b_{c_{S\to T}}, \tag{3}$$

where the relation-specific bias reduces to a single b when only one relation is present. Across all headline K_T, the main route directly decodes $\widetilde{z}_T$ with $J = 0$ target-time coordinate-refinement steps and no backbone update. Target-only and full-fine-tuning baselines use their stated target-support budgets. Strict source reuse decodes z_S directly; source-initialised target adaptation is named separately. The matched refinement ablation in Appendix C.3 tests a separate intervention and does not imply that refinement is universally ineffective.

3.1 Meta-training and low-data bias–variance trade-off

Each training episode contains source and target support/query sets for one semantic task. With $z_S = r_\psi(D_S^s)$, $z_T = r_\psi(D_T^s)$ and $\widetilde{z}_T$ from Equation (3), we minimise

$$\mathcal{L}_{\text{meta}} = \mathcal{L}_S(z_S) + \lambda_T \mathcal{L}_T(z_T) + \lambda_{\text{tr}} \mathcal{L}_T(\widetilde{z}_T) + \lambda_z(\|z_S\|_2^2 + \|z_T\|_2^2), \tag{4}$$

where each $\mathcal{L}$ is the standard diffusion denoising loss evaluated on a disjoint query set. Target supports and queries supervise meta-training through separate branches. The transport map takes only the source coordinate and relation; at deployment the main route reads no target support or query and updates neither coordinates nor backbone weights.

Why task-space transfer can help in the low-data regime. A compact coordinate is useful only if abundant source data estimate it reliably and transport errors do not destabilise generation. Two results formalise these requirements.

Proposition 1 (Source-coordinate concentration)**.** *For i.i.d. $X_1, \dots, X_M$, let $\mu = \mathbb{E}h_\psi(X)$ and $z_M = \rho_\psi(M^{-1}\sum_i h_\psi(X_i))$. If $\mathbb{E}\|h_\psi(X)\|_2^2 < \infty$ and ρ_ψ is L_ρ-Lipschitz, then*

$$\mathbb{E}\|z_M - \rho_\psi(\mu)\|_2^2 \le \frac{L_\rho^2 \operatorname{tr}(\operatorname{Cov}[h_\psi(X)])}{M}.$$

Thus source-coordinate estimation error decays at the usual $O(M^{-1})$ mean-squared rate. This motivates the CIFAR source-budget check, but does not establish a generated-distance plateau at any particular M_S.

Proposition 2 (Stability to coordinate perturbations)**.** *For the reverse SDE with drift $b_z(x,t) = f(x,t) - g(t)^2[s_0(x,t) + B(x,t)z]$, assume b_z is uniformly L_x-Lipschitz in x and $\sup_{x,t}\|g(t)^2 B(x,t)\|_{\text{op}} \le L_z$. Under synchronous coupling,*

$$\mathbb{E}\sup_{t\le T}\|X_t^z - X_t^{z'}\|_2^2 \le C_T\|z - z'\|_2^2, \qquad W_2(P_z, P_{z'}) \le \sqrt{C_T}\|z - z'\|_2,$$

where $C_T = L_z^2 \phi_{L_x}(T)^2$ and $\phi_{L_x}(T) = (e^{L_x T} - 1)/L_x$ (with value T at $L_x = 0$).

Together, the propositions capture the intended low-data bias: source evidence can make z_S precise, while transport error cannot be amplified arbitrarily by the reverse process. They do not establish semantic task identity, relation predictability on unseen tasks, or semantic use by the decoder. Section 5.1 tests those missing conditions; full proofs are in Appendix A.

4 Experimental Design

The propositions establish estimation and stability, not content or relational generalisation. Experiments therefore progress from GMMs, where these missing quantities are observable, to CIFAR under controlled image shifts and finally to the natural Product→Clipart relation in Office-Home. Headline comparisons are paired on semantic task, support draw, target query set and sampling noise where possible. Generation repeats are not independent semantic tasks. CIFAR uses mean ± 1.96 SE convention over matched class-by-transformation episodes (60 / target budget), with 1 training seed per arm; 3 transformations share each class, so these are not class-cluster or training-seed intervals. Other benchmarks retain their separately specified statistical units. Hyperparameters for target-only adaptation and fine-tuning are selected on validation tasks and frozen before final test evaluation.

Controlled GMM boundary study. Two-dimensional Gaussian-mixture tasks vary source-task separation through $\alpha \in \{0, .1, .25, .5, .75, 1, 1.5\}$ and relation heterogeneity through $\eta \in \{0, .1, .25, .5, 1, 2\}$. The full 7×6 grid is repeated over five seeds, giving 210 trained cells and 13,440 held-out task evaluations at $K_T = 0$. A separate shift-magnitude sweep at $\alpha = 1, \eta = 0$ contributes a further 25 cells and 1,600 held-out task evaluations. The primary effect is

$$G_{\text{source}} = \text{SW}_{\text{wrong}} - \text{SW}_{\text{correct}},$$

so positive values favour the correct source. Because the data-generating distributions are known, intrinsic Bayes task identification can be measured separately from task identification in the learned coordinate and from held-out affine relation predictability.

CIFAR-100. Fine classes are semantic tasks (Krizhevsky, 2009). Entire superclasses are held out: 60 fine classes from 12 superclasses for meta-training, 20 from four superclasses for validation, and 20 from four disjoint superclasses for final testing. Sources are clean; targets apply Gaussian blur, Gaussian noise or contrast reduction at fixed severity (Hendrycks & Dietterich, 2019). Source and target pools contain disjoint underlying images. We test zero-shot source controls and $K_T \in \{1, 5, 20\}$ against target-only modelling and validation-selected full fine-tuning. Sliced Wasserstein distance (SW) is the headline distribution metric; energy MMD and semantic consistency are complementary. The appendix retains separate GMM denoising-calibration and historical oracle diagnostics; these are not pooled with the image-generation metrics.

Office-Home. Product→Clipart provides a natural relation across 65 object categories. The fixed semantic split is 45/10/10 train/validation/test classes. Real-image audits verify both the semantic evaluator and the Product–Clipart domain distinction before generated images are scored. We first evaluate the same source controls used on CIFAR, then use frozen-checkpoint diagnostics to localise whether weak semantics arise in the encoder, transport or coordinate-to-generation path.

5 Results

5.1 GMM: a controlled operating regime

The analytical results guarantee a stable coordinate, not a useful one. GMMs make the missing ingredients observable: we vary source-task separation α and relation heterogeneity η independently, then evaluate zero-target transfer over the resulting 7×6 grid. Figure 3 shows the first two conclusions. When $\alpha = 0$, correct and wrong sources are unresolved in all six relation conditions; once the source carries task information, a correct-source advantage emerges, but it decreases as the relation becomes less reusable. In total, 34 of 42 cells have positive paired intervals excluding zero, with the unresolved cells confined to the six $\alpha = 0$ conditions and the two most heterogeneous relations at $\alpha = .1$.

The same experiment distinguishes *intrinsic* from *usable* identifiability. At $\alpha = .25$, the Bayes classifier on the true GMM likelihoods already identifies the source task with accuracy .953, yet nearest-centroid identification in the learned coordinate is only .323. At $\alpha = 1.5$, the corresponding values are 1.000 and .920. This gap is not cosmetic: across the 42 cells, G_{source} correlates much more strongly with learned-ID than with Bayes-ID ($\rho = .98$ versus .88). Replacing Bayes-ID with learned-ID in the same descriptive interaction model raises the cell-level fit from $R^2 = .587$ to .850; even within the 24 cells with Bayes-ID $> .99$, learned-ID remains strongly associated with the source-specific gain ($\rho = .97$). We treat these regressions as descriptive summaries of the factorial experiment rather than causal estimates.

Relation predictability is a separate requirement. Along $\alpha = 1$, Bayes-ID remains at 1.0 while increasing η from 0 to 2 lowers held-out affine relation R^2 from 1.000 to .408 and reduces the correct-over-wrong SW advantage from .523 to .229. Across the full grid, $\rho(G_{\text{source}}, R^2_{\text{rel}}) = .53$. Thus source identifiability alone is insufficient: even when the source task is intrinsically unambiguous, source-specific transfer weakens as the relation becomes less reusable. Full diagnostics are in Appendix B.

The coordinate-retrieval analysis makes information loss along the pipeline explicit. For $\alpha \geq .5$, intrinsic identity is perfect, yet learned-ID ranges from .691 to .920 and transported target-coordinate retrieval from only .169 to .341 (chance $1/64$). Averaged over the grid, transported retrieval is .172 versus .061 for the raw cross-domain $z_S \to z_T$ baseline; the transport gain is resolved in every cell with $\alpha \geq .25$ (+.112 overall, paired seed-block interval $[+.106, +.116]$) and correctly vanishes at $\alpha \in \{0, .1\}$. Across the grid it correlates with G_{source} at $\rho = .91$, between Bayes-ID (.88) and learned-ID (.98). Retrieval margins show that the recovery is partial: the learned-z margin rises from $-.002$ to $+.060$ across the sweep, whereas the transported margin remains negative ($-.039$ to $-.079$). Figure 4a and Appendix B report the full identity chain.

The shift-magnitude sweep in Figure 4b isolates a different question: *when is transport itself needed?* We fix $\alpha = 1, \eta = 0$ and interpolate the same source–target transformation by λ. The correct-source advantage over a wrong source remains large even at $\lambda = 0$ ($G_{\text{source}} = .482$), because source identity

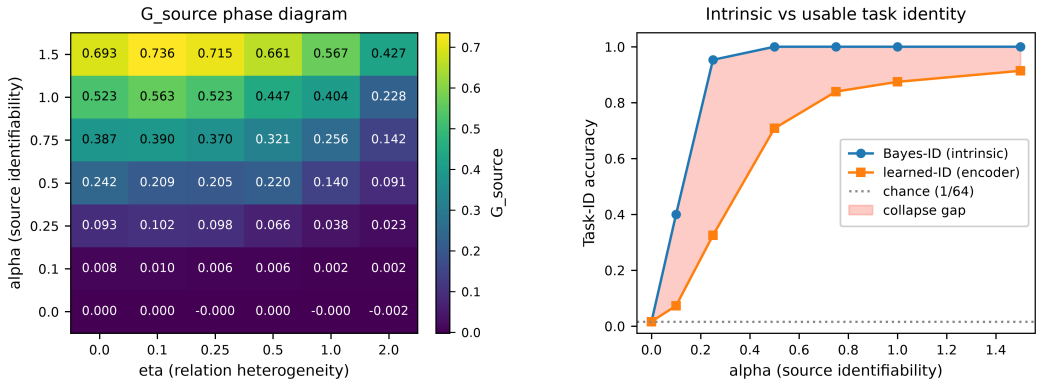

Figure 3: **GMM operating regime.** *Left:* correct-source advantage $G_{\text{source}} = \text{SW}_{\text{wrong}} - \text{SW}_{\text{correct}}$ over the full source-identifiability/relation-heterogeneity grid at $K_T = 0$. The advantage disappears when source identity is absent and attenuates as the relation becomes less predictable. *Right:* intrinsic Bayes task identification and nearest-centroid identification in the learned encoder coordinate, averaged over relation levels. Nearly perfect intrinsic identifiability can coexist with substantially weaker usable identity. Full cell-level results and paired intervals are in Appendix B.

is still recoverable. In contrast, transport improves over direct source reuse by an unresolved .024 SW when there is essentially no shift to correct, but by .541 at $\lambda = 2$; the independently measured source–target headroom increases from .08 to 1.59 SW over the same sweep. Source specificity and transport utility are therefore different quantities: identifying the right source is not enough to make a transport map useful if source and target are already nearly the same.

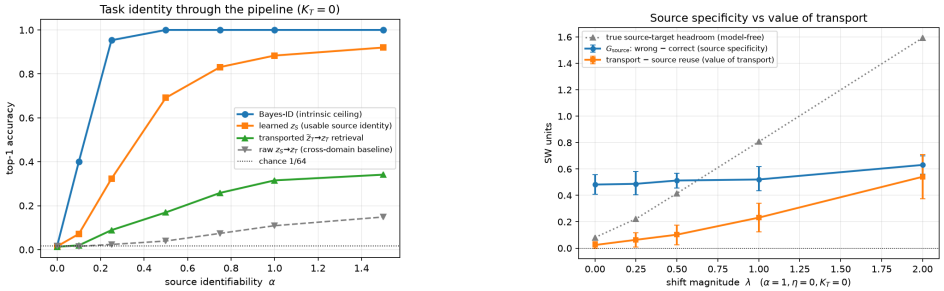

Figure 4: **Two additional GMM boundaries.** *Left:* task identity through the pipeline. Bayes-ID is the intrinsic ceiling, learned z_S measures usable source identity, and transported $\widetilde{z}_T$ measures whether the predicted target coordinate retrieves the correct held-out target task; raw $z_S \to z_T$ retrieval is a cross-domain baseline. *Right:* the shift-magnitude sweep separates source specificity from the value of transport, with hierarchical seed/task 95% intervals. The correct source remains informative as $\lambda \to 0$, whereas the advantage of transport over source reuse grows only when there is nontrivial source–target headroom. Exact tables, regression diagnostics and uncertainty are in Appendix B.

The controlled study therefore adds four empirical requirements beyond the two analytical guarantees: task identity must survive the encoder, the relation must generalise, transport must preserve enough identity to reach the target task, and the source–target discrepancy must leave useful headroom over source reuse. CIFAR now asks whether this mechanism survives when the distributions are images rather than known mixtures.

5.2 CIFAR: SOURCE-SPECIFIC ZERO-SHOT TRANSFER AND A LOW-DATA CROSSOVER

We first evaluate generation without any target input on the 60 held-out validation class–transformation episodes. Correct source + transport reaches SW 0.2061, versus 0.2393 for wrong-source transport and 0.2407 for relation-only prediction (Table 1). Their paired SW and semantic-consistency contrasts favour the correct source; the semantic difference from strict source reuse remains unresolved. Frozen-coordinate diagnostics point in the same direction before generation: transported target-coordinate retrieval reaches 62.5%, 59.5% and 49.0% top-1 for blur, noise and contrast, versus 17.3%, 26.5%

and 14.0% for the raw source coordinate and at most 2.25% for wrong-source transport (Appendix C). A separate $K_T = 1$ deployment grid from the model checkpoint is shown in Figure 10; it retains six labelled routes and paired generation noise. This visual diagnostic is distinct from the zero-target and frozen-coordinate packages.

On the separate final-test episodes, one-target SW is 0.1781 for correct source + transport, 0.2087 for target-only and 0.2648 for full fine-tuning. At $K_T = 20$, the corresponding transport and full-FT values are 0.1844 and 0.1483. Paired transport-versus-target-only intervals favour transport at $K_T = 1$; remain unresolved at $K_T = 5, 20$; transport-versus-full-FT intervals favour transport at $K_T = 1, 5$; favour full fine-tuning at $K_T = 20$. The paired final-test comparisons support a low-data advantage followed by a crossover to full fine-tuning. Energy MMD and its paired contrasts are reported separately in Appendix C. Full-FT settings were selected on validation tasks and frozen before final-test evaluation. The appendix also reports the matched refinement on/off experiment on the correct source + transport checkpoint; it is a separate diagnostic, not part of the headline route.

Table 1: **CIFAR headline comparisons for correct source + transport, $J = 0$.** Panel A uses validation classes; Panel B uses separate final-test classes. Distances: $\Delta = \text{comparator} - \text{main}$; consistency: $\Delta = \text{main} - \text{comparator}$. Positive favours the main route. Entries show episode-level 95% half-widths; stars use unrounded paired intervals excluding zero, in either direction.

A. Zero-target controls: 60 validation episodes

Route	SW	Δ SW	Class consistency	Δ class
correct source + transport	0.2061 ± 0.0194	–	0.1075 ± 0.0235	–
Strict source reuse	0.2416 ± 0.0190	$0.0355 \pm 0.0108^*$	0.1102 ± 0.0254	-0.0028 ± 0.0109
Wrong source	0.2393 ± 0.0205	$0.0333 \pm 0.0113^*$	0.0470 ± 0.0079	$0.0604 \pm 0.0216^*$
Relation-only	0.2407 ± 0.0197	$0.0346 \pm 0.0081^*$	0.0490 ± 0.0094	$0.0585 \pm 0.0206^*$

B. Final-test SW: 180 episodes, 60 per budget

Route	$K_T = 1$	$K_T = 5$	$K_T = 20$
correct source + transport	0.1781 ± 0.0116	0.1834 ± 0.0107	0.1844 ± 0.0118
Target-only	0.2087 ± 0.0157	0.1884 ± 0.0128	0.1799 ± 0.0099
Full fine-tuning	0.2648 ± 0.0184	0.2205 ± 0.0183	0.1483 ± 0.0098
Δ target-only	$0.0306 \pm 0.0105^*$	0.0051 ± 0.0058	-0.0044 ± 0.0067
Δ full FT	$0.0867 \pm 0.0176^*$	$0.0371 \pm 0.0197^*$	$-0.0361 \pm 0.0147^*$

Two further results distinguish source use from architecture. FiLM correct source + transport improves SW and class consistency over mean-source, shuffled-source and relation-only controls at all three budgets in its own validation package. This tests FiLM source specificity, not a complete FiLM full-FT crossover. The matched conditioning comparison favours FiLM at $K_T = 5, 20$, with the other budgets unresolved. In the separate retained transport-structure ablations, Basis linear versus residual MLP remains unresolved, while FiLM linear has lower SW and energy MMD than FiLM nonlinear at all three budgets. These one-seed matched comparisons do not quantify training-seed variation. The basis uses 55,344 conditioning parameters versus 118,784 for FiLM; this does not establish a runtime or memory advantage.

The source-budget sweep shows small changes between $M_S = 16$ and 300, but not exact equivalence: the paired SW difference at $M_S = 64$ relative to 16 is $-0.00287 \pm 0.00236^*$. Other measured increments relative to 16 have intervals containing zero; no $M_S < 16$ was tested. Figure 5 uses the actual final-test target-budget and nested source-budget records. Appendix C contains the validation source-dependence sweep, final-test MMD, discussion of the source-budget sweep, FiLM controls, frozen adaptation settings, the separate refinement ablation, and conditioning/task-scarcity results.

Table 2: **Three separate matched CIFAR packages.** The first row uses the two conditioning runs of correct source + transport; the other rows compare transport structures. Each evaluated transport route uses $J = 0$. Signs follow the first column; stars use unrounded paired episode intervals. One training seed per arm.

Paired SW comparison	$K_T = 1$	$K_T = 5$	$K_T = 20$
FiLM − basis	-0.0094 ± 0.0099	$-0.0148 \pm 0.0107^*$	$-0.0149 \pm 0.0107^*$
Linear − MLP (basis)	0.0022 ± 0.0049	0.0004 ± 0.0048	0.0036 ± 0.0050
MLP − linear (FiLM)	$0.0093 \pm 0.0037^*$	$0.0112 \pm 0.0040^*$	$0.0087 \pm 0.0042^*$

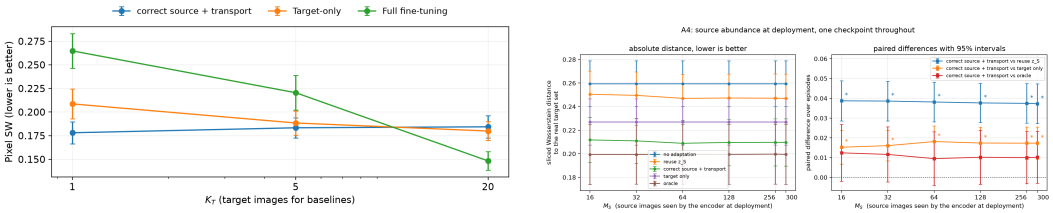

Figure 5: **CIFAR target and source budgets.** Left: final-test SW for correct source + transport and target-adapted baselines. The main route uses no target support at any K_T; its small changes across budgets reflect different sampled source/query episodes and generation noise. Right: $K_T = 1$ with nested source supports and the checkpoint trained at $M_S = 64$. Error bars show marginal episode-level 95% intervals; paired effects and stars are computed separately.

5.3 Office-Home: natural-domain alignment and decoder-side bottleneck

CIFAR establishes source-specific generation under controlled shifts. Office-Home asks whether the distributional effect persists under a natural Product→Clipart shift. Before scoring generated images, the benchmark passes the intended structural checks: CLIP reaches 65.96% top-1 on held-out real Clipart images, DINOv2 distinguishes Product from Clipart with 98.30% accuracy, and same-class cross-domain centroid retrieval is 100% top-1 with median rank 1. Weak generated semantics therefore cannot be dismissed as a broken real-image evaluator.

Office-Home yields a positive source-specific domain-alignment result while exposing a distinct semantic bottleneck. At $K_T = 0$, correct-source transport improves DINO-SW over wrong-source transport (.038969 vs. .039394), relation-only prediction (.039251) and source reuse (.041641); all three paired distance gains exclude zero. It also raises predicted Clipart share from 63.8% under source reuse to 86.5%. Yet requested-object identity is only 1.67%, and its differences from wrong-source and relation-only controls are unresolved. Figure 6 summarises this separation: with one target example, transport beats target-only in DINO-SW and MMD but is unresolved against full fine-tuning in SW; by $K_T = 5$ and 10, full fine-tuning is clearly better.

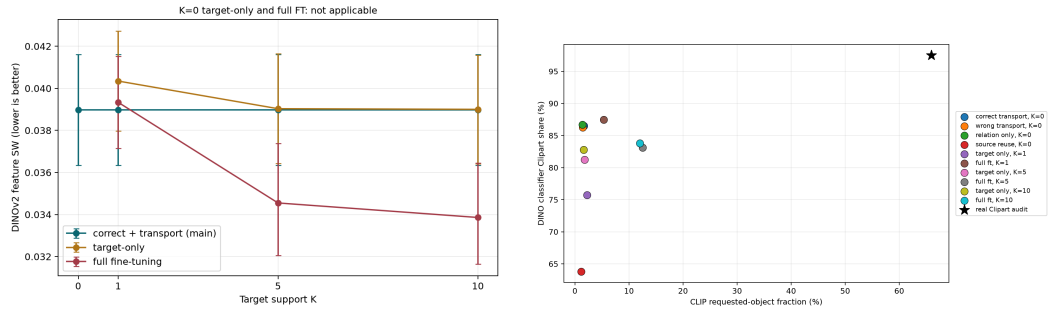

Figure 6: **Office-Home exposes a different boundary from CIFAR.** *Left:* correct-source transport remains nearly flat in DINO feature-SW, while full fine-tuning improves as target support grows. *Right:* generated routes can become strongly Clipart-like without preserving the requested object; the star is the held-out real Clipart audit. The natural-domain limitation is therefore semantic, not simply a failure to change domain appearance.

The distributional gain without reliable object identity creates a localisation question: is semantic information lost in the encoder, the transport map, or the decoder? Frozen-checkpoint diagnostics separate these stages. Independent 16-image Product support sets retrieve the matching held-out class in learned z with 91.0% top-1 accuracy, so the encoder has substantial within-domain object information. Transport improves cross-domain target-coordinate retrieval from 12.0% to 42.0% L2 top-1, although the correct-versus-wrong transport interval remains unresolved over the ten test classes. More decisively, bypassing transport does not restore semantics: a direct target coordinate estimated from abundant Clipart support produces only 2.38% requested-class accuracy on seen classes and 1.56% on held-out classes, despite real-query accuracies around 50%. Correct coordinates do reduce denoising MSE slightly across diffusion time, so the generator is not coordinate-blind, but the effect is weak. An explicit learned class embedding raises generated identity to 6.90% on seen classes, still far below the real-image calibration. Taken together, these diagnostics point primarily to the coordinate-

to-generation interface as the semantic bottleneck, possibly compounded by incomplete semantic encoding, rather than to transport error alone. Full route matrices, frozen-checkpoint diagnostics, and the validation-only checkpoint/baseline selection record are in Appendix D (Table 28).

6 Mechanistic Synthesis and Limitations

Theory and experiments form one diagnostic chain. Propositions 1–2 explain why task-space transfer can be stable; Table 3 shows what else makes it useful. GMMs isolate usable identity, relation predictability, transport preservation and headroom; CIFAR tests its image-generation consequences through paired comparisons; Office-Home adds the final requirement that the generator actually use the transported semantics.

Table 3: **Stages of source-informed transfer.** Representative diagnostics identify conditions enabling transfer and the stage that becomes limiting; values are not intended as a common metric across rows.

Setting	Representation / transport evidence	Generative implication
GMM	Bayes-ID can be perfect while learned-ID is weaker; at $\alpha = 1$, learned-ID $.883$ and transported retrieval $.315$. Shift sweep: transport-over-reuse $.024 \rightarrow .541$ as headroom grows.	Correct-vs-wrong SW gain $.573$ at $(1, 0)$; usable identity tracks the gain more closely than Bayes-ID
CIFAR	Transported target-coordinate retrieval 49–62.5% across relations	Correct source $+$ transport improves on wrong/relation-only at $K_T = 0$; the separate final-test budget comparison is in Table 1
Office-Home	Product-z retrieval 91.0%; transport raises target-coordinate retrieval $12.0\% \rightarrow 42.0\%$	Clipart share 86.5%; low object ID persists even with abundant direct z_T, localising the semantic bottleneck downstream of transport.

This comparison suggests a practical diagnostic hierarchy for source-informed generative transfer. Before training a new route, one can ask whether source tasks are separable, whether the learned coordinate retains that information, whether the source–target relation generalises, whether there is enough headroom for transport to improve on source reuse, and whether the generator uses the resulting coordinate semantically. The GMM study isolates the first four under known ground truth; Office-Home identifies the final generator-side semantic bottleneck. These are operating diagnostics rather than interchangeable post-hoc explanations.

Source specificity survives FiLM under the matched source-control protocol, while the transport-structure ablations retain their own comparison scope. At $k = 16$, a single-relation affine map needs 272 active parameters; the three-relation CIFAR implementation uses a shared matrix plus relation-specific biases (304 active parameters) and performs no deployment update. Its source-budget changes over 16–300 images are small but do not establish a precise plateau at 16. The paired final-test comparisons support a low-data advantage followed by a crossover to full fine-tuning.

Limitations. CIFAR provides positive source-specific image evidence under controlled shifts, conditional on the tested checkpoints and episode-level uncertainty; Office-Home achieves target-domain alignment without reliable object-specific generation. Propositions 1–2, proved in Appendix A, establish concentration and continuity only, not relational generalisation. The clinical studies in Appendix E are retained as boundary analyses and do not establish source-specific clinical transfer.

7 Conclusion

Meta-Diff asks whether a recurring source–target relation can replace repeated high-capacity adaptation when target data are scarce. The theory explains why source coordinates can be estimated stably and why transport error need not be amplified by generation; GMMs identify the additional requirements of usable task identity, relation predictability and headroom. CIFAR supports source-specific generation without target input and quantifies the budget-dependent comparison with target adaptation, while Office-Home exposes a separate decoder-side semantic bottleneck. Together, these results establish a measurable operating regime for source-informed task-space transport and make its success conditions and semantic bottlenecks directly inspectable.

AI USE STATEMENT

We used generative artificial intelligence (AI) tools for language editing, drafting support, literature search, and to support some implementation and analysis tasks. The authors reviewed all AI-assisted text, technical claims, derivations, experimental specifications, analyses, and code, and take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

The paper specifies the task-space encoder, conditioning mechanisms, source–target transport map, meta-training objective, evaluation protocols, baselines, and ablations. Full proofs are in Appendix A; complete GMM diagnostics, CIFAR-100 controls and ablations, Office-Home experiments, and clinical stress tests are in Appendices B, C, D, and E. Implementation details, data splits, adaptation settings, parameter counts, statistical units, and checkpoint-selection procedures are also reported in the appendix.

REFERENCES

Yu Cao and Shaogang Gong. Few-shot image generation by conditional relaxing diffusion inversion. In *European Conference on Computer Vision*, pp. 20–37, 2024. doi: 10.1007/978-3-031-72907-2_2.

Roxana Daneshjou, Mary P. Smith, Mary D. Sun, Veronica Rotemberg, and James Zou. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. *JAMA Dermatology*, 157(11):1362–1369, 2021. doi: 10.1001/jamadermatol.2021.3129.

Roxana Daneshjou, Kailas Vodrahalli, Roberto A. Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M. Swetter, Elizabeth E. Bailey, Olivier Gevaert, Pritam Mukherjee, Michelle Phung, Kiana Yekrang, Bradley Fong, Rachna Sahasrabudhe, Johan A. C. Allerup, Utako Okata-Karigane, James Zou, and Albert S. Chiou. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science Advances*, 8(32):eabq6147, 2022. doi: 10.1126/sciadv.abq6147.

Amil Dravid, Yossi Gandelsman, Kuan-Chieh Wang, Rameen Abdal, Gordon Wetzstein, Alexei A. Efros, and Kfir Aberman. Interpreting the weight space of customized diffusion models. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=DAO2BFzMfy.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1126–1135, 2017.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NAQvF08TcyG.

Giorgio Giannone, Didrik Nielsen, and Ole Winther. Few-shot diffusion models. *arXiv preprint arXiv:2205.15463*, 2022.

Parul Gupta, Munawar Hayat, Abhinav Dhall, and Thanh-Toan Do. Conditional distribution modelling for few-shot image synthesis with diffusion models. In *Proceedings of the Asian Conference on Computer Vision*, pp. 818–834, 2024.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023. doi: 10.1109/CVPR52729.2023.00192.

Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020. doi: 10.1073/pnas.1919012117.

Yidong Ouyang, Liyan Xie, Hongyuan Zha, and Guang Cheng. Transfer learning for diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136962–136989, 2024.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023. doi: 10.1109/CVPR52729.2023.02155.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6527–6536, 2024.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27:2176–2182, 2021. doi: 10.1038/s41591-021-01595-0.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

David Wen, Saad M. Khan, Antonio Ji Xu, Hussein Ibrahim, Luke Smith, Jose Caballero, Luis Zepeda, Carlos de Blas Perez, Alastair K. Denniston, Xiaoxuan Liu, and Rubeta N. Matin. Characteristics of publicly available skin cancer image datasets: A systematic review. *The Lancet Digital Health*, 4(1):e64–e74, 2022. doi: 10.1016/S2589-7500(21)00252-1.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3391–3401, 2017.

IMPACT STATEMENT

This work studies generative modelling under uneven data availability. Source-informed generation may reduce target-specific optimisation where representative target data are difficult to obtain, but synthetic data can also reproduce existing bias or create misleading detail. We therefore evaluate against held-out real data, retain negative stress tests, and do not treat generated clinical images as substitutes for representative collection or clinical validation. Any healthcare deployment would require independent clinical evaluation, governance and privacy review beyond the scope of this work.

A Proofs for the Main-Text Guarantees

The guarantees in Section 3 justify source-coordinate estimation and stability with respect to coordinate error only. They do not establish low-dimensionality, semantic sufficiency, task identifiability, or cross-task relational generalisation.

A.1 Proof of Proposition 1

Let $Y_i = h_\psi(X_i) - \mu$. Then $\mathbb{E}[Y_i] = 0$, the Y_i are independent, and

$$\bar{h}_M - \mu = \frac{1}{M}\sum_{i=1}^{M} Y_i.$$

By the L_ρ-Lipschitz property of ρ_ψ,

$$\|z_M - \rho_\psi(\mu)\|_2^2 \le L_\rho^2 \|\bar{h}_M - \mu\|_2^2.$$

Taking expectations and expanding,

$$\begin{aligned}\mathbb{E}\|\bar{h}_M - \mu\|_2^2 &= \frac{1}{M^2}\mathbb{E}\left\|\sum_{i=1}^{M} Y_i\right\|_2^2 \\ &= \frac{1}{M^2}\left[\sum_{i=1}^{M}\mathbb{E}\|Y_i\|_2^2 + 2\sum_{i<j}\mathbb{E}\langle Y_i, Y_j\rangle\right].\end{aligned}$$

For $i \neq j$, independence and zero mean imply

$$\mathbb{E}\langle Y_i, Y_j\rangle = \langle \mathbb{E}Y_i, \mathbb{E}Y_j\rangle = 0.$$

Hence

$$\mathbb{E}\|\bar{h}_M - \mu\|_2^2 = \frac{1}{M}\mathbb{E}\|h_\psi(X) - \mu\|_2^2 = \frac{\operatorname{tr}(\operatorname{Cov}[h_\psi(X)])}{M}.$$

Combining the last two displays yields

$$\mathbb{E}\|z_M - \rho_\psi(\mu)\|_2^2 \le \frac{L_\rho^2 \operatorname{tr}(\operatorname{Cov}[h_\psi(X)])}{M}.$$

This proves Proposition 1. The result controls mean-squared estimation error around the population feature coordinate; it does not imply that this coordinate is semantically sufficient.

A.2 Proof of Proposition 2

Let X_t^z and $X_t^{z'}$ solve the reverse SDE with the same initial random variable and the same Brownian path, and write

$$\Delta_t = X_t^z - X_t^{z'}.$$

Because the diffusion coefficient does not depend on z, the stochastic increments cancel under this synchronous coupling. With $\Delta_0 = 0$,

$$\Delta_t = \int_0^t \left[b_z(X_s^z, s) - b_{z'}(X_s^{z'}, s)\right] ds.$$

Adding and subtracting $b_z(X_s^{z'}, s)$ gives

$$\begin{aligned}\|\Delta_t\|_2 \le &\int_0^t \|b_z(X_s^z, s) - b_z(X_s^{z'}, s)\|_2 \, ds \\ &+ \int_0^t \|b_z(X_s^{z'}, s) - b_{z'}(X_s^{z'}, s)\|_2 \, ds.\end{aligned}$$

The state-Lipschitz assumption bounds the first integrand by $L_x\|\Delta_s\|_2$. For the second,

$$b_z(x, s) - b_{z'}(x, s) = -g(s)^2 B(x, s)(z - z'),$$

so

$$\|b_z(x, s) - b_{z'}(x, s)\|_2 \le L_z \|z - z'\|_2.$$

Therefore

$$\|\Delta_t\|_2 \le L_x \int_0^t \|\Delta_s\|_2 \, ds + L_z t \|z - z'\|_2.$$

The integral form of Grönwall's inequality yields

$$\|\Delta_t\|_2 \le L_z \|z - z'\|_2 \int_0^t e^{L_x(t-s)} \, ds = L_z \phi_{L_x}(t) \|z - z'\|_2,$$

where

$$\phi_{L_x}(t) = \begin{cases} (e^{L_x t} - 1)/L_x, & L_x > 0, \\ t, & L_x = 0. \end{cases}$$

Since ϕ_{L_x} is nondecreasing,

$$\sup_{t \le T} \|\Delta_t\|_2 \le L_z \phi_{L_x}(T) \|z - z'\|_2.$$

Squaring and taking expectations proves the first claim with $C_T = L_z^2 \phi_{L_x}(T)^2$.

For the Wasserstein statement, the synchronous construction is a valid coupling of the endpoint laws P_z and $P_{z'}$. Hence

$$W_2(P_z, P_{z'})^2 \le \mathbb{E}\|X_T^z - X_T^{z'}\|_2^2 \le C_T \|z - z'\|_2^2,$$

and taking square roots completes the proof.

The proposition establishes continuity of the generated law with respect to z under the stated regularity assumptions and a common initial law coupled identically. Its displayed drift uses additive score conditioning. A general FiLM-conditioned drift would require a corresponding uniform Lipschitz bound in z as an additional assumption; the FiLM experiments alone do not establish that bound. It does not imply that the learned coordinate is low-dimensional, task-identifying, or transferable across unseen semantic tasks.

B Gaussian-Mixture Experiments

B.1 Coordinate calibration

Five separate 2-D Gaussian-mixture families check whether the learned generator responds to task coordinates. We evaluate calibration on each corresponding 80k-step checkpoint using 32 held-out tasks, $K_T = 5$ and eight noise draws per task. The mean-minus-own loss interval is positive in 5 of the five families. These conditioning diagnostics remain separate from the factorial boundary study.

Table 4: GMM calibration on five separate 80k-step checkpoints. Each uses 32 held-out tasks; eight noise draws are averaged within task. $\Delta_{\text{task}} = L(\bar{z}) - L(z_y)$ has a paired task-level 95% half-width. This is a coordinate intervention, not a deployment comparison.

Task family	Own	Mean	Δ_{task}
Perturbation .10	0.4398	0.5956	$0.1558 \pm 0.0286^*$
Perturbation .25	0.4912	0.6125	$0.1212 \pm 0.0232^*$
Perturbation .50	0.5847	0.6729	$0.0882 \pm 0.0272^*$
Perturbation 1.00	0.6684	0.7591	$0.0907 \pm 0.0423^*$
Unrelated	0.5427	0.6386	$0.0959 \pm 0.0341^*$

Figure 7 has two different checkpoint sources: the GMM panel uses the 80k-step perturbation-0.1 checkpoint, and the CIFAR panel uses the 50k-step `cifar_ds3` checkpoint. They interpolate encoder coordinates and diagnose conditioning; neither panel directly measures correct-source deployment superiority.

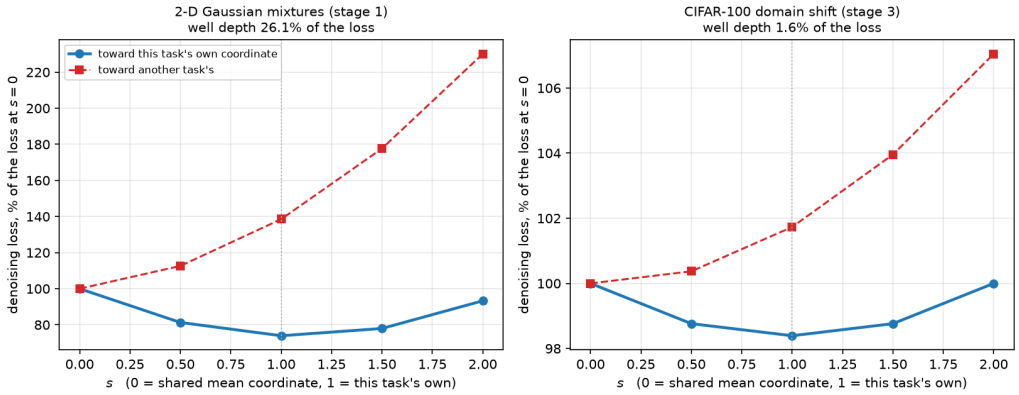

Figure 7: **Task-coordinate interpolation.** The GMM and CIFAR panels use their respective model checkpoints. Losses are normalised by the shared-coordinate loss; well depth is the percentage drop from this loss to the minimum along the own-task direction. Own-task and wrong-task directions keep their diagnostic meanings. Their shapes do not replace the paired deployment-route comparisons.

B.2 Complete factorial diagnostics

The boundary experiment uses 64/16/64 train/validation/test semantic tasks, five model/data seeds, 8,000 training steps, linear transport and $K_T = 0$. Each task is a four-component 2-D GMM with fixed mixture weights and covariances; only component means vary across tasks. Let $\bar{\mu}_{S,j}$ be the component-wise mean over meta-training tasks. Source-task identifiability is controlled by

$$\mu^y_{S,j}(\alpha) = \bar{\mu}_{S,j} + \alpha\big(\mu^y_{S,j} - \bar{\mu}_{S,j}\big),$$

with $\alpha \in \{0, .1, .25, .5, .75, 1, 1.5\}$. Thus $\alpha = 0$ makes source tasks distributionally identical, $\alpha = 1$ recovers the original task separation, and $\alpha > 1$ increases it.

Targets share the affine base relation

$$T_0(u) = Ru + b, \qquad R = 1.10 \begin{bmatrix} \cos 25^\circ & -\sin 25^\circ \\ \sin 25^\circ & \cos 25^\circ \end{bmatrix}, \qquad b = \begin{bmatrix} 0.75 \\ -0.50 \end{bmatrix}.$$

For each task y and component j, a perturbation $\xi_{y,j} \sim \mathcal{N}(0, I_2)$ is drawn once and reused across the heterogeneity sweep. After normalising these perturbations to unit RMS over training tasks and components, target means are

$$\mu^y_{T,j}(\alpha, \eta) = T_0\Big(\mu^y_{S,j}(\alpha)\Big) + \eta s_R \xi_{y,j},$$

where s_R is the RMS magnitude of the common source–target transformation over meta-training components and $\eta \in \{0, .1, .25, .5, 1, 2\}$. Hence $\eta = 0$ gives a perfectly shared relation, while increasing η introduces task-specific displacement that cannot be inferred from other source–target pairs. The same latent tasks and perturbations are reused across neighbouring cells to support paired comparisons. The component index is $j \in \{1, \ldots, 4\}$; it is unrelated to the refinement-step count J. The transformation scale depends on the source-separation condition:

$$s_R(\alpha)^2 = \frac{1}{4|\mathcal{Y}_{\text{train}}|} \sum_{y \in \mathcal{Y}_{\text{train}}} \sum_{j=1}^{4} \big\| T_0(\mu^y_{S,j}(\alpha)) - \mu^y_{S,j}(\alpha) \big\|_2^2.$$

Training-task centres and normalisation constants are applied unchanged to validation and test tasks. Bayes task identification uses the known source-GMM likelihoods over the 64 held-out candidates; learned identification uses the trained encoder coordinate. Neither identification accuracy is a generated-image fidelity metric.

The full grid makes three distinctions visible. First, G_{source} is effectively zero at $\alpha = 0$ even though a shared relation can still move the source distribution. Second, learned task discrimination increases much more slowly than Bayes identifiability around $\alpha = .1$–$.25$. Third, relation predictability falls with η even when Bayes-ID is saturated. The mechanism analysis makes the second point sharper: the rank association with G_{source} is $\rho = .98$ for learned-ID versus .89 for Bayes-ID, and replacing Bayes-ID with learned-ID in the same interaction model increases cell-level fit from $R^2 = .587$ to .843. These regressions are descriptive summaries of the factorial grid rather than causal estimates.

Table 5: Complete GMM factorial diagnostics. Top: correct-source advantage $G_{\text{source}} = \text{SW}_{wrong} - \text{SW}_{correct}$. Middle: learned-coordinate task-ID accuracy. Bottom: held-out oracle affine relation R^2. Each cell averages five seeds; generation effects further average 64 held-out tasks per seed.

Correct-source advantage G_{source}

$\alpha\backslash\eta$	0	0.1	0.25	0.5	1	2
0	0.000	0.000	-0.000	0.000	-0.000	-0.002
0.1	0.008	0.010	0.006	0.006	0.002	0.002
0.25	0.093	0.102	0.098	0.066	0.038	0.023
0.5	0.242	0.209	0.205	0.220	0.140	0.091
0.75	0.387	0.390	0.370	0.321	0.256	0.142
1	0.523	0.563	0.523	0.447	0.404	0.228
1.5	0.693	0.736	0.715	0.661	0.567	0.427

Learned-coordinate task ID

$\alpha\backslash\eta$	0	0.1	0.25	0.5	1	2
0	0.014	0.016	0.016	0.020	0.018	0.013
0.1	0.074	0.061	0.057	0.069	0.078	0.098
0.25	0.338	0.340	0.316	0.282	0.324	0.349
0.5	0.718	0.737	0.717	0.728	0.670	0.678
0.75	0.866	0.861	0.849	0.845	0.825	0.791
1	0.873	0.894	0.891	0.886	0.871	0.830
1.5	0.928	0.920	0.925	0.920	0.908	0.882

Oracle affine relation R^2

$\alpha\backslash\eta$	0	0.1	0.25	0.5	1	2
0	1.000	0.849	0.509	0.227	0.067	0.007
0.1	1.000	0.918	0.652	0.327	0.105	0.019
0.25	1.000	0.975	0.861	0.609	0.276	0.077
0.5	1.000	0.992	0.949	0.824	0.538	0.218
0.75	1.000	0.995	0.971	0.892	0.672	0.333
1	1.000	0.996	0.978	0.919	0.738	0.408
1.5	1.000	0.997	0.984	0.939	0.794	0.487

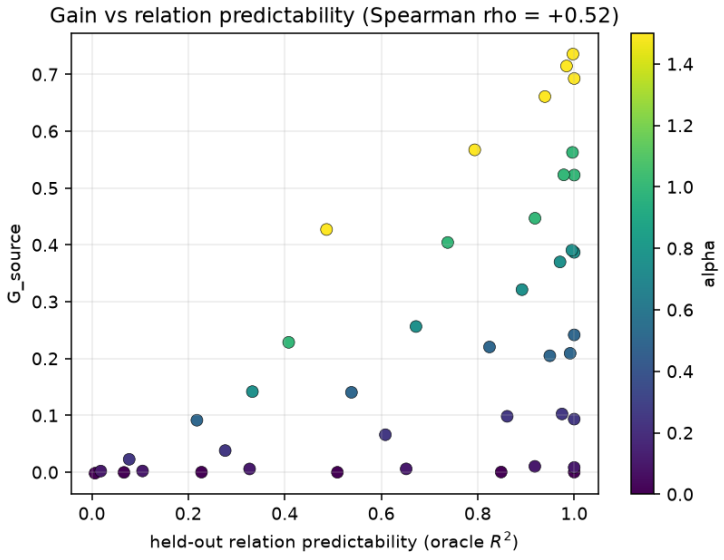

Figure 8: **Relation predictability is a separate boundary from task identifiability.** Each colour fixes source-task separation α while η varies. The horizontal axis is the held-out R^2 of an affine map fitted to true training-task component means; it is therefore independent of Meta-Diff's learned coordinate. At $\alpha = 1$, Bayes-ID remains 1.0 while relation R^2 falls from 1.000 to .408 and the correct-over-wrong SW advantage falls from .523 to .228. Over the full grid, $\rho(G_{\text{source}}, R^2_{\text{rel}}) = .54$.

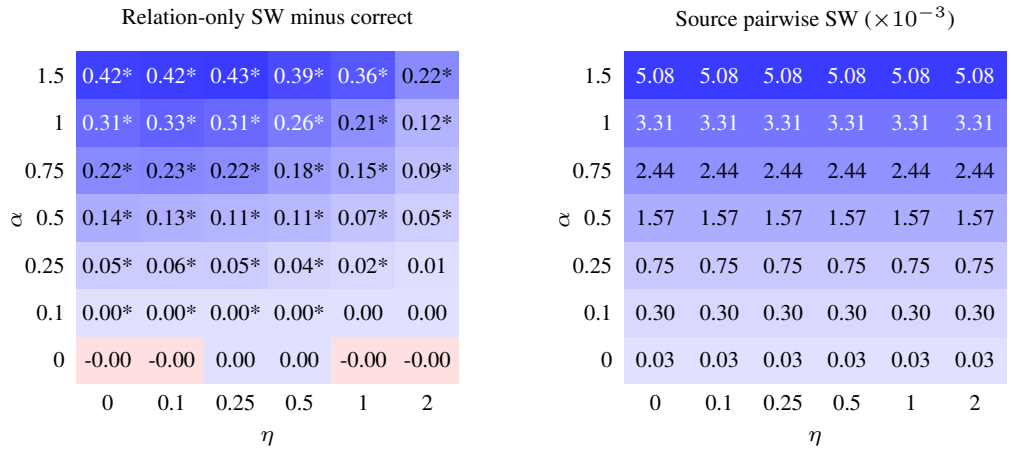

Figure 9: **Full-grid GMM controls.** *Left:* relation-only minus correct-source SW; positive values favour the source-conditioned route and stars mark hierarchical paired intervals excluding zero. *Right:* true-source pairwise SW (in units of 10^{-3}), an outcome-independent source-separation diagnostic. The source-conditioned route has a resolved advantage over relation-only in 33 of 42 cells; Table 7 also contains 34 resolved correct-versus-wrong cells, while source-reuse and oracle comparisons are not uniformly favourable.

Complete route means and paired uncertainty. The following tables give SW/MMD means and paired route contrasts for all 42 factorial cells. Route intervals use the hierarchical seed/task bootstrap and remain separate from the independent pipeline and shift-sweep summaries below.

Table 6: Complete factorial GMM SW route means. Lower is better; five seeds and 64 held-out tasks per seed.

α	η	Correct	Wrong	Relation-only	Source reuse	Oracle
0	0	0.09386	0.09396	0.09361	0.64158	0.09311
0	0.1	0.14144	0.14170	0.14129	0.70917	0.14840
0	0.25	0.16930	0.16911	0.16933	0.63886	0.16775
0	0.5	0.32266	0.32270	0.32317	0.77757	0.31470
0	1	0.58670	0.58635	0.58467	0.71809	0.52292
0	2	1.04219	1.04036	1.04108	1.16237	0.90496
0.1	0	0.18620	0.19456	0.19078	0.64653	0.19528
0.1	0.1	0.16964	0.17991	0.17319	0.71509	0.18359
0.1	0.25	0.23343	0.23926	0.23713	0.68183	0.23469
0.1	0.5	0.32876	0.33447	0.33166	0.76656	0.34392
0.1	1	0.61856	0.62068	0.62086	0.77128	0.56109
0.1	2	1.13177	1.13375	1.13227	1.19641	0.99374
0.25	0	0.26717	0.36053	0.32118	0.70593	0.29249
0.25	0.1	0.25727	0.35972	0.31534	0.68822	0.25517
0.25	0.25	0.30459	0.40290	0.35736	0.73701	0.31244
0.25	0.5	0.38146	0.44737	0.41840	0.73248	0.34137
0.25	1	0.66614	0.70417	0.69010	0.82028	0.55004
0.25	2	1.16372	1.18639	1.17351	1.23343	0.98578
0.5	0	0.43521	0.67676	0.57230	0.75540	0.48756
0.5	0.1	0.46706	0.67618	0.59320	0.66103	0.45364
0.5	0.25	0.50701	0.71183	0.62122	0.76954	0.52467
0.5	0.5	0.53815	0.75841	0.64413	0.78782	0.52228
0.5	1	0.83608	0.97654	0.91079	0.89080	0.68141
0.5	2	1.31188	1.40333	1.36229	1.32739	1.08919
0.75	0	0.63087	1.01769	0.85555	0.87930	0.68385
0.75	0.1	0.59480	0.98529	0.82756	0.83604	0.63539
0.75	0.25	0.65450	1.02438	0.87596	0.87217	0.68470
0.75	0.5	0.77528	1.09644	0.95094	0.93380	0.73713
0.75	1	0.98062	1.23690	1.13105	1.05599	0.85252
0.75	2	1.57659	1.71834	1.66407	1.47961	1.30518
1	0	0.83601	1.35871	1.14405	1.00989	0.87475
1	0.1	0.77718	1.33978	1.10764	1.01242	0.83143
1	0.25	0.84056	1.36367	1.14931	1.06450	0.88271
1	0.5	0.98325	1.42997	1.24690	1.04953	0.93214
1	1	1.19519	1.59930	1.40758	1.25568	1.07821
1	2	1.83706	2.06552	1.95703	1.68288	1.53973
1.5	0	1.30175	1.99432	1.71676	1.58434	1.45669
1.5	0.1	1.32602	2.06157	1.75072	1.56571	1.39372
1.5	0.25	1.32210	2.03668	1.74836	1.58367	1.39352
1.5	0.5	1.38495	2.04579	1.77314	1.57787	1.41911
1.5	1	1.66776	2.23484	2.02376	1.67006	1.50820
1.5	2	2.47424	2.90125	2.69910	2.32351	2.11900

Table 7: Complete factorial GMM SW source controls: comparator minus correct-source transport, with hierarchical seed/task 95% intervals. Positive favours correct-source transport.

α	η	Wrong − correct	Relation-only − correct
0	0	$+0.00010\,[-0.00035, +0.00059]$	$-0.00025\,[-0.00086, +0.00036]$
0	0.1	$+0.00025\,[-0.00054, +0.00093]$	$-0.00015\,[-0.00087, +0.00044]$
0	0.25	$-0.00019\,[-0.00106, +0.00062]$	$+0.00003\,[-0.00105, +0.00102]$
0	0.5	$+0.00004\,[-0.00138, +0.00160]$	$+0.00051\,[-0.00113, +0.00227]$
0	1	$-0.00035\,[-0.00319, +0.00254]$	$-0.00203\,[-0.00437, +0.00043]$
0	2	$-0.00182\,[-0.00695, +0.00300]$	$-0.00110\,[-0.00633, +0.00385]$
0.1	0	$+0.00835^{*}\,[+0.00423, +0.01379]$	$+0.00458^{*}\,[+0.00175, +0.00758]$
0.1	0.1	$+0.01027^{*}\,[+0.00584, +0.01483]$	$+0.00355^{*}\,[+0.00065, +0.00663]$
0.1	0.25	$+0.00583^{*}\,[+0.00176, +0.01033]$	$+0.00370^{*}\,[+0.00007, +0.00708]$
0.1	0.5	$+0.00572^{*}\,[+0.00102, +0.01033]$	$+0.00291^{*}\,[+0.00013, +0.00566]$
0.1	1	$+0.00213\,[-0.00186, +0.00623]$	$+0.00231\,[-0.00150, +0.00582]$
0.1	2	$+0.00198\,[-0.00546, +0.00931]$	$+0.00051\,[-0.00778, +0.00873]$
0.25	0	$+0.09336^{*}\,[+0.07494, +0.11153]$	$+0.05401^{*}\,[+0.03773, +0.07027]$
0.25	0.1	$+0.10245^{*}\,[+0.07687, +0.12837]$	$+0.05807^{*}\,[+0.04549, +0.06963]$
0.25	0.25	$+0.09831^{*}\,[+0.07986, +0.11555]$	$+0.05277^{*}\,[+0.04184, +0.06411]$
0.25	0.5	$+0.06591^{*}\,[+0.04754, +0.08641]$	$+0.03694^{*}\,[+0.01948, +0.05338]$
0.25	1	$+0.03803^{*}\,[+0.02468, +0.05266]$	$+0.02396^{*}\,[+0.01248, +0.03724]$
0.25	2	$+0.02266^{*}\,[+0.00366, +0.04314]$	$+0.00979\,[-0.01160, +0.03503]$
0.5	0	$+0.24155^{*}\,[+0.21516, +0.26783]$	$+0.13709^{*}\,[+0.10877, +0.16139]$
0.5	0.1	$+0.20912^{*}\,[+0.16496, +0.25024]$	$+0.12614^{*}\,[+0.09744, +0.15530]$
0.5	0.25	$+0.20482^{*}\,[+0.15696, +0.25306]$	$+0.11421^{*}\,[+0.08465, +0.14390]$
0.5	0.5	$+0.22026^{*}\,[+0.17131, +0.27742]$	$+0.10598^{*}\,[+0.07229, +0.13699]$
0.5	1	$+0.14047^{*}\,[+0.08743, +0.20163]$	$+0.07471^{*}\,[+0.04927, +0.10087]$
0.5	2	$+0.09145^{*}\,[+0.05889, +0.12524]$	$+0.05041^{*}\,[+0.00794, +0.08689]$
0.75	0	$+0.38682^{*}\,[+0.31401, +0.46122]$	$+0.22468^{*}\,[+0.18242, +0.26639]$
0.75	0.1	$+0.39049^{*}\,[+0.32193, +0.44730]$	$+0.23276^{*}\,[+0.18354, +0.27327]$
0.75	0.25	$+0.36989^{*}\,[+0.30159, +0.44673]$	$+0.22146^{*}\,[+0.17257, +0.26753]$
0.75	0.5	$+0.32116^{*}\,[+0.24744, +0.40265]$	$+0.17567^{*}\,[+0.11925, +0.23146]$
0.75	1	$+0.25629^{*}\,[+0.18494, +0.34726]$	$+0.15043^{*}\,[+0.09492, +0.20354]$
0.75	2	$+0.14175^{*}\,[+0.09699, +0.18694]$	$+0.08748^{*}\,[+0.04723, +0.13009]$
1	0	$+0.52270^{*}\,[+0.44979, +0.59553]$	$+0.30804^{*}\,[+0.25922, +0.35910]$
1	0.1	$+0.56260^{*}\,[+0.47423, +0.66587]$	$+0.33046^{*}\,[+0.27791, +0.38118]$
1	0.25	$+0.52310^{*}\,[+0.44510, +0.59904]$	$+0.30874^{*}\,[+0.24722, +0.36915]$
1	0.5	$+0.44672^{*}\,[+0.32666, +0.56506]$	$+0.26366^{*}\,[+0.20126, +0.33012]$
1	1	$+0.40410^{*}\,[+0.31186, +0.50894]$	$+0.21238^{*}\,[+0.15675, +0.27564]$
1	2	$+0.22846^{*}\,[+0.13314, +0.34093]$	$+0.11997^{*}\,[+0.05312, +0.18728]$
1.5	0	$+0.69257^{*}\,[+0.56014, +0.82813]$	$+0.41501^{*}\,[+0.33127, +0.50103]$
1.5	0.1	$+0.73555^{*}\,[+0.63170, +0.84237]$	$+0.42470^{*}\,[+0.35280, +0.49942]$
1.5	0.25	$+0.71458^{*}\,[+0.60007, +0.83214]$	$+0.42626^{*}\,[+0.33255, +0.51560]$
1.5	0.5	$+0.66084^{*}\,[+0.53810, +0.81305]$	$+0.38819^{*}\,[+0.31138, +0.47313]$
1.5	1	$+0.56708^{*}\,[+0.43503, +0.71278]$	$+0.35600^{*}\,[+0.25317, +0.46499]$
1.5	2	$+0.42700^{*}\,[+0.28685, +0.58549]$	$+0.22485^{*}\,[+0.14329, +0.30859]$

Table 8: Complete factorial GMM SW reuse/oracle contrasts: comparator minus correct-source transport, with hierarchical seed/task 95% intervals. Positive favours correct-source transport. The oracle uses abundant target data.

α	η	Source reuse − correct	Oracle − correct
0	0	$+0.54772^{*}$ $[+0.47915, +0.61436]$	-0.00075 $[-0.02123, +0.01295]$
0	0.1	$+0.56772^{*}$ $[+0.50726, +0.63460]$	$+0.00695$ $[-0.00126, +0.01545]$
0	0.25	$+0.46955^{*}$ $[+0.40729, +0.51238]$	-0.00156 $[-0.01588, +0.01273]$
0	0.5	$+0.45492^{*}$ $[+0.37794, +0.53592]$	-0.00796 $[-0.02093, +0.00691]$
0	1	$+0.13139$ $[-0.01521, +0.26379]$	-0.06379^{*} $[-0.09976, -0.02633]$
0	2	$+0.12019$ $[-0.02686, +0.29146]$	-0.13723^{*} $[-0.18348, -0.08521]$
0.1	0	$+0.46033^{*}$ $[+0.40389, +0.52582]$	$+0.00908^{*}$ $[+0.00413, +0.01490]$
0.1	0.1	$+0.54545^{*}$ $[+0.47017, +0.62192]$	$+0.01395^{*}$ $[+0.00489, +0.02307]$
0.1	0.25	$+0.44839^{*}$ $[+0.37088, +0.51481]$	$+0.00125$ $[-0.00893, +0.00999]$
0.1	0.5	$+0.43781^{*}$ $[+0.32473, +0.53652]$	$+0.01517$ $[-0.02963, +0.07979]$
0.1	1	$+0.15272^{*}$ $[+0.03401, +0.24160]$	-0.05746^{*} $[-0.08008, -0.03720]$
0.1	2	$+0.06465$ $[-0.14536, +0.25890]$	-0.13803^{*} $[-0.18973, -0.08234]$
0.25	0	$+0.43875^{*}$ $[+0.34914, +0.52549]$	$+0.02532$ $[-0.01838, +0.06959]$
0.25	0.1	$+0.43095^{*}$ $[+0.33990, +0.51491]$	-0.00210 $[-0.01978, +0.01578]$
0.25	0.25	$+0.43242^{*}$ $[+0.31778, +0.56028]$	$+0.00784$ $[-0.01576, +0.03842]$
0.25	0.5	$+0.35102^{*}$ $[+0.25553, +0.44285]$	-0.04009 $[-0.07272, +0.00166]$
0.25	1	$+0.15415^{*}$ $[+0.01199, +0.25514]$	-0.11609^{*} $[-0.14699, -0.08798]$
0.25	2	$+0.06971$ $[-0.06061, +0.19539]$	-0.17795^{*} $[-0.23321, -0.13095]$
0.5	0	$+0.32019^{*}$ $[+0.26384, +0.36867]$	$+0.05235^{*}$ $[+0.02156, +0.08589]$
0.5	0.1	$+0.19397^{*}$ $[+0.08944, +0.29378]$	-0.01342 $[-0.07453, +0.04412]$
0.5	0.25	$+0.26253^{*}$ $[+0.16573, +0.34602]$	$+0.01765$ $[-0.02270, +0.06038]$
0.5	0.5	$+0.24967^{*}$ $[+0.12643, +0.35329]$	-0.01587 $[-0.06285, +0.03293]$
0.5	1	$+0.05472$ $[-0.10061, +0.15118]$	-0.15467^{*} $[-0.22629, -0.09391]$
0.5	2	$+0.01552$ $[-0.13468, +0.15717]$	-0.22269^{*} $[-0.29576, -0.16079]$
0.75	0	$+0.24843^{*}$ $[+0.12217, +0.36943]$	$+0.05298$ $[-0.01921, +0.13013]$
0.75	0.1	$+0.24124^{*}$ $[+0.12439, +0.34690]$	$+0.04059$ $[-0.01339, +0.08357]$
0.75	0.25	$+0.21767^{*}$ $[+0.12208, +0.31442]$	$+0.03021$ $[-0.02956, +0.09356]$
0.75	0.5	$+0.15852^{*}$ $[+0.05433, +0.24901]$	-0.03815 $[-0.07614, +0.00502]$
0.75	1	$+0.07538$ $[-0.05367, +0.16205]$	-0.12810^{*} $[-0.20445, -0.06870]$
0.75	2	-0.09698 $[-0.30044, +0.04797]$	-0.27141^{*} $[-0.39106, -0.17787]$
1	0	$+0.17388^{*}$ $[+0.07600, +0.27364]$	$+0.03874$ $[-0.02206, +0.09591]$
1	0.1	$+0.23524^{*}$ $[+0.14588, +0.32692]$	$+0.05425$ $[-0.00352, +0.10887]$
1	0.25	$+0.22394^{*}$ $[+0.08812, +0.33918]$	$+0.04214$ $[-0.02872, +0.10773]$
1	0.5	$+0.06628$ $[-0.12114, +0.20964]$	-0.05111 $[-0.15309, +0.03730]$
1	1	$+0.06048$ $[-0.00260, +0.12129]$	-0.11699^{*} $[-0.18859, -0.04980]$
1	2	-0.15418 $[-0.33414, +0.00775]$	-0.29733^{*} $[-0.43268, -0.17561]$
1.5	0	$+0.28258^{*}$ $[+0.17342, +0.39528]$	$+0.15494^{*}$ $[+0.10056, +0.21521]$
1.5	0.1	$+0.23969^{*}$ $[+0.09724, +0.38696]$	$+0.06770$ $[-0.07017, +0.19717]$
1.5	0.25	$+0.26157^{*}$ $[+0.15161, +0.36427]$	$+0.07142^{*}$ $[+0.00223, +0.14246]$
1.5	0.5	$+0.19292^{*}$ $[+0.10353, +0.28814]$	$+0.03416$ $[-0.02510, +0.08968]$
1.5	1	$+0.00230$ $[-0.08436, +0.09102]$	-0.15957^{*} $[-0.24140, -0.07676]$
1.5	2	-0.15073^{*} $[-0.29265, -0.04305]$	-0.35525^{*} $[-0.46329, -0.24579]$

Table 9: Complete factorial GMM MMD route means. Lower is better; five seeds and 64 held-out tasks per seed.

α	η	Correct	Wrong	Relation-only	Source reuse	Oracle
0	0	0.01827	0.01820	0.01813	0.88275	0.01782
0	0.1	0.03767	0.03772	0.03767	1.01943	0.04387
0	0.25	0.06638	0.06640	0.06653	0.87212	0.06448
0	0.5	0.14388	0.14391	0.14401	1.02813	0.13898
0	1	0.32266	0.32176	0.32035	0.71514	0.24480
0	2	0.72649	0.72400	0.72402	1.19824	0.52956
0.1	0	0.07334	0.08149	0.07906	0.83916	0.07990
0.1	0.1	0.05986	0.06983	0.06498	0.97429	0.06869
0.1	0.25	0.09329	0.10050	0.09806	0.91302	0.09706
0.1	0.5	0.16318	0.17106	0.16820	0.95430	0.18661
0.1	1	0.34343	0.34776	0.34619	0.80935	0.25908
0.1	2	0.75629	0.75893	0.75859	1.22893	0.54214
0.25	0	0.10815	0.23019	0.17891	0.78865	0.13855
0.25	0.1	0.10634	0.23746	0.17717	0.80646	0.10232
0.25	0.25	0.12515	0.25525	0.19413	0.81587	0.12691
0.25	0.5	0.18401	0.28003	0.23306	0.81571	0.14977
0.25	1	0.37590	0.43806	0.41686	0.81969	0.25679
0.25	2	0.78925	0.82407	0.81257	1.25444	0.56455
0.5	0	0.17616	0.49284	0.35407	0.59608	0.23008
0.5	0.1	0.22647	0.52662	0.39873	0.45728	0.20212
0.5	0.25	0.23612	0.52913	0.40286	0.59270	0.25045
0.5	0.5	0.25410	0.57246	0.40274	0.62153	0.23771
0.5	1	0.49623	0.72537	0.62030	0.73690	0.33088
0.5	2	0.82885	0.98155	0.91753	1.21317	0.55722
0.75	0	0.28914	0.83530	0.59069	0.60096	0.33253
0.75	0.1	0.27496	0.82089	0.58035	0.56537	0.29812
0.75	0.25	0.31701	0.85119	0.62212	0.60848	0.35899
0.75	0.5	0.39967	0.87885	0.64763	0.62969	0.34415
0.75	1	0.53643	0.93519	0.77855	0.75011	0.41116
0.75	2	0.99672	1.24134	1.15750	1.17534	0.68976
1	0	0.38453	1.13998	0.81764	0.61778	0.42747
1	0.1	0.34852	1.17486	0.82466	0.62534	0.37613
1	0.25	0.40165	1.17876	0.84802	0.69373	0.42724
1	0.5	0.53272	1.20578	0.95292	0.63116	0.45816
1	1	0.64184	1.28125	0.99301	0.83097	0.54703
1	2	1.13301	1.52550	1.35384	1.22301	0.82498
1.5	0	0.66734	1.76367	1.28258	1.04101	0.82003
1.5	0.1	0.68992	1.82891	1.34389	1.01392	0.73695
1.5	0.25	0.64330	1.74992	1.28208	1.00599	0.71198
1.5	0.5	0.68368	1.67984	1.25541	0.96404	0.69489
1.5	1	0.86344	1.75000	1.38795	0.94367	0.67324
1.5	2	1.62377	2.34254	1.99685	1.68128	1.17444

Table 10: Complete factorial GMM MMD source controls: comparator minus correct-source transport, with hierarchical seed/task 95% intervals. Positive favours correct-source transport.

α	η	Wrong − correct	Relation-only − correct
0	0	$-0.00007\,[-0.00025, +0.00007]$	$-0.00014\,[-0.00033, +0.00003]$
0	0.1	$+0.00005\,[-0.00008, +0.00018]$	$+0.00000\,[-0.00013, +0.00012]$
0	0.25	$+0.00003\,[-0.00022, +0.00026]$	$+0.00015\,[-0.00009, +0.00044]$
0	0.5	$+0.00004\,[-0.00072, +0.00090]$	$+0.00014\,[-0.00066, +0.00098]$
0	1	$-0.00090\,[-0.00334, +0.00176]$	$-0.00231^{*}\,[-0.00389, -0.00061]$
0	2	$-0.00249\,[-0.00761, +0.00164]$	$-0.00247\,[-0.00659, +0.00100]$
0.1	0	$+0.00815^{*}\,[+0.00443, +0.01245]$	$+0.00573^{*}\,[+0.00352, +0.00875]$
0.1	0.1	$+0.00998^{*}\,[+0.00621, +0.01473]$	$+0.00512^{*}\,[+0.00289, +0.00797]$
0.1	0.25	$+0.00720^{*}\,[+0.00217, +0.01418]$	$+0.00476^{*}\,[+0.00159, +0.00878]$
0.1	0.5	$+0.00788^{*}\,[+0.00284, +0.01387]$	$+0.00502^{*}\,[+0.00219, +0.00798]$
0.1	1	$+0.00433\,[-0.00037, +0.00924]$	$+0.00277\,[-0.00149, +0.00691]$
0.1	2	$+0.00264\,[-0.00755, +0.01284]$	$+0.00230\,[-0.00857, +0.01487]$
0.25	0	$+0.12203^{*}\,[+0.09047, +0.15798]$	$+0.07076^{*}\,[+0.05145, +0.08979]$
0.25	0.1	$+0.13112^{*}\,[+0.09258, +0.18448]$	$+0.07083^{*}\,[+0.05726, +0.08734]$
0.25	0.25	$+0.13011^{*}\,[+0.10825, +0.15320]$	$+0.06898^{*}\,[+0.05395, +0.08484]$
0.25	0.5	$+0.09601^{*}\,[+0.07087, +0.12419]$	$+0.04904^{*}\,[+0.02913, +0.06673]$
0.25	1	$+0.06216^{*}\,[+0.04336, +0.08356]$	$+0.04096^{*}\,[+0.02735, +0.05778]$
0.25	2	$+0.03481^{*}\,[+0.00644, +0.06557]$	$+0.02331\,[-0.00206, +0.05358]$
0.5	0	$+0.31668^{*}\,[+0.27528, +0.36086]$	$+0.17791^{*}\,[+0.15063, +0.20589]$
0.5	0.1	$+0.30015^{*}\,[+0.22862, +0.37459]$	$+0.17226^{*}\,[+0.13303, +0.22478]$
0.5	0.25	$+0.29300^{*}\,[+0.23177, +0.36024]$	$+0.16674^{*}\,[+0.13550, +0.19907]$
0.5	0.5	$+0.31836^{*}\,[+0.24775, +0.40319]$	$+0.14864^{*}\,[+0.10246, +0.19293]$
0.5	1	$+0.22914^{*}\,[+0.16015, +0.30839]$	$+0.12407^{*}\,[+0.08790, +0.15940]$
0.5	2	$+0.15270^{*}\,[+0.09963, +0.20514]$	$+0.08868^{*}\,[+0.03489, +0.13962]$
0.75	0	$+0.54616^{*}\,[+0.46194, +0.63855]$	$+0.30155^{*}\,[+0.24394, +0.37747]$
0.75	0.1	$+0.54593^{*}\,[+0.44574, +0.63694]$	$+0.30539^{*}\,[+0.23962, +0.37953]$
0.75	0.25	$+0.53418^{*}\,[+0.44127, +0.64401]$	$+0.30511^{*}\,[+0.23772, +0.38154]$
0.75	0.5	$+0.47917^{*}\,[+0.38579, +0.57822]$	$+0.24796^{*}\,[+0.16661, +0.33848]$
0.75	1	$+0.39875^{*}\,[+0.30187, +0.52259]$	$+0.24212^{*}\,[+0.16768, +0.31574]$
0.75	2	$+0.24462^{*}\,[+0.17113, +0.31805]$	$+0.16078^{*}\,[+0.09541, +0.22461]$
1	0	$+0.75545^{*}\,[+0.65833, +0.85396]$	$+0.43311^{*}\,[+0.34737, +0.53878]$
1	0.1	$+0.82634^{*}\,[+0.68290, +0.99393]$	$+0.47615^{*}\,[+0.38587, +0.56505]$
1	0.25	$+0.77711^{*}\,[+0.66190, +0.89134]$	$+0.44637^{*}\,[+0.35150, +0.57146]$
1	0.5	$+0.67306^{*}\,[+0.52330, +0.82360]$	$+0.42019^{*}\,[+0.31515, +0.52882]$
1	1	$+0.63941^{*}\,[+0.49330, +0.81341]$	$+0.35117^{*}\,[+0.25545, +0.46088]$
1	2	$+0.39249^{*}\,[+0.24244, +0.57382]$	$+0.22082^{*}\,[+0.11081, +0.32567]$
1.5	0	$+1.09634^{*}\,[+0.91938, +1.28449]$	$+0.61525^{*}\,[+0.49514, +0.75308]$
1.5	0.1	$+1.13899^{*}\,[+0.97726, +1.31227]$	$+0.65397^{*}\,[+0.55980, +0.75650]$
1.5	0.25	$+1.10662^{*}\,[+0.96072, +1.25630]$	$+0.63877^{*}\,[+0.49773, +0.79731]$
1.5	0.5	$+0.99616^{*}\,[+0.82566, +1.20117]$	$+0.57174^{*}\,[+0.44409, +0.70192]$
1.5	1	$+0.88656^{*}\,[+0.68842, +1.10578]$	$+0.52451^{*}\,[+0.37423, +0.69403]$
1.5	2	$+0.71877^{*}\,[+0.48858, +0.97687]$	$+0.37308^{*}\,[+0.23704, +0.50955]$

Table 11: Complete factorial GMM MMD reuse/oracle contrasts: comparator minus correct-source transport, with hierarchical seed/task 95% intervals. Positive favours correct-source transport. The oracle uses abundant target data.

α	η	Source reuse − correct	Oracle − correct
0	0	$+0.86448^{*}$ $[+0.62391, +1.08397]$	-0.00044 $[-0.01105, +0.00643]$
0	0.1	$+0.98176^{*}$ $[+0.86073, +1.10203]$	$+0.00620^{*}$ $[+0.00008, +0.01270]$
0	0.25	$+0.80574^{*}$ $[+0.62634, +0.92682]$	-0.00190 $[-0.01245, +0.00877]$
0	0.5	$+0.88426^{*}$ $[+0.69128, +1.05914]$	-0.00489 $[-0.02169, +0.01614]$
0	1	$+0.39248^{*}$ $[+0.18649, +0.59469]$	-0.07786^{*} $[-0.11287, -0.04096]$
0	2	$+0.47175^{*}$ $[+0.24454, +0.74321]$	-0.19693^{*} $[-0.24489, -0.15161]$
0.1	0	$+0.76582^{*}$ $[+0.64899, +0.88055]$	$+0.00656^{*}$ $[+0.00010, +0.01308]$
0.1	0.1	$+0.91444^{*}$ $[+0.73172, +1.06691]$	$+0.00883^{*}$ $[+0.00461, +0.01360]$
0.1	0.25	$+0.81973^{*}$ $[+0.60398, +1.01282]$	$+0.00377$ $[-0.00503, +0.01209]$
0.1	0.5	$+0.79112^{*}$ $[+0.62793, +0.91909]$	$+0.02343$ $[-0.02890, +0.10372]$
0.1	1	$+0.46593^{*}$ $[+0.31219, +0.61078]$	-0.08435^{*} $[-0.11503, -0.05602]$
0.1	2	$+0.47265^{*}$ $[+0.17378, +0.73851]$	-0.21415^{*} $[-0.24325, -0.18304]$
0.25	0	$+0.68050^{*}$ $[+0.53245, +0.83404]$	$+0.03039$ $[-0.01902, +0.07962]$
0.25	0.1	$+0.70012^{*}$ $[+0.56395, +0.82675]$	-0.00402 $[-0.02492, +0.01340]$
0.25	0.25	$+0.69072^{*}$ $[+0.54898, +0.83160]$	$+0.00176$ $[-0.02138, +0.03217]$
0.25	0.5	$+0.63169^{*}$ $[+0.45808, +0.82498]$	-0.03424 $[-0.07724, +0.01442]$
0.25	1	$+0.44378^{*}$ $[+0.25424, +0.62415]$	-0.11912^{*} $[-0.14697, -0.09456]$
0.25	2	$+0.46519^{*}$ $[+0.34597, +0.59740]$	-0.22470^{*} $[-0.27629, -0.17115]$
0.5	0	$+0.41992^{*}$ $[+0.34712, +0.49157]$	$+0.05392^{*}$ $[+0.02198, +0.09153]$
0.5	0.1	$+0.23082^{*}$ $[+0.09143, +0.36996]$	-0.02435 $[-0.09469, +0.04308]$
0.5	0.25	$+0.35657^{*}$ $[+0.20655, +0.49022]$	$+0.01433$ $[-0.03347, +0.06629]$
0.5	0.5	$+0.36743^{*}$ $[+0.19529, +0.50937]$	-0.01639 $[-0.07890, +0.05097]$
0.5	1	$+0.24067^{*}$ $[+0.01953, +0.39879]$	-0.16535^{*} $[-0.24434, -0.09497]$
0.5	2	$+0.38432^{*}$ $[+0.25790, +0.51440]$	-0.27164^{*} $[-0.32902, -0.21952]$
0.75	0	$+0.31182^{*}$ $[+0.13990, +0.47463]$	$+0.04339$ $[-0.01995, +0.10980]$
0.75	0.1	$+0.29041^{*}$ $[+0.14437, +0.42873]$	$+0.02316$ $[-0.02126, +0.06039]$
0.75	0.25	$+0.29147^{*}$ $[+0.15950, +0.42208]$	$+0.04198$ $[-0.02242, +0.10815]$
0.75	0.5	$+0.23002^{*}$ $[+0.10975, +0.34290]$	-0.05552^{*} $[-0.09451, -0.01572]$
0.75	1	$+0.21367^{*}$ $[+0.05158, +0.32043]$	-0.12527^{*} $[-0.20447, -0.06212]$
0.75	2	$+0.17862$ $[-0.02082, +0.34877]$	-0.30697^{*} $[-0.42875, -0.19865]$
1	0	$+0.23325^{*}$ $[+0.10184, +0.37085]$	$+0.04294$ $[-0.01879, +0.10034]$
1	0.1	$+0.27682^{*}$ $[+0.15706, +0.40179]$	$+0.02761$ $[-0.03696, +0.08544]$
1	0.25	$+0.29208^{*}$ $[+0.11990, +0.45826]$	$+0.02559$ $[-0.04356, +0.09289]$
1	0.5	$+0.09844$ $[-0.10099, +0.24785]$	-0.07456 $[-0.18434, +0.01108]$
1	1	$+0.18914^{*}$ $[+0.11140, +0.27233]$	-0.09481^{*} $[-0.16812, -0.02574]$
1	2	$+0.08999$ $[-0.09839, +0.25686]$	-0.30804^{*} $[-0.45103, -0.16747]$
1.5	0	$+0.37367^{*}$ $[+0.21737, +0.53198]$	$+0.15269^{*}$ $[+0.08160, +0.23430]$
1.5	0.1	$+0.32400^{*}$ $[+0.10381, +0.55118]$	$+0.04703$ $[-0.12328, +0.20882]$
1.5	0.25	$+0.36268^{*}$ $[+0.20650, +0.52492]$	$+0.06868$ $[-0.01017, +0.16233]$
1.5	0.5	$+0.28036^{*}$ $[+0.14932, +0.42798]$	$+0.01121$ $[-0.05211, +0.07335]$
1.5	1	$+0.08023$ $[-0.01360, +0.17687]$	-0.19020^{*} $[-0.24430, -0.13708]$
1.5	2	$+0.05751$ $[-0.06773, +0.17574]$	-0.44933^{*} $[-0.60695, -0.29172]$

B.3 USABLE IDENTITY EXPLAINS THE SOURCE-SPECIFIC GAIN

The mechanism analysis compares intrinsic Bayes task identification with nearest-centroid identification in the learned source coordinate. Bayes-ID saturates rapidly, while learned-ID catches up only at much larger α. Averaging across relation levels gives the following identity profile.

Table 12: Intrinsic versus usable source-task identifiability. Chance is $1/64 = .0156$. The collapse gap is Bayes-ID minus learned-ID.

α	Bayes-ID	learned-ID	Collapse gap
0.00	.016	.016	$-.001$
0.10	.400	.073	.327
0.25	.953	.325	**.628**
0.50	1.000	.708	.292
0.75	1.000	.839	.161
1.00	1.000	.874	.126
1.50	1.000	.914	.086

The source-specific gain tracks usable learned identity more closely than the intrinsic ceiling. Across the 42 cells, $\rho(G_{\text{source}}, I_{\text{learned}}) = .98$ versus $\rho(G_{\text{source}}, I_{\text{Bayes}}) = .89$. Restricting to the 24 cells with Bayes-ID $> .99$ leaves $\rho(G_{\text{source}}, I_{\text{learned}}) = .97$. Descriptive cell-level regressions give: The dedicated identity-pipeline retrieval uses an independently sampled retrieval protocol, so its learned-z_S values below differ slightly from the phase-diagram learned-ID summary above; the qualitative ordering is unchanged.

Table 13: Descriptive GMM regressions predicting G_{source}. OLS fits use the 42 unrounded seed-mean cells. Displayed coefficient intervals are the conventional homoskedastic normal approximation $\widehat{\beta} \pm 1.96\,\text{SE}_{\text{OLS}}$ with residual variance estimated using $42 - p$ degrees of freedom ($p = 4$ including the intercept). These descriptive cell-level intervals do not account for matched-seed dependence across cells and are not the hierarchical seed/task intervals for route comparisons.

Predictors	Fit R^2	Key coefficient	95% CI
Bayes-ID $+R^2_{\text{rel}}+$ interaction	.587	$I_{\text{Bayes}} \times R^2_{\text{rel}} = +.54$	$[+.17, +.92]$
learned-ID $+R^2_{\text{rel}}+$ interaction	**.843**	$I_{\text{learned}} \times R^2_{\text{rel}} = +.61$	$[+.33, +.89]$
Bayes-ID + learned-ID $+R^2_{\text{rel}}$	.814	$I_{\text{learned}} = +.73$	$[+.55, +.91]$
		$I_{\text{Bayes}} = -.27$	$[-.44, -.10]$

The negative Bayes coefficient in the joint descriptive regression should not be interpreted causally; Bayes-ID and learned-ID share strong variation with α. The robust point is that variation remaining in the learned coordinate continues to track the transfer benefit after intrinsic identification has saturated.

B.4 IDENTITY THROUGH THE TRANSPORT PIPELINE

To locate where task information is lost, we retrieve the correct target task at successive stages of the pipeline. Transported-coordinate retrieval asks whether $\widetilde{z}_T = Az_S + b$ is closest to the correct held-out target coordinate. Raw $z_S \to z_T$ retrieval is a cross-domain baseline that omits transport.

Table 14: Mean top-1 task identity through the GMM pipeline, averaged over relation levels and seeds. Chance is $1/64 = .0156$.

α	Bayes-ID	learned z_S	transported $\widetilde{z}_T$	raw $z_S \to z_T$
0.00	.016	.015	.013	.016
0.10	.400	.071	.019	.015
0.25	.953	.323	.089	.023
0.50	1.000	.691	.169	.039
0.75	1.000	.831	.258	.074
1.00	1.000	.883	.315	.109
1.50	1.000	.920	.341	.148

Identity is therefore compressed at both learned stages. Averaged over the grid, transported retrieval is .172, compared with .061 for raw cross-domain retrieval. The reported overall gain is $+.112$ (paired seed-block interval $[+.106, +.116]$). Mean transported retrieval is close to chance at $\alpha = 0$ and .1. These summaries do not provide cellwise retrieval intervals. Across the 42 cells, $\rho(G_{\text{source}}, I_{\widetilde{z}_T}) = .91$. The nearest-centroid margin for learned z_S grows from $-.002$ at $\alpha = 0$ to $+.060$ at $\alpha = 1.5$,

while the transported margin remains negative ($-.039$ to $-.079$), showing that transport improves task alignment without fully recovering the correct target coordinate.

B.5 Source–target shift magnitude and transport headroom

The phase diagram varies task identity and relation consistency. A separate sweep holds $\alpha = 1, \eta = 0$ fixed and changes only the magnitude of the shared transformation,

$$T_\lambda(u) = u + \lambda(T_0(u) - u), \qquad \lambda \in \{0, .25, .5, 1, 2\}.$$

This separates the value of choosing the correct source from the value of transporting that source.

Table 15: GMM shift-magnitude sweep: route means averaged over tasks and five seeds. $G_{\text{source}} = \text{SW}_{\text{wrong}} - \text{SW}_{\text{correct}}$; transport-over-reuse is $\text{SW}_{\text{reuse}} - \text{SW}_{\text{correct}}$.

λ	Correct	Wrong	Relation-only	Source reuse	G_{source}	Transport over reuse
0.00	.759	1.241	1.062	.783	+.482	+.024
0.25	.765	1.252	1.048	.828	+.487	+.063
0.50	.802	1.315	1.115	.905	+.513	+.102
1.00	.835	1.355	1.157	1.066	+.520	+.231
2.00	1.037	1.667	1.443	1.577	+.631	+.541

The source-specific gain remains large across the sweep because the semantic source task is identifiable even when $\lambda = 0$. The mean transport-over-reuse advantage grows from $.024$ at $\lambda = 0$ to $.541$ at $\lambda = 2$. These shift-sweep comparisons are descriptive mean differences. The independent model-free source–target SW headroom increases from approximately $.08$ to 1.59 over the same sweep. This dissociation shows that source specificity and transport utility answer different questions: a correct source can matter even when there is almost nothing to transport.

B.6 Historical representation-capacity and relatedness diagnostics

The earlier MLP-era GMM suite is retained because it probes the representational ceiling, but it is not used as evidence for the final correct source + transport route. Its aggregate summaries lack the paired task-level records needed to recover new confidence intervals.

Table 16: Historical GMM representation-capacity diagnostic. The oracle fits an abundant-target coordinate; captured fraction is $(e_0 - e_{\text{oracle}})/(e_0 - e_{\text{FT}})$.

Configuration	No adapt.	Oracle	Conv. full FT	Captured
$k = 16$, 80k	.8821	.6320	.1276	33.2%
$k = 16$, 320k	.8741	.5965	.1508	38.4%
$k = 64$, 80k	.8868	.5396	.1267	45.7%
$k = 64$, 320k	.8779	.4690	.1565	56.7%

Table 17: Historical GMM task-relatedness diagnostic at $k = 16$, 80k steps. Larger template perturbations produce less related tasks.

Task family	No adapt.	Oracle	Full FT	Captured
Perturbation .10	.7513	.3835	.1179	58.1%
Perturbation .25	.8910	.5583	.1220	43.3%
Perturbation .50	.9516	.6949	.1359	31.5%
Perturbation 1.00	.9679	.7808	.1239	22.2%

These diagnostics support the intuitive claim that a compact shared representation has a higher ceiling when tasks are more closely related. Because the runs predate the final deployment protocol and report aggregate rather than paired route statistics, they are deliberately separated from the main claim.

C CIFAR-100: Full Controls and Ablations

C.1 Protocol

Entire CIFAR-100 superclasses are assigned to train, validation or test so that close sibling classes do not cross the split. Twelve superclasses (60 fine classes) train the model, four superclasses (20 classes) are used for validation and hyperparameter selection, and four superclasses (20 classes) are held for

final evaluation. A fixed permutation creates disjoint source and target pools: 300 source-support images, 50 source-query images, 20 reserved target-support images and 230 target-query images per class. Target transformations are Gaussian blur ($\sigma = 1.0$), additive Gaussian noise (standard deviation 0.18 on the normalised scale), and contrast reduction (factor 0.45 toward the image mean), all at severity 3.

Meaning of the evaluation packages. In all headline generation packages, $J = 0$ denotes no target-time coordinate-refinement steps for correct-source transport; K_T is the target-support budget made available to adapting baselines, not a refinement-step count. Table 1A uses 60 validation class–transformation episodes at $K_T = 0$; Table 1B uses 60 final-test episodes at each of $K_T = 1, 5, 20$. The source-dependence sweep is a separate 180-episode validation package. The conditioning comparison and the two transport-structure comparisons are three separately matched packages, each with one training seed per arm; their rows are not paired across packages. Historical interpolation/deployment-grid and refinement diagnostics use the identified 50k `cifar_ds3` checkpoint. The frozen FiLM retrieval/direct-target diagnostic below is another independent package and uses its own support resampling, post-hoc target split and semantic evaluator. For the episode-based generation tables, reported $\pm$ values are $1.96\,\mathrm{SE}$ half-widths; three corruption episodes share each class, so they are not class-cluster or training-seed intervals. A star marks a paired interval excluding zero in either direction, not an equivalence test or a multiplicity-adjusted decision.

C.2 Source specificity from zero to twenty targets

This independent validation package adds mean-source and shuffled-source controls to the final-test comparison. 9 of nine main-relative source-control SW intervals favour correct source + transport. Target-only comparisons favour transport at $K_T = 1$; remain unresolved at $K_T = 5, 20$. Rows labelled source-initialised adaptation optimise a source-initialised coordinate using target support; they do not implement direct target-free reuse.

Table 18: CIFAR validation source-dependence sweep on the matched correct source + transport evaluation: 180 total episodes, 60 per budget. Main transport and mean/shuffled/relation controls use $J = 0$, meaning zero target-time coordinate-refinement steps. Source-initialised adaptation starts at z_S and uses target support; it is not strict reuse. SW is lower-is-better; paired Δ is comparator minus main, with episode-level 95% half-widths and unrounded stars.

Route / comparison	$K_T = 1$	$K_T = 5$	$K_T = 20$
correct source + transport	0.2071 ± 0.0192	0.2077 ± 0.0171	0.2129 ± 0.0193
Source-init. adaptation	0.2436 ± 0.0197	0.2351 ± 0.0182	0.2203 ± 0.0220
Target-only	0.2255 ± 0.0191	0.2130 ± 0.0191	0.2087 ± 0.0221
Mean source	0.2309 ± 0.0201	0.2347 ± 0.0187	0.2379 ± 0.0204
Shuffled source	0.2417 ± 0.0206	0.2453 ± 0.0191	0.2469 ± 0.0208
Relation-only	0.2413 ± 0.0197	0.2427 ± 0.0185	0.2476 ± 0.0200
Δ Source-init. adaptation	$0.0366 \pm 0.0101^*$	$0.0274 \pm 0.0082^*$	$0.0074 \pm 0.0070^*$
Δ Target-only	$0.0184 \pm 0.0076^*$	0.0053 ± 0.0055	-0.0043 ± 0.0084
Δ Mean source	$0.0239 \pm 0.0067^*$	$0.0270 \pm 0.0075^*$	$0.0250 \pm 0.0074^*$
Δ Shuffled source	$0.0347 \pm 0.0105^*$	$0.0376 \pm 0.0114^*$	$0.0340 \pm 0.0111^*$
Δ Relation-only	$0.0342 \pm 0.0074^*$	$0.0350 \pm 0.0073^*$	$0.0347 \pm 0.0079^*$

The same final-test package gives energy MMD $2.3306, 2.3386, 2.4041$ for correct source + transport, $3.4434, 2.5351, 2.4137$ for target-only, and $5.8971, 4.0219, 1.9086$ for full fine-tuning at $K_T = 1, 5, 20$. Paired full-FT-minus-main differences are $3.5666 \pm 0.7496^*$, $1.6833 \pm 0.6629^*$, $-0.4955 \pm 0.3955^*$, favouring transport at $K_T = 1, 5$ and full fine-tuning at $K_T = 20$. Paired target-only-minus-main differences are $1.1129 \pm 0.3763^*$, $0.1965 \pm 0.1611^*$ and 0.0096 ± 0.1781, favouring transport at $K_T = 1, 5$ and remaining unresolved at $K_T = 20$. They use the same 60 matched episodes per budget and are not inferred from the SW intervals.

C.3 Validation-frozen full fine-tuning and coordinate refinement

Identity of the current CIFAR result packages. The archived A1 source-dependence, A4 source-budget, A6 zero-target and A9 final-test outputs identify the same 50k-step basis-conditioned `cifar_ds3` EMA checkpoint, with linear transport and $k = 16$ (recorded SHA256 prefix `be2b86b158cf`). Its training seed is 12345; these evaluations use seed 4321, 96 generated samples



Figure 10: CIFAR $K_T = 1$ deployment grid from the 50k-step checkpoint. The seven rows, from top to bottom, show no adaptation, source-initialised target adaptation (labelled reuse-z_S in the image), correct source + transport with $J = 0$, transport plus refinement, target-only, oracle, and real target. Here J counts target-time coordinate-refinement steps: $J = 0$ is the headline transport route, while transport plus refinement is a separate target-using diagnostic. Generated rows share initial noise within each column; the bottom row is real target data, not generated.

per episode and 50-step DDIM. A1 and A6 use $M_S = 64$ with independently drawn source sets; A4 varies M_S using nested prefixes. The separate frozen FiLM diagnostics below use their own checkpoint and protocol.

Coordinate updates in the validation controls. The A1 and A4 reports log $J = 25$, coordinate learning rate 0.01 and $\beta_0 = 1.0$ for their target-using coordinate routes; the corresponding checkpoint configuration records a noise-batch setting of 16. Target-only starts from the target-support encoding, while source-initialised adaptation starts from z_S. In A4, the raw label `reuse z_S` refers to this source-initialised target adaptation. Correct-source transport remains at $J = 0$. These validation-control settings are separate from the full-FT budgets below.

The full-fine-tuning settings were selected on validation tasks and frozen before correct source + transport was evaluated on the final-test classes. They remain 25 steps at learning rate 10^{-5} for $K_T = 1$, 25 at 10^{-4} for $K_T = 5$, and 400 at 10^{-5} for $K_T = 20$. No test result selects a new budget or model.

The five-budget A3 refinement experiment compares matched on/off conditions on the 50k-step `cifar_ds3` checkpoint. It uses 300 validation episodes (60 at each K_T), matched target support/query sets and paired refinement/generation noise. The metric is transformation-statistic distance. Gain is transport alone minus transport plus 25 refinement steps; positive favours refinement. This separate diagnostic is not pooled with final-test SW or MMD.

Table 19: CIFAR validation-selected full-fine-tuning settings frozen before the matched final-test evaluation.

Target budget	Full-FT steps	Learning rate
$K_T = 1$	25	10^{-5}
$K_T = 5$	25	10^{-4}
$K_T = 20$	400	10^{-5}

Table 20: Matched refinement ablation on the correct source + transport checkpoint. $J = 0$ remains the main route; the ablation uses $J = 25$, learning rate 0.01 and the same prior/noise budget. Half-widths are $1.96\,\mathrm{SE}$ over 60 paired class–transformation episodes per budget; stars use unrounded gain intervals.

	$K_T = 1$	$K_T = 2$	$K_T = 5$	$K_T = 10$	$K_T = 20$
Refinement gain	-0.0244^*	-0.0071	-0.0267	$+0.0144$	-0.0065
95% half-width	0.0136	0.0169	0.0320	0.0348	0.0434

The gain interval is negative at $K_T = 1$: refinement worsens transformation-statistic distance in that comparison. The other four intervals contain zero; this establishes neither equivalence nor universal absence of benefit. The separate source-initialised target-adaptation baseline starts from z_S and uses target evidence. Its behaviour cannot be inferred from the post-transport gains in this table, and it must not be called strict direct reuse.

C.4 Source budget and conditioning generality

The source-budget sweep shows small changes between $M_S = 16$ and 300, but not exact equivalence: the paired SW difference at $M_S = 64$ relative to 16 is $-0.00287 \pm 0.00236^*$. Other measured increments relative to 16 have intervals containing zero; no $M_S < 16$ was tested. The model was trained with $M_S = 64$ and evaluated at $K_T = 1$ with nested source prefixes and matched query/noise. Proposition 1 concerns coordinate estimation error, not equivalence of generated SW at all source sizes.

In the independent 180-episode FiLM validation package, SW for correct source + transport is $0.1973, 0.1942, 0.1987$ and fine-class consistency is $0.1306, 0.1278, 0.1311$ at $K_T = 1, 5, 20$. Mean-source SW is $0.2417, 0.2441, 0.2462$, shuffled-source $0.2567, 0.2578, 0.2616$, and relation-only $0.2366, 0.2373, 0.2414$. All 9/9 SW and 9/9 semantic main-relative intervals favour the correct source. Paired consistency gains are $0.0740 \pm 0.0234^*$, $0.0786 \pm 0.0253^*$, $0.0780 \pm 0.0245^*$ against mean source, $0.0835 \pm 0.0233^*$, $0.0799 \pm 0.0258^*$, $0.0835 \pm 0.0255^*$ against shuffled source, and $0.0757 \pm 0.0218^*$, $0.0773 \pm 0.0238^*$, $0.0773 \pm 0.0220^*$ against relation-only. These source controls do not test the full FiLM target-only/full-FT crossover.

Table 21: Three independent matched CIFAR SW packages. Upper block: correct source + transport with two conditioning architectures. Lower blocks: transport-structure ablations, with direct $J = 0$ generation in both compared arms. Pairing is within block only; all intervals are over validation episodes at one training seed per arm.

Configuration / comparison	$K_T = 1$	$K_T = 5$	$K_T = 20$
Basis: correct source + transport	0.2067 ± 0.0192	0.2090 ± 0.0177	0.2136 ± 0.0196
FiLM: correct source + transport	0.1973 ± 0.0206	0.1942 ± 0.0164	0.1987 ± 0.0193
FiLM − basis	-0.0094 ± 0.0099	$-0.0148 \pm 0.0107^*$	$-0.0149 \pm 0.0107^*$
Basis residual MLP	0.2102 ± 0.0190	0.2109 ± 0.0165	0.2138 ± 0.0193
Basis linear	0.2124 ± 0.0211	0.2113 ± 0.0183	0.2174 ± 0.0210
Linear − MLP	0.0022 ± 0.0049	0.0004 ± 0.0048	0.0036 ± 0.0050
FiLM nonlinear	0.2066 ± 0.0224	0.2054 ± 0.0185	0.2074 ± 0.0214
FiLM linear	0.1973 ± 0.0206	0.1942 ± 0.0164	0.1987 ± 0.0193
Nonlinear − linear	$0.0093 \pm 0.0037^*$	$0.0112 \pm 0.0040^*$	$0.0087 \pm 0.0042^*$

In the retained FiLM transport-structure ablation, nonlinear-minus-linear MMD differences are $0.2479 \pm 0.1658^*$, $0.2810 \pm 0.1588^*$, $0.1958 \pm 0.1640^*$, all favouring linear transport. They are paired effects, not raw MMD values. The matched conditioning comparison favours FiLM at $K_T = 5, 20$, with the other budgets unresolved. The basis and FiLM conditioning counts remain

55,344 and 118,784 (46.6% ratio). Their evaluated score paths have 27,798,195 and 27,916,979 parameters, respectively; the shared transport module adds the same 352 registered parameters (304 active). No latency, FLOP or memory conclusion is inferred.

C.5 META-TRAINING-TASK SCARCITY

Both task-scarcity arms use correct source + transport and $J = 0$, with 50% (30/60 classes, all 12 superclasses) and 25% training subsets. At 50%, FiLM-minus-basis SW effects are -0.0098 ± 0.0119, -0.0102 ± 0.0128, -0.0104 ± 0.0135; all three intervals contain zero. FiLM-minus-basis MMD effects are $-0.3484 \pm 0.3090^*$, $-0.5448 \pm 0.3468^*$, $-0.5468 \pm 0.4052^*$, and superclass-consistency effects are $0.0363 \pm 0.0211^*$, $0.0380 \pm 0.0232^*$, $0.0335 \pm 0.0225^*$. The MMD and consistency contrasts favour FiLM. At 25%, SW effects are 0.0027 ± 0.0076, 0.0007 ± 0.0080, -0.0003 ± 0.0076, all unresolved. These SW intervals do not establish an architecture ranking or a ranking reversal. Each arm has one training seed; episode intervals do not cover training-seed uncertainty.

C.6 COORDINATE RETRIEVAL AND DIRECT-TARGET DIAGNOSTICS

The frozen 50k EMA FiLM model with affine transport (checkpoint hash prefix `f7c1369c`) was probed on 60 seen and 20 held-out fine classes. Retrieval uses 20 fixed support resamples, with 64 source and 64 target images per class in each resample; candidates are confined to the corresponding seen or held-out class group. This diagnostic support pool is distinct from the default 20-image target-support pool used for the headline budget comparison. Its bootstrap resamples the 20 support draws conditional on the same classes, so significance describes stability to support choice and does not establish population-level class significance. With 20 held-out candidates, correct-source transport achieves L2 retrieval of 62.5%, 59.5%, and 49.0% for blur, noise, and contrast, respectively; raw z_S achieves 17.25%, 26.5%, and 14.0%, and wrong-source transport achieves 1.25%, .75%, and 2.25%. Tables 22 and 23 give retrieval rates, margins and paired contrasts for the reported class groups and transformations.

Table 22: Complete CIFAR frozen-coordinate retrieval. Each cell is top-1 percentage / mean nearest-wrong margin. Positive margins favour the requested class; relation-only chance/tie behaviour is not task discrimination.

Group	Relation	Metric	Raw z_S	Main	Wrong-source + transport	Relation only
Seen	Blur	L2	$36.50/-0.01050$	$71.50/+0.02241$	$0.42/-0.24324$	$1.67/-0.13592$
		Cosine	$26.58/-0.06849$	$72.58/+0.01675$	$0.42/-0.24228$	$1.67/-0.10330$
Seen	Noise	L2	$56.83/+0.00548$	$74.50/+0.02872$	$0.33/-0.26493$	$1.67/-0.13956$
		Cosine	$30.50/-0.04843$	$78.08/+0.01791$	$0.25/-0.24565$	$1.67/-0.09906$
Seen	Contrast	L2	$36.00/-0.01447$	$68.83/+0.02719$	$0.50/-0.26657$	$1.67/-0.12781$
		Cosine	$25.58/-0.07178$	$72.08/+0.02048$	$0.67/-0.28837$	$1.67/-0.11249$
Held-out	Blur	L2	$17.25/-0.02427$	$62.50/+0.01114$	$1.25/-0.12900$	$5.00/-0.08812$
		Cosine	$16.75/-0.13231$	$64.50/+0.00806$	$1.25/-0.11517$	$5.00/-0.07663$
Held-out	Noise	L2	$26.50/-0.01098$	$59.50/+0.00985$	$0.75/-0.11476$	$5.00/-0.08370$
		Cosine	$15.50/-0.09900$	$60.00/+0.00693$	$0.00/-0.09392$	$5.00/-0.06372$
Held-out	Contrast	L2	$14.00/-0.04431$	$49.00/-0.00423$	$2.25/-0.09975$	$5.00/-0.07584$
		Cosine	$10.25/-0.16097$	$48.25/-0.00313$	$2.00/-0.09921$	$5.00/-0.07595$

Table 23: CIFAR L2 retrieval gains in percentage points, main minus comparator. Intervals bootstrap the 20 paired support resamples conditional on the fixed class set; positive favours correct-source transport.

Group	Relation	Main − raw source (pp) [95% CI]	Main − wrong source (pp) [95% CI]
Seen	Blur	$+35.00\,[+32.25, +37.67]^*$	$+71.08\,[+69.08, +73.08]^*$
Seen	Noise	$+17.67\,[+15.33, +20.00]^*$	$+74.17\,[+72.25, +75.92]^*$
Seen	Contrast	$+32.83\,[+30.67, +35.08]^*$	$+68.33\,[+66.58, +70.08]^*$
Held-out	Blur	$+45.25\,[+42.25, +48.25]^*$	$+61.25\,[+58.25, +64.00]^*$
Held-out	Noise	$+33.00\,[+27.25, +38.25]^*$	$+58.75\,[+54.50, +63.25]^*$
Held-out	Contrast	$+35.00\,[+30.25, +39.50]^*$	$+46.75\,[+43.00, +50.50]^*$

Direct-target generation and timestep probes. For direct-z_T generation, a separate post-hoc repartition supplies 200 target-support images and 50 disjoint real target queries per class; it does not reuse the headline 20/230 target-support/query partition. Each class has 96 generated samples using 50-step DDIM. The frozen DINO nearest-prototype evaluator ranks all 80 seen-plus-held-out classes and is calibrated on the diagnostic real queries. These identity percentages are not directly comparable with headline semantic consistency because both the partition and the identity instrument differ. Generation and timestep intervals resample classes, not the fixed support draws used for

retrieval. Tables 24 and 25 report the raw metrics and the descriptive held-out-minus-seen identity contrasts.

Table 24: CIFAR direct-target generation using abundant target support. Identity rates are percentages; SW and MMD use the diagnostic DINO feature space, not the headline pixel-space metrics.

Group	Relation	Generated identity (%)	Real identity (%)	DINO-SW ↓	DINO-MMD ↓
Seen	Blur	5.52	50.80	0.02176	0.15937
Seen	Noise	3.78	29.27	0.01671	0.09777
Seen	Contrast	6.84	78.10	0.02983	0.27962
Held-out	Blur	3.59	51.20	0.02338	0.18884
Held-out	Noise	3.85	34.50	0.01888	0.13402
Held-out	Contrast	2.71	70.90	0.03059	0.28967

Table 25: CIFAR direct-target generated-identity differences, held-out minus seen, in percentage points. The two different class groups are independently resampled; these are descriptive group contrasts, not paired causal effects.

Relation	Identity Δ (pp) [95% CI]
Blur	$-1.93\,[-6.70, +4.11]$
Noise	$+0.07\,[-4.34, +6.18]$
Contrast	$-4.13\,[-7.99, +0.14]$

The timestep probe pairs correct and cyclic wrong-class target coordinates at identical real query, noise and timestep. The five bins contain 200 diffusion steps each, with 800–999 the highest-noise bin. In Table 26, positive $\Delta_z = \mathcal{L}_\epsilon(z_{\text{wrong}}) - \mathcal{L}_\epsilon(z_{\text{correct}})$ favours the correct target coordinate; this does not specifically test transport. MSE values are unscaled. Stars follow the source report's unrounded class-bootstrap limits, including limits displayed as zero after rounding.

Table 26: CIFAR coordinate use across diffusion time: absolute denoising MSE and wrong-minus-correct differences with class-bootstrap 95% intervals.

Group	Relation	Timesteps	Correct MSE	Wrong MSE	Δ_z [95% CI]
Seen	Blur	0–199	0.069118	0.069843	$+0.00073\,[+0.00058, +0.00089]^*$
		200–399	0.025385	0.026199	$+0.00081\,[+0.00065, +0.00099]^*$
		400–599	0.014950	0.015695	$+0.00074\,[+0.00061, +0.00089]^*$
		600–799	0.008320	0.008864	$+0.00054\,[+0.00045, +0.00065]^*$
		800–999	0.002333	0.002566	$+0.00023\,[+0.00019, +0.00028]^*$
Seen	Noise	0–199	0.575592	0.576098	$+0.00051\,[+0.00040, +0.00062]^*$
		200–399	0.149464	0.150564	$+0.00110\,[+0.00087, +0.00135]^*$
		400–599	0.055730	0.056779	$+0.00105\,[+0.00083, +0.00129]^*$
		600–799	0.020958	0.021725	$+0.00077\,[+0.00063, +0.00092]^*$
		800–999	0.004284	0.004563	$+0.00028\,[+0.00023, +0.00033]^*$
Seen	Contrast	0–199	0.097140	0.098230	$+0.00109\,[+0.00084, +0.00137]^*$
		200–399	0.025000	0.026042	$+0.00104\,[+0.00082, +0.00130]^*$
		400–599	0.011904	0.012583	$+0.00068\,[+0.00055, +0.00083]^*$
		600–799	0.005251	0.005642	$+0.00039\,[+0.00032, +0.00047]^*$
		800–999	0.001187	0.001322	$+0.00014\,[+0.00011, +0.00016]^*$
Held-out	Blur	0–199	0.079225	0.079248	$+0.00002\,[-0.00008, +0.00012]$
		200–399	0.030886	0.030951	$+0.00006\,[+0.00000, +0.00013]^*$
		400–599	0.018033	0.018145	$+0.00011\,[+0.00005, +0.00018]^*$
		600–799	0.009747	0.009893	$+0.00015\,[+0.00005, +0.00026]^*$
		800–999	0.002668	0.002732	$+0.00006\,[+0.00002, +0.00012]^*$
Held-out	Noise	0–199	0.584385	0.584397	$+0.00001\,[-0.00004, +0.00006]$
		200–399	0.157766	0.157801	$+0.00004\,[-0.00005, +0.00013]$
		400–599	0.060890	0.060990	$+0.00010\,[+0.00005, +0.00016]^*$
		600–799	0.023057	0.023172	$+0.00012\,[+0.00004, +0.00021]^*$
		800–999	0.004634	0.004695	$+0.00006\,[+0.00002, +0.00011]^*$
Held-out	Contrast	0–199	0.111840	0.111861	$+0.00002\,[-0.00011, +0.00016]$
		200–399	0.030179	0.030231	$+0.00005\,[-0.00006, +0.00015]$
		400–599	0.013888	0.013959	$+0.00007\,[-0.00000, +0.00015]$
		600–799	0.005905	0.005997	$+0.00009\,[+0.00003, +0.00017]^*$
		800–999	0.001368	0.001395	$+0.00003\,[+0.00001, +0.00005]^*$

Bypassing transport does not guarantee semantic generation. Direct target coordinates estimated from 200 target images give held-out generated identity of 3.59% (blur), 3.85% (noise), and 2.71% (contrast), compared with real-query identity of 51.2%, 34.5%, and 70.9%. Diffusion-time probes nevertheless show that correct target coordinates lower denoising MSE in most held-out timestep bins, including all three highest-noise bins. The evidence therefore supports coordinate use without implying that coordinate use alone guarantees class-correct generation.

D OFFICE-HOME: FULL NATURAL-DOMAIN STRESS TEST

D.1 PROTOCOL AND EVALUATOR AUDIT

The corrective Office-Home experiment uses the Product→Clipart relation, 45/10/10 train/validation/test semantic classes, 64×64 images, FiLM conditioning, linear transport and the effective-batch-64 training run. Ten held-out classes are the independent inference units; eight generation repeats and wrong-source assignments are averaged within class before the 10,000-draw class bootstrap.

Table 27: Office-Home real-image evaluator audit.

Check	Result
CLIP top-1, Product / Clipart	89.03% / 65.96%
Test classes with Clipart top-1 $\geq 30\%$	10/10
DINOv2 held-out domain accuracy	98.30%
Same-class centroid top-1 / median rank	100.0% / 1
Same-minus-different centroid cosine gap	.5077

D.2 VALIDATION-ONLY CHECKPOINT AND ADAPTATION SELECTION

All Office-Home model and adapting-baseline choices were fixed on the ten validation classes before the outer test. Validation DINO-SW decreased monotonically across the saved checkpoint sequence, and the 50,000-step checkpoint was selected for the final route comparisons.

Table 28: Office-Home validation-only selection record. The upper block reports the main-route checkpoint trajectory. The lower block reports the budget-specific target-only and full-fine-tuning settings subsequently frozen for final testing. These validation values are not pooled with test results.

Checkpoint step	Main $K_T = 0$ DINO-SW
20,000	.040525
25,000	.039609
30,000	.038712
35,000	.038019
40,000	.037572
45,000	.037351
50,000	.037283

K_T	Target steps	Target lr	Target val SW	Full-FT steps	Full-FT lr	Full-FT val SW
1	5	.030	.037660	100	3×10^{-5}	.036521
5	5	.010	.038021	400	3×10^{-5}	.034206
10	5	.030	.037397	400	3×10^{-5}	.032314

A separate validation-only zero-target code-path check gave DINO-SW .037239 for correct-source transport, .038168 for wrong-source transport, .037950 for relation-only prediction and .040021 for source reuse. These values are selection/audit evidence only; all Office-Home headline superiority claims use the held-out test route matrices below.

D.3 ZERO-TARGET AND TARGET-BUDGET RESULTS

Table 29: Office-Home $K_T = 0$ route matrix. Distance metrics are lower-is-better; requested-class and Clipart fractions are higher-is-better.

Route	DINO-SW	MMD	Object ID (%)	Clipart (%)
Correct + transport	.038969	.407846	1.67	86.50
Wrong + transport	.039394	.419016	1.44	86.27
Relation-only	.039251	.415280	1.37	86.71
Source reuse	.041641	.457524	1.18	63.80

The paired SW gain over wrong-source transport is .000425 [.000213, .000643]; over relation-only it is .000282 [.000109, .000456]; over source reuse it is .002673 [.002270, .003036]. Requested-class differences from wrong and relation-only are unresolved. Thus the zero-target claim is distributional/domain alignment rather than reliable semantic transfer.

Table 30: Office-Home target-budget route means. The same zero-target correct-source generations are reused across budgets; adapting baselines receive the stated target support. SW and MMD are lower-is-better; identity and Clipart rates are percentages.

K_T	Route	DINO-SW	MMD	Object ID (%)	Clipart (%)
1	Correct + transport	0.038969	0.407846	1.67	86.50
1	Target-only	0.040346	0.439860	2.23	75.73
1	Full FT	0.039336	0.468223	5.33	87.49
5	Correct + transport	0.038969	0.407846	1.67	86.50
5	Target-only	0.039031	0.409734	1.81	81.22
5	Full FT	0.034550	0.349838	12.55	83.11
10	Correct + transport	0.038969	0.407846	1.67	86.50
10	Target-only	0.039001	0.406858	1.61	82.79
10	Full FT	0.033859	0.321982	12.01	83.82

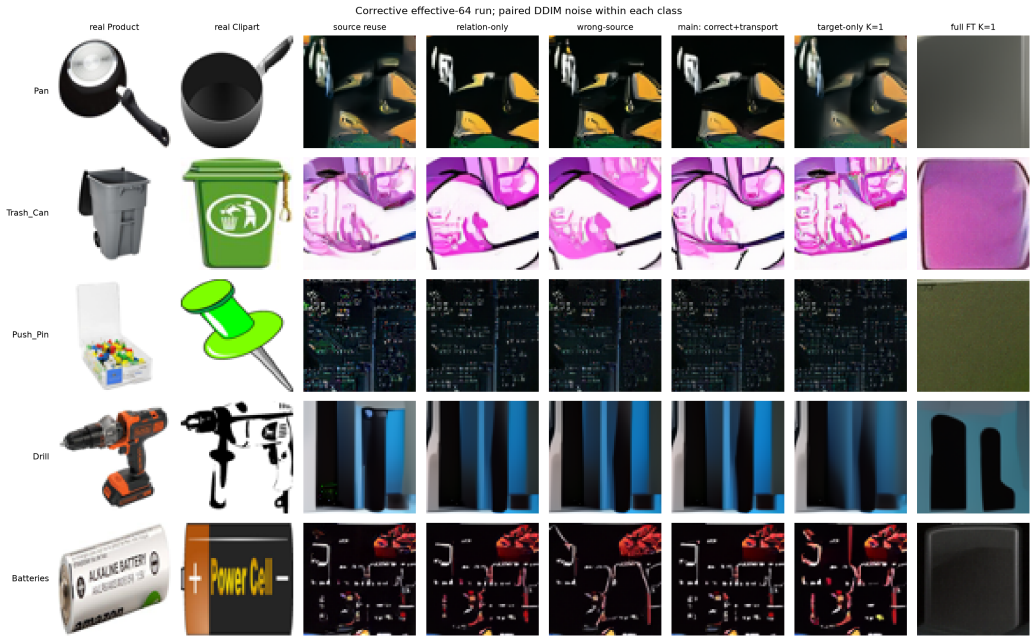

Figure 11: Office-Home generation grid. Rows are held-out classes. Columns contain real Product, real Clipart, source reuse, relation-only, wrong-source transport, correct-source transport, one-shot target-only and one-shot full fine-tuning. Generated routes share DDIM noise within class.

At $K_T = 1$, transport improves DINO-SW over target-only with paired difference .001377 $[.000568, .002311]$; its SW comparison with full fine-tuning is unresolved, although its MMD is better. At $K_T = 5$ and 10, full fine-tuning is significantly better in SW, MMD and requested-object accuracy.

Complete main-relative uncertainty. For distance metrics, Δ is comparator minus main; for requested identity and Clipart share it is main minus comparator. Thus positive values always favour correct-source transport in Tables 31 and 32. Eight generation repeats and wrong-source assignments are averaged within each of the ten held-out classes before a paired 10,000-draw class bootstrap. Identity/domain differences are percentage points. Stars mark unrounded 95% intervals excluding zero in either direction; the tests are exploratory and unadjusted for multiplicity.

D.4 Localising the semantic bottleneck

Three frozen-checkpoint probes and a separately trained class-conditioning capacity control probe where object identity is lost.

The cross-domain retrieval probe uses 20 fixed support resamples over the ten held-out classes, with 16 Product images and at most 32 disjoint Clipart support images per class. Retrieval is among ten target-class coordinates. L2 margins compare the correct target distance with the nearest wrong target; cosine margins compare corresponding similarities, with positive margins favourable in both

Table 31: Office-Home paired distributional gains with 95% class-bootstrap intervals. Positive favours correct-source transport; negative favours the comparator.

K_T	Comparator	Δ DINO-SW [95% CI]	Δ MMD [95% CI]
0	Wrong + transport	0.000425^* [0.000213, 0.000643]	0.011170^* [0.006870, 0.015706]
0	Relation-only	0.000282^* [0.000109, 0.000456]	0.007434^* [0.002360, 0.012075]
0	Source reuse	0.002673^* [0.002270, 0.003036]	0.049678^* [0.039741, 0.060430]
1	Target-only	0.001377^* [0.000568, 0.002311]	0.032014^* [0.015413, 0.047276]
1	Full FT	0.000367 [-0.001578, 0.002312]	0.060376^* [0.019137, 0.101637]
5	Target-only	0.000062 [-0.000754, 0.000800]	0.001888 [-0.019453, 0.018779]
5	Full FT	-0.004419^* [-0.006279, -0.002922]	-0.058008^* [-0.097989, -0.024569]
10	Target-only	0.000032 [-0.000685, 0.000719]	−0.000988 [-0.021510, 0.016291]
10	Full FT	-0.005110^* [-0.006639, -0.003887]	-0.085864^* [-0.124581, -0.053092]

Table 32: Office-Home paired identity and domain-share gains with 95% class-bootstrap intervals, in percentage points. Positive favours correct-source transport.

K_T	Comparator	Δ object ID [95% CI]	Δ Clipart [95% CI]
0	Wrong + transport	0.230 [-0.043, 0.543]	0.226 [-0.438, 0.955]
0	Relation-only	0.299 [-0.078, 0.885]	−0.208 [-0.885, 0.443]
0	Source reuse	0.482 [-0.221, 1.406]	22.695^* [20.820, 24.661]
1	Target-only	−0.560 [-2.031, 0.404]	10.768^* [2.760, 19.062]
1	Full FT	−3.659 [-8.737, 0.195]	−0.990 [-4.844, 2.734]
5	Target-only	−0.143 [-0.703, 0.286]	5.273^* [1.667, 9.532]
5	Full FT	-10.885^* [-20.261, -3.984]	3.385 [-1.289, 8.581]
10	Target-only	0.052 [-0.638, 0.573]	3.711^* [0.482, 7.161]
10	Full FT	-10.339^* [-18.229, -3.802]	2.682 [-1.289, 8.815]

cases. Table 34 averages support resamples within class and reports paired class-bootstrap contrasts. This target-coordinate task is distinct from the within-Product retrieval summary in the main text.

Correct transport significantly improves L2 retrieval over raw z_S and both L2/cosine retrieval over relation-only, but the correct-versus-wrong transport class intervals include zero. This is partial coordinate alignment rather than a resolved source-specific retrieval result.

For direct-target generation, all available Clipart images outside 20 fixed real queries per class form the target support. The frozen EMA checkpoint (hash prefix `44e2710a`) generates 96 images per class using 50-step DDIM. Seen classes were used in original Meta-Diff training, and their query images were not necessarily excluded from that training. This is a target-coordinate diagnostic, not the zero-target main route. The seen and held-out groups contain different classes, so their identity contrast independently resamples the two groups rather than pairing them.

Bypassing transport therefore does not restore object identity. The held-out-minus-seen generated-identity contrast is $-.82$ percentage points with 95% interval $[-2.45, .76]$.

The diffusion-time probe replaces the correct target coordinate with a cyclic wrong-class coordinate while holding query image, noise and timestep fixed. On seen classes, wrong-minus-correct MSE is positive in all five timestep bins; on held-out classes it is positive with intervals excluding zero in three of five bins. The highest-noise differences are only 5.5×10^{-5} (seen) and 1.7×10^{-5} (held-out), showing weak but nonzero coordinate use.

Finally, an explicit 16-D learned class embedding provides a separately trained capacity control with the same FiLM interface. On seen Clipart classes, generated identity rises from 2.38% with direct learned z_T to 6.90%, a paired gain of 4.51 percentage points $[2.85, 6.37]$; DINO-SW and MMD also improve. Because 6.90% remains far below the 50.33% real-image calibration, the control demonstrates partial semantic conditioning capacity rather than a complete solution.

The capacity control is newly initialised, has 45 learned 16-D training-class embeddings, and uses neither the set encoder nor transport. It completes 50,000 updates with 64 independently noised observations per update; evaluation uses the same 20 fixed real queries and 96 generated images per class with DDIM-50. Those query IDs are explicitly withheld from this new control, whereas they may have appeared during training of the original direct-z_T model. This exposure difference limits causal attribution of the gain to the conditioning representation alone. Tables 37 and 38 report the raw metrics and contrasts; positive effects in the latter favour the explicit class control, not Meta-Diff's main route.

Table 33: Office-Home target-coordinate retrieval on ten held-out classes.

Query route	L2 top-1	Cosine top-1	L2 margin	Cosine margin
Raw z_S	12.0%	26.5%	-.0500	-.1749
Correct source + transport	42.0%	44.5%	-.0153	-.0620
Wrong-source transport	20.0%	19.0%	-.0897	-.2654
Relation-only	10.0%	10.0%	-.0545	-.1510

Table 34: Office-Home target-coordinate retrieval gains, main minus comparator, in percentage points with paired class-bootstrap 95% intervals.

Comparator	L2 Δ [95% CI]	Cosine Δ [95% CI]
Raw z_S	$+30.00\,[+4.00, +58.50]^*$	$+18.00\,[-2.50, +43.50]$
Wrong-source transport	$+22.00\,[-11.50, +56.00]$	$+25.50\,[-10.00, +60.00]$
Relation only	$+32.00\,[+5.50, +59.50]^*$	$+34.50\,[+7.00, +61.50]^*$

Table 35: Office-Home direct-z_T generation using abundant non-query Clipart support.

Group	Generated ID	Real ID	DINO-SW	DINO-MMD	Clipart
Seen (45 classes)	2.38%	50.33%	.03679	.35175	87.34%
Held-out (10 classes)	1.56%	52.00%	.03902	.40642	88.33%

Table 36: Office-Home direct-target coordinate use across diffusion time. MSEs use identical query/noise/timestep pairs; positive wrong-minus-correct differences favour the correct coordinate. Intervals resample classes.

Group	Timesteps	Correct MSE	Wrong MSE	Δ_z [95% CI]
Seen	0–199	0.056513	0.056724	$+0.000211\,[+0.000141, +0.000294]^*$
	200–399	0.019655	0.019785	$+0.000131\,[+0.000086, +0.000183]^*$
	400–599	0.013028	0.013164	$+0.000136\,[+0.000089, +0.000189]^*$
	600–799	0.008825	0.008963	$+0.000138\,[+0.000095, +0.000187]^*$
	800–999	0.003287	0.003342	$+0.000055\,[+0.000039, +0.000072]^*$
Held-out	0–199	0.059910	0.059928	$+0.000018\,[-0.000044, +0.000073]$
	200–399	0.022160	0.022180	$+0.000020\,[-0.000070, +0.000098]$
	400–599	0.014911	0.014948	$+0.000037\,[+0.000006, +0.000068]^*$
	600–799	0.009441	0.009505	$+0.000063\,[+0.000006, +0.000144]^*$
	800–999	0.003140	0.003157	$+0.000017\,[+2.15\times10^{-7}, +0.000034]^*$

Table 37: Office-Home separately trained class-conditioning capacity control on 45 seen classes. Identity and Clipart columns are percentages.

Seen-class route	Generated identity (%)	Real identity (%)	DINO-SW ↓	DINO-MMD ↓	Clipart (%)
Original direct z_T	2.38	50.33	0.03679	0.35175	87.34
Explicit class embedding	6.90	50.33	0.03507	0.32768	87.45

Table 38: Office-Home capacity-control effects with 95% intervals paired over 45 classes. Identity/domain gains are control minus direct-z_T; SW/MMD reductions are direct-z_T minus control.

Metric	Favor-control Δ [95% CI]
Requested-class identity	$+4.51$ pp $[+2.85, +6.37]^*$
DINO-SW reduction	$+0.00172\,[+0.00121, +0.00229]^*$
DINO-MMD reduction	$+0.02407\,[+0.01332, +0.03671]^*$
Clipart-share increase	$+0.12$ pp $[-1.30, +1.53]$

E Clinical and External-Domain Stress Tests

These studies are retained for transparency and boundary analysis. They are not used as positive evidence for clinical or demographic source-specific transfer.

E.1 ChestX-ray14 age-group transfer

Tasks are single-positive thoracic findings, source ages are 45–60, and target ages are 0–30, 30–45 and 60–75. The fixed 4/2/3 finding split uses whole-patient allocation. Evaluation covers three held-out findings, three age relations and four resamples, with $M_S = 64$.

Table 39: ChestX-ray14 SW route means. Lower is better. The independently reported wrong-minus-correct paired interval at $K_T = 1$ is $7.7 \times 10^{-6} \pm 2.3 \times 10^{-5}$; other displayed route differences lack paired covariance information.

Method	$K_T = 1$	2	5	10	20
No adaptation	.3696	.3696	.3696	.3696	.3696
Target-only	.1397	.1406	.1398	.1405	.1408
Relation-only	.3680	.3680	.3680	.3680	.3680
Wrong-source transport	.1407	.1407	.1407	.1407	.1407
Correct source + transport	.1407	.1407	.1407	.1407	.1407
Full fine-tuning	.2981	.2802	.2304	.1800	.1591

Correct-source transport reaches the broad target-image distribution far better than no adaptation or relation-only, but wrong-source transport has the same rounded mean and the available paired correct-versus-wrong contrast is unresolved. The finding classifier has only 12.1% accuracy and an abundant-target age diagnostic has slope $.039 \pm .064$, so neither image distance nor predictor output validates the requested age/finding identity. This branch therefore establishes broad distribution matching, not source-specific demographic transfer.

E.2 Fitzpatrick17k disease and pathology-family transfer

All 16,577 official rows were recovered from the SFU renamed archive, matched to the official CSV by MD5, and filtered through the published cleaned allow-list. The retained set contains 11,394 images, of which 11,098 have Fitzpatrick type I–VI. The disease-level split is 8/4/4 train/validation/test diseases. Pathology-family follow-ups use six test families for I–III→IV–VI and five for I–III→V–VI.

Table 40: Fitzpatrick zero-target route means. Invalidated phenotype/semantic instruments are omitted here; statistic distance, SW and energy MMD are lower-is-better.

Protocol	Route	Statistic	SW	MMD
Disease IV–VI	Source reuse	2.5248	.3642	7.1765
	Wrong disease, same family	2.5020	.3638	7.1466
	Wrong disease, different family	2.5025	.3639	7.1444
	Correct source + transport	2.5030	.3638	7.1429
Family IV–VI	Source reuse	1.7452	.3081	5.7946
	Wrong-family transport	1.7185	.3064	5.7468
	Correct source + transport	1.7188	.3063	5.7436

Disease-level correct-source and wrong-source routes show no resolved difference on the valid distributional metrics under the reported intervals; this is not an equivalence result. The same conclusion holds at pathology-family resolution: correct source has no resolved advantage over wrong-family transport. The stronger I–III→V–VI follow-up also fails to recover source specificity; the wrong-task-minus-correct paired effects are $-.00068 \pm .00273$ for statistic distance, $-.00003 \pm .00134$ for SW and $-.00107 \pm .03907$ for MMD.

The interpretation is further constrained by the evaluators. The pathology-family semantic gate fails on every outer fold in the stronger split, and an exploratory phenotype audit finds real-target recall of 0% for IV–VI and 0–6.3% for V–VI despite high aggregate accuracy. A source-only disease-separability analysis reports $\rho = -.775$ over four diseases, but three of the four semantic effects are exactly zero and only 6,793/10,000 bootstrap draws have defined rank variance. We therefore do not use task separability as the central explanation for the Fitzpatrick result. Downstream demographic augmentation was not run because the pre-specified positive family-semantic gate was not met.

E.3 CRC HISTOLOGY PREFLIGHT

CRC uses NCT-CRC-HE-100K as source and CRC-VAL-HE-7K as external target over nine tissue classes and three outer folds. A prespecified gate required both a valid tissue evaluator and sufficient source–target domain separation before any Meta-Diff training.

Table 41: CRC real-image preflight and stopping decision.

Check	Result
Target tissue balanced accuracy	88.25%
Minimum target tissue recall	60.57%
Domain balanced accuracy, fold 1	59.01%
Domain balanced accuracy, fold 2	58.24%
Domain balanced accuracy, fold 3	57.00%
Same-tissue centroid retrieval	100% top-1 in each fold
Diffusion training / generation	Not run: structural gate failed

The tissue evaluator and cross-domain tissue alignment pass, but all three domain-separation folds fall below the locked 70% threshold. No CRC generation result exists. The stop therefore concerns benchmark suitability under the frozen representation and provides no evidence for or against Meta-Diff generation performance.

F IMPLEMENTATION AND REPRODUCIBILITY

The CIFAR score model uses a 128-base-channel U-Net and $k = 16$ coordinates; its encoder has approximately 0.33M parameters. The basis contributes 55,344 conditioning parameters. The actual three-relation linear transport is $Az_S + b_c$: $16^2 + 3 \times 16 = 304$ active parameters. Its saved module has 352 registered parameters because a 48-value compatibility embedding is not used by its forward map. A single-relation affine map has $16^2 + 16 = 272$ parameters. The headline CIFAR correct source + transport route directly decodes the predicted coordinate at all K_T, without target input, coordinate refinement or backbone update. Office-Home retains its separate 64-base-channel, 64×64 FiLM model, DDIM-50 and no-refinement headline protocol.

Semantic splits are fixed and target support/query images are disjoint. Within each paired CIFAR generation or ablation package, routes share the task, support/query episode and generation noise; the reported episode-level $1.96\,\mathrm{SE}$ intervals retain nested class dependence. The independent frozen-coordinate retrieval diagnostic instead bootstraps fixed support resamples conditional on the class set, and its direct-target generation and timestep diagnostics resample classes. The direct-target diagnostic uses the explicitly stated post-hoc support/query repartition. Independent conditioning packages are not pooled. Validation-selected full-FT budgets are frozen before matched final-test evaluation, and no test score selects a new model, budget or subset. Other benchmarks retain their own frozen protocols and statistical units.

The paper distinguishes final-model interventions from historical evidence. The GMM calibration table, two-panel interpolation profile and main-text own/mean/shuffled/zero generation grid use their separately identified checkpoints; these diagnose coordinates, not deployment superiority. Historical GMM capacity/relatedness aggregates remain accurately labelled in the appendix, and transport-structure ablations compare MLP and linear maps. The refinement table reports matched on/off measurements on the correct source + transport checkpoint. Frozen-coordinate and direct-target semantic probes are reported as independent diagnostics.