WHAT GOVERNS UNIFIED COMPREHENSION AND GENERATION ON GRAPHS? A CONTROLLED STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision and language have converged on unified models that both understand and generate, yet graph pretraining usually discards its generative decoder and deploys the encoder alone. We ask what governs a single graph model's ability to support both tasks. A controlled evaluation of a compact graph autoencoder across five domains and eight training conditions separates three effects. First, the generation degradation seen with frozen decoders comes from refitting the latent prior on the fine-tuned encoder. It shrinks as a weight anchor limits encoder drift, and the tested generator families shift the trade-off less than this prior deployment on three of five domains. Second, a null-calibrated learned metric exposes generation discrepancies that descriptor MMD hides and changes which conditions are not significantly dominated. Third, the comprehension gain of retaining the decoder, significant only on Synthetic and CiteSeer, is reproduced by a weight anchor with no generative loss on Synthetic but not on CiteSeer. Only an active generative loss reduces the mismatch between the current encoder and the deployed decoder relative to frozen decoders, on every domain under descriptor MMD and outside QM9 under C2ST, without pinning the encoder to its pretrained weights.

1 INTRODUCTION

Graph self-supervised learning has settled on a standard pipeline. It pretrains a graph neural network encoder with a generative pretext task such as masked feature reconstruction (Hou et al., 2022) or attribute-edge prediction (Hu et al., 2020b), then discards the decoder and fine-tunes the encoder for downstream classification. This pretrain-then-discard convention is now the default. Masked graph autoencoders pretrain with a decoder and discard it (Hou et al., 2022; 2023), while contrastive methods (Veličković et al., 2019; Zhu et al., 2020) and encoder-side graph foundation models (Liu et al., 2024; Zhao et al., 2025) forgo a generative decoder altogether. Either way, the deployed model carries no generation capability, because the generative pathway shapes the encoder during pretraining and is absent at deployment.

Discarding the decoder, however, forecloses a dual-use deployment in which one model serves both comprehension and generation. The trade-off introduced by this choice has rarely been examined, because the two tasks are studied by separate communities with separate models. In vision and language, the field has moved toward unified models that both comprehend and generate. Janus (Wu et al., 2025) decouples visual-encoder pathways for understanding and generation, Show-o (Xie et al., 2025) unifies autoregressive understanding with discrete diffusion, and Transfusion (Zhou et al., 2025) merges language modeling with continuous diffusion in one model. GenCeption (Wang et al., 2026) goes further, arguing that large-scale video generation is itself the pretraining paradigm for general-purpose vision, so that the generative pathway is the representation learner rather than an auxiliary pretraining component. Graph learning has adopted the opposite default without testing it. Graph generators (You et al., 2018; Jo et al., 2022; Vignac et al., 2023) are standalone systems with no comprehension capability, and comprehension-side graph models keep discarding their decoders.

Whether this unification transfers to graphs is not obvious. Graphs are discrete and permutation-invariant. Comprehension pools node embeddings into a single graph-level label, whereas generation must reconstruct an entire adjacency structure, and the two objectives pull the shared encoder in different directions during fine-tuning. Rather than propose another unified graph model, we ask a question that comes first: in a compact dual-use graph model under controlled training and

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.

evaluation, how effective is unification, which mechanisms govern it, and which constraints should future graph models preserve? We answer this question through a controlled evaluation with safeguards against metric sensitivity and deployment confounds. We complement descriptor MMD with a learned classifier two-sample test (C2ST) calibrated against a density-matched random-graph null that the test must reject. We also vary the generative mechanism and graph scale, so that no conclusion rests on a single setting.

Our main claim is that, in the compact graph autoencoders we test, the trade-off traces more to interfaces than to the generator family. Explanations in vision and language usually point to architecture, representation granularity, or the modeling objective, and we do not compare modalities. A deployed generator separates into a latent prior and a decoder, and comprehension fine-tuning moves the encoder that defines the prior. Caching the prior of a frozen decoder restores the Phase-1 generator by construction, so we measure how the cost of the post-hoc refit grows with encoder drift, whether a weight-space anchor reproduces the comprehension gain of retention, and how compatible the current encoder and deployed decoder remain. The experimental grid uses a compact graph autoencoder across five graph domains and eight conditions. It crosses decoder handling with refit or cached-prior generator deployment, compares all methods over 10 paired seeds with paired-bootstrap 95% confidence intervals, and extends the same logic to learned metrics, generator families, ogbn-arxiv scale, and a single-GPU external generator. This evaluation yields five findings.

- **Null-calibrated C2ST changes the generation assessment.** Descriptor MMD understates the gap. Under C2ST, active-loss generators fall significantly below the density-matched null on synthetic and citation graphs yet stay at least 0.21 above the real-versus-real floor on citation graphs, and on QM9 only JOINT-RETAINED significantly clears the null, by less than 0.01.
- **The refit penalty tracks encoder drift.** Refitting a frozen decoder's prior after fine-tuning raises C2ST by 0.04 to 0.29. The cached prior or a tight anchor removes most of it.
- **Anchoring reproduces the comprehension gain except on CiteSeer.** Retention helps significantly only on Synthetic and CiteSeer. An L2-SP anchor reproduces the Synthetic gain and also gains on Cora, but on CiteSeer only a generative loss that trains the decoder helps.
- **An active generative loss keeps encoder and decoder compatible.** Keeping the generative loss active during Phase 2, with or without decoder updates, reduces the round-trip mismatch between the current encoder and the deployed decoder relative to frozen decoders, on every domain under descriptor MMD and outside QM9 under C2ST. This compatibility is relative, and per graph it covers node attributes but not edges.
- **Generator families matter less than prior deployment.** Diffusion and VAE priors shift C2ST significantly less than prior deployment on three of five domains.

2 RELATED WORK

This section covers the lines of work the argument needs. Appendix A gives the full discussion. Graph self-supervised learning often trains a decoder but rarely deploys one. Generative pretext methods such as VGAE (Kipf & Welling, 2016), GraphMAE (Hou et al., 2022), GraphMAE2 (Hou et al., 2023), MaskGAE (Li et al., 2023), and GPT-GNN (Hu et al., 2020b) use reconstruction or attribute-edge prediction to shape an encoder, while contrastive methods such as DGI (Veličković et al., 2019) and GRACE (Zhu et al., 2020) build no decoder at all. Recent graph foundation models likewise emphasize transferable encoders (Liu et al., 2024; Zhao et al., 2025). Our study treats the discard-or-retain choice itself as the experimental variable.

Graph generation and graph comprehension are usually evaluated as separate capabilities. Dedicated generators such as GraphRNN (You et al., 2018), GDSS (Jo et al., 2022), and DiGress (Vignac et al., 2023) do not deploy a comprehension head, while unified graph systems such as GraphsGPT (Gao et al., 2024), Latent Graph Diffusion (Zhou et al., 2024), and InstructGraph (Wang et al., 2024a) validate capabilities at scale. We instead isolate decoder retention with paired seeds and explicit regularization and update-count controls.

Unified vision and language models motivate the comparison but not the answer. Janus (Wu et al., 2025), Show-o (Xie et al., 2025), and Transfusion (Zhou et al., 2025) attribute comprehension-generation trade-offs to representation pathways or objective design, and GenCeption (Wang et al., 2026) argues that generation can be the representation learner itself. We do not test these explana-

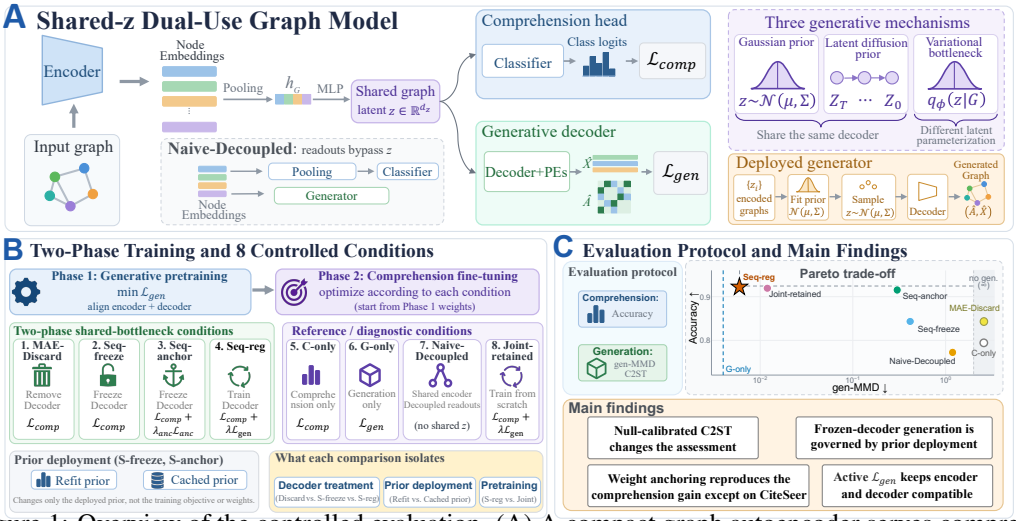


Figure 1: Overview of the controlled evaluation. (A) A compact graph autoencoder serves comprehension and generation through one shared latent. (B) Two-phase training yields eight controlled conditions that vary decoder treatment and prior deployment. (C) A fixed protocol scores both tasks, and the main findings preview the governing factors.

tions across modalities, and we evaluate generation with both descriptor MMD and a null-calibrated classifier two-sample test (Lopez-Paz & Oquab, 2017), motivated by known instability in descriptor-based graph metrics (O'Bray et al., 2022; Thompson et al., 2022).

3 Model and Controlled Conditions

We study models in which one set of weights supports two capabilities. Given a graph, the comprehension objective predicts a label, and the generation objective synthesizes new graphs whose summary statistics match a reference distribution. Write $\mathcal{L}_{\text{comp}}$ for the cross-entropy classification loss and $\mathcal{L}_{\text{gen}}$ for the graph reconstruction loss defined below. A method is single-model if one parameter vector $\boldsymbol{\theta}$ supports both capabilities at inference. A two-model reference trains separate comprehension and generation networks and serves only as a single-task comparison point.

A deployed generator consists of a latent prior and a decoder, which are specified by different choices. Decoder retention acts on the decoder, whereas comprehension fine-tuning moves the encoder that induces the prior. This decomposition identifies which component governs generation quality and structures the controlled conditions below. Our goal is to identify which single-model training conditions are not significantly dominated in comprehension accuracy and generation quality, under a controlled protocol whose per-condition budgets are stated in Sec. 3.2.

3.1 Model

To keep the controlled conditions comparable, we use one compact architecture throughout. Fig. 1 summarizes the model, condition design space, and evaluation protocol. A graph is given by an adjacency matrix $\mathbf{A}$ and node features $\mathbf{X}$, and a GIN (Xu et al., 2019) encoder $f_{\boldsymbol{\theta}}$ maps them to node embeddings $\mathbf{H} = f_{\boldsymbol{\theta}}(\mathbf{A}, \mathbf{X})$. Graphs of different sizes are batched by padding each to a shared node budget and carrying a real-node mask that confines message passing, the pooled readout, and the reconstruction loss to the occupied slots and leaves the aggregator unchanged.

Shared-z bottleneck. A linear projection on the pool of $\mathbf{H}$ produces a graph-level latent $\mathbf{z} \in \mathbb{R}^{d_z}$. The comprehension head is an MLP from $\mathbf{z}$ to K class logits. The generative decoder is an MLP from $\mathbf{z}$ plus positional embeddings to per-node features $\hat{\mathbf{X}}$, with bilinear edge logits $\hat{\mathbf{A}}$. Both heads read the same $\mathbf{z}$. The exception is NAIVE-DECOUPLED, which removes this shared bottleneck and uses decoupled readouts. This bottleneck is imposed so that both objectives operate through the same graph-level representation and the shared-path conditions can test coupling against the

decoupled-readouts pathway. The generative loss is a deterministic graph-reconstruction objective

$$\mathcal{L}_{\text{gen}}(\mathbf{A}, \mathbf{X}) = \frac{1}{nF} \sum_{i=1}^{n} \|\hat{\mathbf{X}}_i - \mathbf{X}_i\|^2 + \frac{1}{n(n-1)} \sum_{i \neq j} \text{BCE}(\hat{\mathbf{A}}_{ij}, \mathbf{A}_{ij}), \tag{1}$$

where n is the node number and F is the feature dimension. A samplable generator is obtained after training. We fit a Gaussian $\mathcal{N}(\mu, \Sigma)$ to the encoded latents, sample $\mathbf{z} \sim \mathcal{N}(\mu, \Sigma)$, decode, and threshold edges by Bernoulli sampling on $\sigma(\hat{\mathbf{A}})$, masking to a size drawn from the empirical reference-size distribution. The main-grid default is a deterministic autoencoder with a Gaussian latent prior fitted after training. Sec. 4.5 evaluates latent diffusion and a variational bottleneck under the same dual-use protocol and separates generator family from the prior and decoder factors.

Three generative mechanisms. The default mechanism estimates $\mathcal{N}(\mu, \Sigma)$ from encoded latents after training. The latent-diffusion prior instead trains a denoising diffusion model on the latents of the same checkpoint and decodes its samples with the same decoder, so the two differ only in the prior. The variational bottleneck replaces the deterministic latent with a sampled one and adds a KL term during training. It changes the latent parameterization and is trained on the SEQ-REG schedule.

Two prior specifications for generation. A graph autoencoder generator is a sampling distribution over $\mathbf{z}$ plus the decoder applied to those samples. The source of that distribution matters whenever comprehension fine-tuning separates the current encoder from a frozen decoder. The refit deployment fits $\mathcal{N}(\mu, \Sigma)$ on the fine-tuned encoder, the ex-post density estimation of Ghosh et al. (2020) applied after fine-tuning. The cached deployment reuses the prior snapshotted after generative pretraining, before the encoder drifts, so for a frozen decoder it recovers the pretrained generator. The cache stores a mean and a Cholesky factor. We report the cached deployment for the frozen-decoder conditions, where this interface mismatch is the experimental question.

Decoupled-readouts pathway for NAIVE-DECOUPLED. The NAIVE-DECOUPLED condition ablates the shared $\mathbf{z}$. The comprehension head reads $\text{softmax}(\text{MLP}(\text{pool}(\mathbf{H})))$, while the decoder reads $\mathbf{H}$ node-wise, with a per-node feature head and a bilinear edge head taken directly from $\mathbf{H}$. NAIVE-DECOUPLED keeps a shared encoder but separates the two readouts, which removes the shared graph-level bottleneck, a readout-side analogue of the encoder-side decoupling of Janus (Wu et al., 2025). At generation time, we fit a single Gaussian to the pooled node embeddings and sample each node embedding i.i.d. from it, then decode. This i.i.d. sampler makes NAIVE-DECOUPLED a pathway-decoupling diagnostic under the same compact protocol.

3.2 TWO-PHASE TRAINING PROTOCOL

The two-phase schedule aligns encoder and decoder by pretraining on $\mathcal{L}_{\text{gen}}$ alone. It then fine-tunes for comprehension. The experimental variable is how this phase treats the decoder and generative loss. The decoder may be discarded, frozen, frozen with an L2-SP anchor on the encoder, or kept trainable with its loss active. When the loss remains active, the Phase 2 objective becomes

$$\mathcal{L}_{\text{comp}} + \lambda \mathcal{L}_{\text{gen}}, \tag{2}$$

which makes the decoder co-adapt with the drifting encoder. The remaining conditions isolate the role of pretraining: a comprehension-only baseline, a generation-only reference, and a from-scratch joint model that optimizes Eq. (2) with no Phase 1. Every condition except the two-model reference has the same encoder and parameter count. We report the exact optimizer-update budget because the schedules differ across families. A two-phase condition takes S Phase-1 updates, one per 64-graph pool mini-batch, and then S Phase-2 outer iterations, each with one update per labeled mini-batch, where B is the number of labeled mini-batches. Single-phase conditions take S outer iterations that sum the comprehension loss over the B mini-batches, add $\lambda\mathcal{L}_{\text{gen}}$ on one pool mini-batch, and take a single update. The two families therefore differ in optimizer updates by a factor of $1 + B$, which is two on Synthetic and PubMed, three on CiteSeer and QM9, and four on Cora (Table A.14). The primary claims use within-family contrasts under one schedule, and Appendix D.6 reports update-matched single-phase controls that equalize the count and the per-update loss weight.

3.3 EIGHT CONTROLLED CONDITIONS

The central constraint is a single learned parameterization for both tasks. Comprehension and generation are evaluated from that parameterization, with the explicit two-model reference as the only

exception. Within this constraint, graph generation introduces two separable components, a latent prior and a decoder. In a smaller grid, decoder treatment, prior specification, bottleneck sharing, and Phase-1 pretraining would remain confounded. The controlled grid therefore uses eight training conditions, namely MAE-DISCARD, SEQ-FREEZE, SEQ-ANCHOR, SEQ-REG, C-ONLY, G-ONLY, NAIVE-DECOUPLED, and JOINT-RETAINED. Appendix B gives the design matrix.

Decoder treatment. Four conditions share Phase 1 and differ only in Phase-2 decoder treatment.

- **MAE-DISCARD** discards the decoder and fine-tunes the encoder plus a fresh comprehension head on $\mathcal{L}_{\text{comp}}$ only. This adapts the discard-the-decoder convention of GraphMAE (Hou et al., 2022) and GPT-GNN (Hu et al., 2020b) to our backbone.
- **SEQ-FREEZE** fixes the Phase-1 decoder and fine-tunes the encoder with a fresh comprehension head on $\mathcal{L}_{\text{comp}}$. Because a frozen decoder contributes no gradient, its Phase-2 training path is identical to MAE-DISCARD's, so the two conditions coincide on comprehension metrics under paired seeds and differ only in whether the decoder is deployed.
- **SEQ-ANCHOR** is the encoder-anchoring control. It uses the same fixed Phase-1 decoder as SEQ-FREEZE and optimizes $\mathcal{L}_{\text{comp}} + \lambda_{\text{anc}}\mathcal{L}_{\text{anc}}$, where $\mathcal{L}_{\text{anc}} = \|\boldsymbol{\theta}_{\text{enc}} - \boldsymbol{\theta}_{\text{enc}}^{(1)}\|^2 + \|\boldsymbol{\theta}_{\mathbf{z}} - \boldsymbol{\theta}_{\mathbf{z}}^{(1)}\|^2$ (Li et al., 2018). The penalty anchors the encoder and the projection into the shared latent $\mathbf{z}$ to their Phase-1 weights. No generative loss is evaluated during Phase 2, which isolates weight anchoring from the active generative objective.
- **SEQ-REG** retains the decoder as trainable. The encoder, comprehension head, and decoder are fine-tuned jointly with $\mathcal{L}_{\text{comp}} + \lambda\mathcal{L}_{\text{gen}}$ kept active.

Reference and diagnostic conditions. C-ONLY and G-ONLY set the single-task endpoints, while NAIVE-DECOUPLED and JOINT-RETAINED remove one modeling assumption each.

- **C-ONLY** trains the encoder and comprehension head on labels only, with no generator.
- **G-ONLY** trains the encoder and decoder with $\mathcal{L}_{\text{gen}}$ and is the generation reference, with no comprehension head.
- **NAIVE-DECOUPLED** keeps the encoder shared but assigns separate comprehension and generation readouts, removing the shared graph-level bottleneck used by SEQ-REG.
- **JOINT-RETAINED** is the shared-$\mathbf{z}$ pathway trained jointly from scratch with no Phase 1, optimizing $\mathcal{L}_{\text{comp}} + \lambda\mathcal{L}_{\text{gen}}$. This tests whether Phase-1 pretraining matters.

The two-model reference combines C-ONLY and G-ONLY as two separate models with twice the parameters, so the tables report it through the C-ONLY and G-ONLY columns. Single-model methods are evaluated by their Pareto balance under one parameter vector.

Prior deployment. When the deployed decoder remains tied to Phase-1 latents, the refit and cached deployments defined above differ only in the prior paired with that decoder. Where the distinction matters we write SEQ-FREEZE ^{c} and SEQ-ANCHOR ^{c} for the cached-prior variants of the frozen conditions. The decoder-prior interface is absent for MAE-DISCARD, which retains no decoder, and SEQ-REG uses the refit deployment because its encoder and decoder are co-adapted during Phase 2. This deployment factor changes no training objective, model weights, or budget. It changes only which prior is used with the trained decoder at generation.

What the design isolates. The contrast among MAE-DISCARD, SEQ-FREEZE, and SEQ-REG varies only the Phase-2 treatment of the decoder, and the refit-versus-cached contrast varies only the deployed prior of a trained model. SEQ-ANCHOR is the main control: it anchors the encoder with L2-SP but computes no generative loss. NAIVE-DECOUPLED is a decoupled-readout diagnostic rather than an isolated ablation, because it changes the bottleneck, the decoder inputs, and the node sampler together, and JOINT-RETAINED removes Phase 1. Appendix D.9 adds controls outside the grid, including a frozen decoder with the generative loss kept active and a linear probe on the frozen Phase-1 model, to separate the generative gradient from decoder updates.

3.4 EVALUATION

Protocol summary. We adopt fixed scoring rules. Comprehension uses accuracy. Generation is reported with descriptor MMD and with a null-calibrated classifier two-sample test, C2ST (Lopez-Paz & Oquab, 2017). Descriptor MMD is the standard structural score (You et al., 2018; Liao et al.,

Table 1: Generation C2ST for conditions with a deployed generator and the density-matched Erdős-Rényi null, with G-ONLY ^{m} the update-matched G-ONLY and SEQ-ANCHOR $_1$ the anchor at $\lambda_{\text{anc}}=1$. C2ST is balanced discriminator accuracy, lower is better, and 0.5 is chance.

Dataset	G-ONLY ^{m}	SEQ-REG	JOINT-RETAINED	SEQ-ANCHOR	SEQ-ANCHOR $_1$	SEQ-FREEZE	NAIVE-DECOUPLED	ER null
Synthetic SBM	0.560 ± 0.040	0.538 ± 0.041	0.569 ± 0.046	0.802 ± 0.143	0.569 ± 0.045	0.837 ± 0.123	0.981 ± 0.058	0.772 ± 0.038
CiteSeer	0.719 ± 0.038	0.716 ± 0.038	0.742 ± 0.038	0.891 ± 0.047	0.756 ± 0.029	0.905 ± 0.046	0.942 ± 0.019	0.908 ± 0.029
Cora	0.753 ± 0.031	0.816 ± 0.026	0.849 ± 0.031	0.909 ± 0.025	0.818 ± 0.040	0.899 ± 0.041	0.906 ± 0.022	0.938 ± 0.027
PubMed	0.811 ± 0.038	0.806 ± 0.025	0.828 ± 0.058	0.876 ± 0.041	0.817 ± 0.048	0.901 ± 0.059	0.884 ± 0.044	0.937 ± 0.014
QM9	0.972 ± 0.017	0.963 ± 0.023	0.967 ± 0.035	0.999 ± 0.003	0.975 ± 0.035	0.999 ± 0.003	0.999 ± 0.003	0.975 ± 0.040

2019). C2ST trains a held-out discriminator to separate generated graphs from real graphs and is calibrated against a density-matched Erdős-Rényi null. Appendix C gives the metric definitions, auxiliary generation diagnostics, discriminator protocol, null construction, prior-deployment procedure, and round-trip procedure.

Comparison rule. Comparisons use paired splits and seeds, and significance is assessed with paired-bootstrap confidence intervals. Tables report mean $\pm$ standard deviation over the 10 paired seeds. We summarize the trade-off by single-model Pareto fronts over accuracy $\uparrow$ and either gen-MMD $\downarrow$ or C2ST $\downarrow$, and Appendix D.5 gives the dominance rule and the fronts per metric, deployment, and comparison set. Conditions without a comprehension head are not front candidates, and conditions without a deployable generator are assigned a generation score of $+\infty$.

Mechanism diagnostics. We use linear CKA (Kornblith et al., 2019) to measure encoder drift from the Phase-1 representation and round-trip reconstruction to measure how compatible the current encoder and the deployed decoder remain. These diagnostics separate representation preservation from encoder-decoder compatibility, which is essential for interpreting cached-prior generation. Appendix C specifies both procedures.

4 EXPERIMENTS

The experiments are organized as five research questions that match the five findings in the introduction. All experiments use the protocol of Sec. 3.4. Appendix B specifies the datasets, node budgets, model instantiation, compute, and sampling protocol. [1]

4.1 RQ1. DOES NULL-CALIBRATED C2ST CHANGE THE GENERATION ASSESSMENT?

Descriptor MMD, the common graph-generation score (You et al., 2018; O'Bray et al., 2022), measures only predefined statistics, and RQ1 compares its assessment with that of a null-calibrated learned metric. The ER-null column in Table 1 is an Erdős-Rényi sampler with empirical sizes and one pooled edge density, not a model condition. Under descriptor MMD it reaches 0.245 on QM9, second only to G-ONLY at 0.236 to 0.242 across budgets (Appendix D.2).

The learned metric gives a stricter assessment. Table 1 shows that the active-loss conditions and G-ONLY come close to real graphs only on Synthetic, with C2ST 0.54 to 0.57. For citation, they rise to 0.72 to 0.85, and on QM9 every reported condition lands at 0.96 to 1.00. The same null is separable in every domain, with C2ST 0.77 to 0.98. Table A.18 applies the paired rule of Appendix C. The active-loss conditions, G-ONLY, and SEQ-ANCHOR $_1$ are significantly less separable than the null on Synthetic and the three citation domains, by 0.09 to 0.23, whereas the other frozen and decoupled conditions clear it only on Cora and PubMed. On QM9 only JOINT-RETAINED significantly clears the null, by less than 0.01. Every condition outside Synthetic lies at least 0.21 above the real-versus-real floor of 0.48 to 0.52. Clearing the null shows structure beyond graph size and a pooled edge density, but neither structure beyond per-graph density nor distributional equality.

4.2 RQ2. MUST THE PRIOR AND DECODER BE SEPARATED?

Retained-generation degradation can arise from the decoder itself or from the prior used with it at deployment. Under the refit deployment (Table 2), SEQ-REG is not worse than G-ONLY in C2ST at either budget except on Cora, where the matched G-ONLY is lower by 0.06, and the matched G-ONLY has significantly lower MMD on CiteSeer, Cora, and QM9, by 0.003 to 0.033 (Appendix D.6).

[1] The code will be public after acceptance.

Table 2: Generation MMD, averaged over degree, clustering, and spectral MMD, for conditions with a deployable generator. Lower is better, G-ONLY ^{m} is the update-matched G-ONLY, and SEQ-ANCHOR $_1$ is SEQ-ANCHOR at λ_{anc}=1. The frozen-decoder conditions share one cached-prior deployment, the unchanged Phase-1 generator.

		refit prior						cached prior
Dataset	G-ONLY ^{m}	SEQ-REG	JOINT-RETAINED	NAIVE-DECOUPLED	SEQ-FREEZE	SEQ-ANCHOR	SEQ-ANCHOR $_1$	frozen
Synthetic SBM	0.004 ± 0.002	0.006 ± 0.003	0.012 ± 0.009	1.193 ± 0.044	0.416 ± 0.323	0.301 ± 0.260	0.010 ± 0.006	0.009 ± 0.005
CiteSeer	0.009 ± 0.003	0.012 ± 0.005	0.020 ± 0.010	0.294 ± 0.101	0.406 ± 0.150	0.353 ± 0.124	0.026 ± 0.017	0.018 ± 0.010
Cora	0.020 ± 0.003	0.036 ± 0.006	0.056 ± 0.014	0.422 ± 0.104	0.425 ± 0.186	0.418 ± 0.148	0.043 ± 0.011	0.043 ± 0.008
PubMed	0.068 ± 0.015	0.075 ± 0.014	0.104 ± 0.018	0.301 ± 0.119	0.426 ± 0.098	0.364 ± 0.086	0.095 ± 0.023	0.109 ± 0.030
QM9	0.236 ± 0.015	0.269 ± 0.021	0.286 ± 0.031	0.583 ± 0.037	0.572 ± 0.189	0.519 ± 0.187	0.287 ± 0.035	0.290 ± 0.029

Table 3: Test accuracy for the seven comprehension-capable conditions and SEQ-ANCHOR $_1$, which is SEQ-ANCHOR at λ_{anc}=1 (Appendix D.4). G-ONLY has no comprehension head.

Dataset	C-ONLY	MAE-DISCARD	SEQ-FREEZE	SEQ-ANCHOR	SEQ-REG	NAIVE-DECOUPLED	JOINT-RETAINED	SEQ-ANCHOR $_1$
Synthetic SBM	0.794 ± 0.033	0.843 ± 0.034	0.843 ± 0.034	0.915 ± 0.020	0.924 ± 0.028	0.772 ± 0.058	0.919 ± 0.016	0.935 ± 0.008
CiteSeer	0.524 ± 0.038	0.571 ± 0.033	0.571 ± 0.033	0.574 ± 0.034	0.616 ± 0.024	0.575 ± 0.025	0.602 ± 0.030	0.592 ± 0.029
Cora	0.659 ± 0.039	0.682 ± 0.027	0.682 ± 0.027	0.683 ± 0.029	0.683 ± 0.033	0.682 ± 0.029	0.656 ± 0.030	0.709 ± 0.030
PubMed	0.646 ± 0.028	0.690 ± 0.030	0.690 ± 0.030	0.692 ± 0.025	0.671 ± 0.047	0.684 ± 0.035	0.689 ± 0.030	0.675 ± 0.038
QM9	0.825 ± 0.043	0.835 ± 0.036	0.835 ± 0.036	0.828 ± 0.040	0.840 ± 0.034	0.852 ± 0.038	0.830 ± 0.037	0.771 ± 0.048

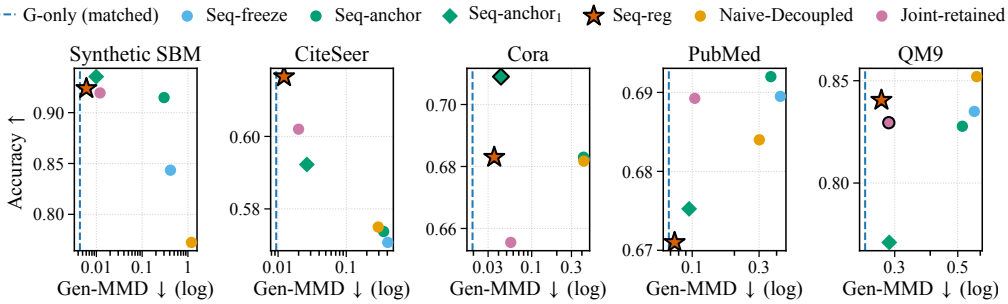


Figure 2: Per-dataset accuracy (higher is better) against generation MMD (log scale, lower is better) under the refit deployment. ★ marks SEQ-REG, a black outline marks conditions not significantly dominated among the grid and SEQ-ANCHOR $_1$ (Appendix D.5), and the dashed line marks the update-matched G-ONLY. MAE-DISCARD and C-ONLY have no generator.

SEQ-FREEZE and SEQ-ANCHOR have far higher MMD, 0.30 to 0.57, so under refit alone the generative loss would appear necessary.

The cached prior changes the causal attribution. For frozen conditions, the degradation comes from refitting the prior on the drifted Phase-2 encoder while the decoder still expects Phase-1 latents. The cached Phase-1 prior removes most of it on every domain, bringing SEQ-FREEZE and SEQ-ANCHOR within 0.005 to 0.054 of the matched G-ONLY in MMD and lowering their C2ST significantly on all five domains (Table A.20). Limiting drift works under refit too: across the anchor sweep the refit penalty shrinks as the anchor tightens, to at most 0.02 in C2ST at $\lambda_{\text{anc}} \geq 1$ (Appendix D.4), and SEQ-ANCHOR $_1$ comes within 0.006 to 0.051 of the matched G-ONLY in MMD with no active loss. For SEQ-REG, whose decoder co-adapts, the refit prior is better (Appendix D.8). A retained decoder generates well when paired with the prior it was calibrated to.

Fig. 2 summarizes the trade-off under the refit deployment, and front membership depends on the metric and the comparison set (Table A.12). SEQ-REG is on every front of the eight-condition grid, alone under gen-MMD except on QM9. With SEQ-ANCHOR $_1$ added, SEQ-REG keeps every front except C2ST on Cora, where the anchor is more accurate at no C2ST cost. Against update-matched JOINT-RETAINED controls (Appendix D.6), SEQ-REG keeps a significant C2ST advantage on Synthetic, CiteSeer, and PubMed but no accuracy or gen-MMD advantage, and on CiteSeer the more accurate control displaces it from the gen-MMD front.

4.3 RQ3. CAN ANCHORING REPLACE DECODER RETENTION FOR COMPREHENSION?

RQ3 asks whether the comprehension gain of decoder retention can be reproduced without a generative loss. In Table 3, against MAE-DISCARD, SEQ-REG gains significantly only on Synthetic and CiteSeer, by 0.080 and 0.046 in paired accuracy, and is within noise on Cora, PubMed, and QM9. The SEQ-ANCHOR control, which computes no generative loss, closes all but 0.009 of the Synthetic gain but only 0.003 of the CiteSeer gain.

Table 4: Linear CKA between Phase 1 and 2 pooled encoder representations as encoder-drift probe, where higher means less drift from the pretrained representation. JOINT-RETAINED has no Phase 1, so its reference is the initialization. C-ONLY and G-ONLY are single-task endpoints and are omitted.

Dataset	MAE-DISCARD	SEQ-FREEZE	SEQ-ANCHOR	SEQ-REG	NAIVE-DECOUPLED	JOINT-RETAINED
Synthetic SBM	0.515 ± 0.053	0.515 ± 0.053	0.865 ± 0.066	0.863 ± 0.031	0.836 ± 0.040	0.599 ± 0.077
CiteSeer	0.313 ± 0.077	0.313 ± 0.077	0.338 ± 0.074	0.518 ± 0.054	0.518 ± 0.188	0.362 ± 0.102
Cora	0.394 ± 0.088	0.394 ± 0.088	0.425 ± 0.071	0.492 ± 0.075	0.388 ± 0.067	0.261 ± 0.064
PubMed	0.261 ± 0.061	0.261 ± 0.061	0.347 ± 0.077	0.650 ± 0.081	0.550 ± 0.073	0.347 ± 0.095
QM9	0.308 ± 0.097	0.308 ± 0.097	0.360 ± 0.123	0.708 ± 0.087	0.630 ± 0.096	0.434 ± 0.264

Table 4 gives the mechanism probe. MAE-DISCARD and SEQ-FREEZE drift most from the pretrained representation. The L2-SP penalty raises CKA as intended, and SEQ-REG reaches comparable or higher CKA without one. SEQ-ANCHOR $_1$ in Table 3 tests a stronger anchor (Appendix D.4). It matches SEQ-REG on Synthetic and PubMed and beats both SEQ-REG and MAE-DISCARD on Cora, by 0.026 and 0.027, where retention gains nothing. On CiteSeer no weight closes the gap, and the gain appears whenever the generative loss trains the decoder (Appendices D.6 and D.8) but not under a contrastive pool loss (Appendix D.9). On QM9 SEQ-ANCHOR $_1$ loses 0.064 to MAE-DISCARD. Weight anchoring therefore reproduces the retention gain on Synthetic but not on CiteSeer, where only a generative loss that trains the decoder gives it.

4.4 RQ4. DOES THE GENERATIVE LOSS KEEP ENCODER AND DECODER COMPATIBLE?

Cached-prior deployment recovers unconditional generation for a frozen decoder, but it pairs a Phase-2 encoder with a Phase-1 decoder. Any task that must encode a real graph and then decode it needs the latents of the current encoder to decode well. RQ4 asks whether the active generative loss keeps them decodable by the deployed decoder.

Table 5: Round-trip reconstruction MMD from held-out graphs via deployed encoder and decoder. The table reports active-loss and frozen-decoder conditions, which define the encoder-decoder compatibility contrast. Lower is better.

	active loss		frozen decoder	
Dataset	SEQ-REG	JOINT-RETAINED	SEQ-FREEZE	SEQ-ANCHOR
Synthetic SBM	0.003 ± 0.004	0.005 ± 0.004	0.386 ± 0.271	0.288 ± 0.237
Cora	0.034 ± 0.006	0.054 ± 0.014	0.337 ± 0.168	0.325 ± 0.153
CiteSeer	0.017 ± 0.011	0.014 ± 0.007	0.325 ± 0.172	0.272 ± 0.132
PubMed	0.074 ± 0.036	0.099 ± 0.015	0.396 ± 0.107	0.338 ± 0.077
QM9	0.273 ± 0.020	0.283 ± 0.025	0.561 ± 0.200	0.506 ± 0.187

Table 5 tests this requirement at the distribution level, comparing the decoded graphs with the inputs in descriptor MMD. SEQ-REG reaches MMD 0.003 to 0.273 and JOINT-RETAINED 0.005 to 0.283, against 0.325 to 0.561 for SEQ-FREEZE and 0.272 to 0.506 for SEQ-ANCHOR. The learned metric narrows this result (Table A.24). The decoded set of SEQ-REG is at least as separable from its inputs as its samples are from real graphs, and on QM9 no condition's round-trip clears the null. Outside QM9, the round-trip C2ST of SEQ-REG is still significantly below that of both frozen conditions, as is that of JOINT-RETAINED on all but Cora. Among the tested conditions, the active generative loss reduces the encoder-decoder mismatch relative to frozen decoders while the encoder adapts, on every domain under descriptor MMD and outside QM9 under C2ST, which we call compatibility rather than distributional equality. Pinning the encoder by a linear probe (Appendix D.9) or the largest anchor weight (Appendix D.4) also keeps round-trip MMD within 0.02 of SEQ-REG except on PubMed, but costs accuracy on CiteSeer and QM9. Two checks in Appendix D.9 sharpen this statement. A frozen decoder with the loss kept active reaches round-trip MMD 0.008 to 0.279, below both frozen conditions on every domain, so the compatibility comes from the generative gradient reaching the encoder rather than from decoder updates. Decoding each test graph's own latent against a size-matched foreign latent recovers node attributes per graph under the active loss on every domain but edge structure only weakly after node matching, so the structural round-trip is a distribution-level statement.

4.5 RQ5. DOES THE GENERATOR FAMILY OUTWEIGH PRIOR DEPLOYMENT?

The trade-off could instead reflect the choice of generator family, so RQ5 compares three variants: a deterministic autoencoder with a fitted Gaussian prior, the same decoder with a latent-diffusion prior, and a dual-use VAE. The Gaussian and latent-diffusion priors share one checkpoint and differ only in C2ST, where the diffusion prior is significantly worse on PubMed (Table A.22). The VAE of Table 6, trained on the SEQ-REG schedule, has significantly higher C2ST than SEQ-REG on

Table 6: Accuracy and generation C2ST for three generator families with the same encoder architecture. Higher accuracy and lower C2ST are better. Diffusion reuses the Gaussian column's SEQ-REG checkpoint, and the VAE is trained on the SEQ-REG schedule and sampled from its fixed prior.

	Gaussian prior		Latent-diffusion prior		VAE	
Dataset	acc	C2ST	acc	C2ST	acc	C2ST
Synthetic SBM	0.924 ± 0.028	0.538 ± 0.041	0.924 ± 0.028	0.533 ± 0.038	0.894 ± 0.068	0.639 ± 0.059
CiteSeer	0.616 ± 0.024	0.716 ± 0.038	0.616 ± 0.024	0.727 ± 0.022	0.612 ± 0.030	0.681 ± 0.056
Cora	0.683 ± 0.033	0.816 ± 0.026	0.683 ± 0.033	0.826 ± 0.032	0.689 ± 0.033	0.825 ± 0.041
PubMed	0.671 ± 0.047	0.806 ± 0.025	0.671 ± 0.047	0.836 ± 0.032	0.665 ± 0.030	0.820 ± 0.049
QM9	0.840 ± 0.034	0.963 ± 0.023	0.840 ± 0.034	0.934 ± 0.103	0.839 ± 0.030	0.988 ± 0.005

Synthetic and QM9 and lower C2ST on CiteSeer. Both family effects are significantly smaller than the prior-deployment effect on Synthetic, CiteSeer, and Cora.

The tested family changes thus shift the trade-off less than prior deployment on three of five domains. A frozen VAE decoder shows a similar refit penalty, but that refit ignores posterior variance (Appendix D.8). Appendix D groups the supporting experiments: an external generator, decoupled readouts, backbone choice, hyperparameter sensitivity, and ogbn-arxiv scale.

5 DISCUSSION AND CONCLUSION

The generator is a prior plus a decoder. A graph autoencoder generator is a decoder plus a prior over latents. VGAE (Kipf & Welling, 2016) fixes a standard normal prior through its KL term, whereas the deterministic autoencoders studied here fit the prior to encoded latents after training (Ghosh et al., 2020). In both cases fine-tuning moves the encoder that supplies the decoder's latents while a frozen decoder stays calibrated to Phase-1 latents, which is why Sec. 4.2 attributes the degradation to the deployed prior. The effect belongs to a broader class of readout-dependent phenomena, including final-readout bias in class-incremental forgetting (Wu et al., 2019) and readout-dependent fine-tuning distortion (Kumar et al., 2022). Any report that frozen retention degrades graph generation should therefore state the deployed prior.

Comprehension gain and generation capability are separable. The anchoring control in Sec. 4.3, which computes no generative loss, obtains the Synthetic comprehension gain without generative ability and a Cora gain that retention does not give, consistent with diffusion representations whose benefit need not track generation quality (Chen et al., 2025) and with quadratic penalties replacing rehearsal (Chen et al., 2020), but not that the generative loss acts through the same mechanism. Only on CiteSeer does the gain resist every tested L2-SP weight, and on both domains it resists a contrastive loss on the same unlabeled pool (Appendix D.9), which leaves other regularizers untested.

A cached prior covers sampling only. A cached prior suffices only when no input graph must be encoded before decoding. Tasks that encode an input and decode it need a decoder compatible with the deployed encoder, which among the tested conditions only the active generative loss improves over frozen decoders without pinning the encoder (Sec. 4.4). CKA misses the distinction: on Synthetic the anchor matches the CKA of SEQ-REG yet round-trips far worse. Input-conditioned structural reconstruction and latent editing remain untested.

We asked what governs a compact graph autoencoder's ability to both classify and generate. The controlled evaluation traces the apparent trade-off to three separable sources, latent-prior mismatch, the auxiliary generative loss, and the generation metric. Prior mismatch grows with encoder drift and shifts the trade-off more than the tested generator families on three of five domains, so report the deployed prior with any frozen-decoder generation result. In practice, keep the generative loss active with a trainable decoder when an L2-SP anchor does not reproduce its comprehension gain, as on CiteSeer, and whenever the model must decode encoded inputs. For a deterministic autoencoder with a prior fitted to its latents, pair a decoder frozen at Phase 1 with the cached Phase-1 prior and a decoder trained in Phase 2 with a refit prior. Check compatibility by round-trip reconstruction, not linear CKA, and score generators with a learned discriminator against a density-matched null. The open problem is a dual-use graph model that reaches the real-versus-real floor on structurally rich and molecular graphs, as a dedicated generator does on QM9, while sharing one latent space with a strong classifier.

AI USE STATEMENT

Generative AI tools were used to revise text.

ETHICS STATEMENT

This work is a controlled evaluation of compact graph autoencoders on public benchmark data: synthetic stochastic block model graphs, the Cora, CiteSeer, and PubMed citation graphs, ogbn-arxiv, and the QM9 molecular set. It involves no human subjects, personal data, or newly collected data. The molecular generators studied are small models scored against standard benchmark distributions, and the findings concern training and deployment interfaces rather than the design of specific molecules. We do not foresee ethical concerns beyond those common to graph representation learning research.

REPRODUCIBILITY STATEMENT

All experiments use public datasets. Appendix B specifies dataset construction and splits, the encoder and decoder architecture, training schedules and hyperparameters, the seed and confidence-interval protocol, and the compute used. Appendix C specifies every evaluation metric, including the C2ST discriminator and null construction, the prior-deployment procedure, and the round-trip protocol.

REFERENCES

Runjin Chen, Tong Zhao, Ajay Jaiswal, Neil Shah, and Zhangyang Wang. Llaga: large language and graph assistant. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 7809–7823, 2024.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7870–7881, 2020.

Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. In *International Conference on Learning Representations*, volume 2025, pp. 55458–55472, 2025.

Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. An embarrassingly simple approach for transfer learning from pretrained language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2089–2095, 2019.

Bin Dai and David Wipf. Diagnosing and enhancing VAE models. *arXiv preprint arXiv:1903.05789*, 2019.

Zhangyang Gao, Daize Dong, Cheng Tan, Jun Xia, Bozhen Hu, and Stan Z Li. A graph is worth k words: Euclideanizing graph using pure transformer. In *International Conference on Machine Learning*, pp. 14681–14701. PMLR, 2024.

Partha Ghosh, Mehdi S. M. Sajjadi, Antonio Vergari, Michael Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. In *International Conference on Learning Representations*, 2020.

Xueting Han, Zhenhuan Huang, Bang An, and Jing Bai. Adaptive transfer learning on graph neural networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 565–574, 2021.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 594–604, 2022.

Zhenyu Hou, Yufei He, Yukuo Cen, Xiao Liu, Yuxiao Dong, Evgeny Kharlamov, and Jie Tang. Graphmae2: A decoding-enhanced masked self-supervised graph learner. In *Proceedings of the ACM web conference 2023*, pp. 737–746, 2023.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22118–22133, 2020a.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. Gpt-gnn: Generative pre-training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1857–1867, 2020b.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9307–9315, 2024.

Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. Score-based generative modeling of graphs via the system of stochastic differential equations. In *International conference on machine learning*, pp. 10362–10383. PMLR, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Diederik P Kingma, Danilo J Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Lecheng Kong, Jiarui Feng, Hao Liu, Chengsong Huang, Jiaxin Huang, Yixin Chen, and Muhan Zhang. Gofa: A generative one-for-all model for joint graph language modeling. In *International Conference on Learning Representations*, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Markus Krimmel, Philip Hartout, Karsten Borgwardt, and Dexiong Chen. Polygraph discrepancy: a classifier-based metric for graph generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=nGizKVBCgn.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Jintang Li, Ruofan Wu, Wangbin Sun, Liang Chen, Sheng Tian, Liang Zhu, Changhua Meng, Zibin Zheng, and Weiqiang Wang. What's behind the mask: Understanding masked graph modeling for graph autoencoders. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 1268–1279, 2023.

Teng Li, Quanfeng Lu, Lirui Zhao, Hao Li, Xizhou Zhu, Yu Qiao, Jun Zhang, and Wenqi Shao. Unifork: Exploring modality alignment for unified multimodal understanding and generation. *arXiv preprint arXiv:2506.17202*, 2025.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International conference on machine learning*, pp. 2825–2834. PMLR, 2018.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *14th European Conference on Computer Vision, ECCV 2016*, pp. 614–629. Springer, 2016.

Renjie Liao, Yujia Li, Yang Song, Shenlong Wang, Will Hamilton, David K Duvenaud, Raquel Urtasun, and Richard Zemel. Efficient graph generation with graph recurrent attention networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. One for all: Towards training one graph model for all classification tasks. In *International conference on learning representations*, volume 2024, pp. 20188–20210, 2024.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

Leslie O'Bray, Max Horn, Bastian Rieck, and Karsten Borgwardt. Evaluation metrics for graph generative models: Problems, pitfalls, and practical solutions. In *International Conference on Learning Representations*, 2022.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–106, 2008.

Hamed Shirzad, Kaveh Hassani, and Danica J Sutherland. Evaluating graph generative models with contrastively learned features. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7783–7795, 2022.

Martin Simonovsky and Nikos Komodakis. Graphvae: Towards generation of small graphs using variational autoencoders. In *International conference on artificial neural networks*, pp. 412–422. Springer, 2018.

Yifei Sun, Qi Zhu, Yang Yang, Chunping Wang, Tianyu Fan, Jiajun Zhu, and Lei Chen. Fine-tuning graph neural networks by preserving graph generative patterns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 9053–9061, 2024.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 491–500, 2024.

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

Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Mehdi Azabou, Eva L Dyer, Remi Munos, Petar Veličković, and Michal Valko. Large-scale representation learning on graphs via bootstrapping. In *International Conference on Learning Representations*, 2022.

Rylee Thompson, Boris Knyazev, Elahe Ghalebi, Jungtaek Kim, and Graham W Taylor. On evaluation metrics for graph generative models. In *International Conference on Learning Representations*, 2022.

Jakub M Tomczak and Max Welling. VAE with a VampPrior. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations*, 2019.

Clément Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

Jianing Wang, Junda Wu, Yupeng Hou, Yao Liu, Ming Gao, and Julian McAuley. Instructgraph: Boosting large language models via graph-centric instruction tuning and preference alignment. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13492–13510, 2024a.

Letian Wang, Chuhan Zhang, Rishabh Kabra, Jasper Uijlings, Steven Waslander, Andrew Zisserman, Joao Carreira, Kaiming He, Misha Andriluka, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Video generation models are general-purpose vision learners. In *European Conference on Computer Vision (ECCV)*, 2026.

Zehong Wang, Zheyuan Zhang, Nitesh V Chawla, Chuxu Zhang, and Yanfang Ye. Gft: Graph foundation model with transferable tree vocabulary. In *Advances in Neural Information Processing Systems*, volume 37, pp. 107403–107443, 2024b.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12966–12977, 2025.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 374–382, 2019.

Lianghao Xia and Chao Huang. Anygraph: Graph foundation model in the wild. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 882–896, 2026.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pp. 40–48. PMLR, 2016.

Jiaxuan You, Rex Ying, Xiang Ren, William Hamilton, and Jure Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models. In *International conference on machine learning*, pp. 5708–5717. PMLR, 2018.

Jianan Zhao, Zhaocheng Zhu, Mikhail Galkin, Hesham Mostafa, Michael Bronstein, and Jian Tang. Fully-inductive node classification on arbitrary graphs. In *International Conference on Learning Representations*, volume 2025, pp. 23968–23986, 2025.

Lingxiao Zhao, Xueying Ding, and Leman Akoglu. Pard: Permutation-invariant autoregressive diffusion for graph generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 7156–7184, 2024.

Cai Zhou, Xiyuan Wang, and Muhan Zhang. Unifying generation and prediction on graphs with latent graph diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 61963–61999, 2024.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, volume 2025, pp. 6446–6469, 2025.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*, 2020.

APPENDIX OVERVIEW

- **A. Related Work** 15
- **B. Implementation Details** 17
- **C. Evaluation Protocol Details** 18
- **D. Additional Experiment Analyses** 21
  - *D.1. GAT Backbone Ablation* 21
  - *D.2. External Graph-Generation Baselines* 21
  - *D.3. Single-GPU DiGress Reference* 22
  - *D.4. Hyperparameter Sensitivity: λ and Anchor-Weight Sweeps* 23
  - *D.5. Uncertainty-Aware Pareto-Front Computation* 25
  - *D.6. Update-Matched Single-Phase Controls* 27
  - *D.7. ogbn-arxiv Compact Few-Shot Regime* 31
  - *D.8. Auxiliary Robustness Checks* 32
  - *D.9. Per-Graph Reconstruction and Additional Controls* 36
- **E. Limitations** 38

A Related Work

Graph pretraining. Graph SSL methods differ in whether the pretext task includes a decoder. Contrastive and bootstrapped encoders such as DGI (Veličković et al., 2019), GRACE (Zhu et al., 2020), and BGRL (Thakoor et al., 2022) train encoder-only representations. Generative-pretext objectives add a decoder through masked reconstruction in GraphMAE (Hou et al., 2022), GraphMAE2 (Hou et al., 2023), and MaskGAE (Li et al., 2023), attribute-edge prediction in GPT-GNN (Hu et al., 2020b), and VGAE (Kipf & Welling, 2016) as the structural ancestor. In these methods, the decoder is a pretext component and is discarded before fine-tuning. The deployed interface is therefore an encoder for comprehension. We treat this discard convention as the experimental variable and ask whether keeping the decoder and its loss active during fine-tuning provides measurable benefit.

Graph foundation models. A recent line pushes toward universal graph encoders. OFA (Liu et al., 2024) unifies tasks across feature schemas, GFT (Wang et al., 2024b) introduces transferable vocabularies, GraphAny (Zhao et al., 2025) and AnyGraph (Xia & Huang, 2026) target zero-shot transfer, and graph-LLM hybrids such as GraphGPT (Tang et al., 2024), LLaGA (Chen et al., 2024), and GOFA (Kong et al., 2025) fuse graph encoders with language models. These encoder-side models primarily expose comprehension interfaces, while our setting requires a structured graph generator in the same deployed model and evaluates the resulting Pareto trade-off explicitly. Closer to our setting, GraphsGPT (Gao et al., 2024) pretrains a Graph2Seq encoder with a GraphGPT decoder on 100M molecules and shows that the pretrained model transfers to both graph classification and few-shot graph generation, and Latent Graph Diffusion (Zhou et al., 2024) trains one latent diffusion framework that addresses generation and prediction tasks. InstructGraph (Wang et al., 2024a) likewise unifies graph reasoning and generation in an instruction-tuned LLM. These works validate each capability separately at scale. Our controlled grid targets the marginal effect of decoder retention itself, with paired seeds, update-matched controls, and an explicit regularization control on the same backbone.

Graph generators. Graph generators such as GraphRNN (You et al., 2018), GraphVAE (Simonovsky & Komodakis, 2018), GDSS (Jo et al., 2022), DiGress (Vignac et al., 2023), and PARD (Zhao et al., 2024) are developed as standalone systems for high-quality generation. We use them as generation-side references while studying a complementary setting: a shared encoder, a single-model deployment constraint, and a comprehension-generation Pareto comparison.

Priors for deterministic autoencoders. Fitting a Gaussian to the encoded latents of a trained deterministic autoencoder is the ex-post density estimation of regularized autoencoders (Ghosh et al., 2020), and the refit deployment applies it after comprehension fine-tuning. The mismatch studied here is not the gap between a fixed VAE prior and the aggregate posterior, which learned priors such as the VampPrior (Tomczak & Welling, 2018) and the two-stage VAE (Dai & Wipf, 2019) address. The refit prior describes the current latents well, but the frozen decoder was trained on Phase-1 latents that the current encoder no longer produces.

Unified models in vision and language. The vision and language community has converged on unified models that both comprehend and generate, and its explanations of the resulting trade-off fall into two families. The first is representation granularity. Janus (Wu et al., 2025) decouples visual-encoder pathways while sharing a downstream transformer, on the grounds that understanding and generation require features at different granularities, and UniFork (Li et al., 2025) refines this by measuring that the two tasks require modality alignment to move in opposite directions with depth, motivating shared shallow layers and forked deep ones. The second is the modeling objective. Show-o (Xie et al., 2025) unifies autoregressive understanding with discrete diffusion generation, and Chameleon (Team, 2024) and Transfusion (Zhou et al., 2025) push end-to-end multimodal pretraining at scale with modality-specific objectives. A stronger position treats generation as the representation source itself. GenCeption (Wang et al., 2026) proposes large-scale video generation as the pretraining paradigm for general-purpose vision. Our compact graph study does not compare modalities. Within it, the controls separate prior deployment, the auxiliary generative loss, and the generation metric. Motivated by the pathway decoupling of the first family, we test a graph variant, Naive-Decoupled, that gives the two readout heads separate paths on a shared encoder. It is comprehension-competitive under our protocol. Its generation deficit is consistent with i.i.d. node sampling, but because it also changes the bottleneck and the decoder inputs, it does not isolate the decoupled-readout principle.

Table A.1: Controlled conditions and deployment status. Blank cells indicate inapplicable fields. The two-model reference uses separate C-ONLY and G-ONLY checkpoints as a reference relative to the single-checkpoint constraint. MAE-DISCARD and SEQ-FREEZE share one Phase-2 training path and differ only in deployment.

Condition	Phase-1	Ph-2 enc	Ph-2 dec	Active $\mathcal{L}_{\text{gen}}$	Single-ckpt	Generator
C-ONLY		train		no	yes	no
G-ONLY		train	train	yes	yes	reference
MAE-DISCARD	gen pretrain	train	discard	no	yes	no
SEQ-FREEZE	gen pretrain	train	freeze	no	yes	refit-limited
SEQ-ANCHOR	gen pretrain	L2-SP	freeze	no	yes	refit-limited
SEQ-REG	gen pretrain	train	train	yes	yes	yes
NAIVE-DECOUPLED	gen pretrain	train	train	yes	yes	diagnostic
JOINT-RETAINED		train	train	yes	yes	yes
Additional controls (Appendix D.9)						
SEQ-FREEZE-ACTIVE	gen pretrain	train	freeze	yes	yes	refit / cached
LINEAR-PROBE	gen pretrain	freeze	freeze	no	yes	Phase-1 (cached = refit)
SEQ-SSL	gen pretrain	train	freeze	no	yes	Phase-1 (cached)
Two-model						
C-ONLY + G-ONLY					no	reference

Evaluating graph generators. Graph generation is conventionally scored by MMD between hand-crafted structural descriptors. This practice has well-documented sensitivity. O'Bray et al. (2022) show that descriptor, kernel, and hyperparameter choices are made without reference to the distribution being estimated and that the resulting scores are unstable, while Thompson et al. (2022) propose features from an untrained random GNN and Shirzad et al. (2022) contrastively learned features as more expressive alternatives. Classifier-based scoring has also been proposed directly, with Poly-Graph Discrepancy (Krimmel et al., 2026) fitting binary classifiers to approximate a Jensen-Shannon distance and make scores comparable across descriptors. The same movement from fixed statistics to learned discriminators has played out in vision, where the Fréchet Inception Distance was found to misrank models and was replaced by a richer-embedding alternative (Jayasumana et al., 2024). We use a classifier two-sample test (Lopez-Paz & Oquab, 2017) in this spirit. Our contribution is the calibrated dual-use evaluation in which C2ST is scored against a density-matched random-graph null, so that the score states whether a generator improves on a density-only null.

Regularization and G-Tuning. L2-SP (Li et al., 2018) anchors fine-tuned weights to their pretrained initialization. EWC (Kirkpatrick et al., 2017) and LwF (Li & Hoiem, 2016) are continual-learning relatives. Keeping a pretraining loss active during fine-tuning as an anti-forgetting device also has direct precedent in language through SiATL (Chronopoulou et al., 2019) and in graphs through AUX-TS (Han et al., 2021). Our SEQ-REG condition belongs to this auxiliary-loss line, whose ancestor is semi-supervised learning with a generative objective on unlabeled data (Kingma et al., 2014). Two results outside graphs are methodological precedents for our attribution findings. RecAdam (Chen et al., 2020) shows that a quadratic penalty toward the pretrained weights recovers the anti-forgetting benefit of rehearsing the pretraining task itself, and deconstructing denoising diffusion models (Chen et al., 2025) finds that representation quality tracks a low-dimensional latent rather than the generative process, and is uncorrelated with generation quality. Related analyses show that apparent catastrophic forgetting in class-incremental learning can arise from a bias in the final readout (Wu et al., 2019), and that fine-tuning can distort pretrained features in ways a different readout protocol avoids (Kumar et al., 2022). Both are analogous to our prior-deployment result. Our controlled dual-use evaluation adds an explicit anchoring control with no generative loss on the same backbone and identifies the apparent generation advantage of frozen retention as largely a prior-deployment confound. We use SEQ-ANCHOR as an L2-SP control that tests whether anchoring the encoder toward its pretrained weights, with no generative loss, reproduces what the active generative loss provides. Most directly related, G-Tuning (Sun et al., 2024) preserves a graph's generative structure during fine-tuning by reconstructing the downstream graph's graphon. G-Tuning uses generative-pattern preservation as an auxiliary regularizer for classification and reports accuracy without deploying or evaluating a generator, whereas in our study the generator is a deployed output whose quality we measure directly through descriptor MMD and a null-calibrated C2ST.

B Implementation Details

Table A.1 defines the condition taxonomy used throughout the appendix. It distinguishes generation pretraining, Phase-2 encoder and decoder updates, the active generative loss, and whether the deployed artifact is a single checkpoint with a usable generator.

Datasets. The primary controlled grid covers five domains spanning synthetic, citation, and molecular graphs, with a common $N{=}24$ graph budget and 32-d node features. Graphs below the budget are padded, and graphs above the budget are subsampled or truncated before both training and evaluation. Every domain has validation and test sets of 400 graphs (500 on Synthetic) and an unlabeled pool of 1500 graphs (1200 on Synthetic), all redrawn for each seed.

Synthetic stochastic-block-model graphs have exactly 24 nodes. A graph has 2 or 3 communities with probability $1/2$ each, of sizes $12{+}12$ or $8{+}8{+}8$, and its label is the community count. Node pairs are connected independently, with a higher probability within a community than between communities. Each community draws a signal vector uniformly from $\{-1, +1\}^2$. Feature dimensions 0 and 1 of a node are its community's signal plus $\mathcal{N}(0, 0.3^2)$ noise, and dimensions 2 to 31 are $\mathcal{N}(0, 1)$ noise, without standardization. The pool is generated from the experiment seed and the validation and test sets from two separate random streams, and the labeled set is the first 16 pool graphs of each class.

Cora, CiteSeer, and PubMed (Sen et al., 2008; Yang et al., 2016) are converted into 2-hop ego-graphs with the center node first. An ego-graph with more than 24 nodes keeps its center and 23 other nodes drawn uniformly without replacement. A random permutation of all nodes, not the public Planetoid split, assigns the centers: the first $K{=}20$ nodes of each class form the labeled set, and the next 400, 400, and 1500 remaining nodes the validation, test, and pool centers. The raw features of all nodes are centered and projected by SVD onto 32 dimensions without labels, and each projected column is scaled to unit standard deviation over all nodes.

QM9 (Ramakrishnan et al., 2014) is evaluated as attributed-graph generation. Each seed draws 8000 of its molecules, and the label is whether the HOMO-LUMO gap exceeds its median over these 8000 molecules, computed before the split. A permutation then assigns 64 labeled molecules per class and 400, 400, and 1500 molecules to validation, test, and pool. Molecules keep their first 24 atoms in QM9 order with the bonds among them, and the 11 QM9 atom features fill the first 11 of the 32 feature dimensions. Features are standardized per column over all 24 slots and are zero on padding slots.

Within the primary grid, the held-out reference set receives the same padding, subsampling, and truncation as the training set, and generated graphs are masked to sizes drawn from the resulting size distribution. Primary-grid C2ST and MMD scores are therefore measured against the budgeted distribution used for training. Paired seeds provide identical inputs across conditions, making direct contrasts comparable. On the real domains the labeled, validation, test, and pool sets are disjoint by construction. On Synthetic the labeled set is drawn from the pool, and the validation and test sets come from separate random streams. Ego-graphs of different centers may share nodes of the underlying citation graph, and on CiteSeer distinct centers occasionally yield identical ego-graphs, at most one pair across any two splits per seed. The ogbn-arxiv extension in Appendix D.7 uses $N{=}64$ ego-graphs and the same 10 paired seeds.

Architecture. The encoder is a three-layer masked dense GIN with hidden dimension 64, mean-plus-max pooling, and an 8-d graph latent. The classifier is a two-layer MLP. The decoder uses a 16-d positional embedding, a two-layer node-feature MLP, and a symmetric bilinear edge head.

Training. We train with Adam (Kingma & Ba, 2015), learning rate 2×10^{-3}, weight decay 10^{-5}, and batch size 64. Phase-1 and Phase-2 step budgets are matched within each domain, with 500 steps for synthetic graphs, 250 for citation graphs, and 300 for QM9. Single-phase conditions run one phase of that budget. Table A.14 gives the resulting optimizer-update counts, and Appendix D.6 reports update-matched single-phase controls. Unless otherwise stated, we set the generative-loss weight to $\lambda = 1.0$ and the SEQ-ANCHOR penalty weight to $\lambda_{\text{anc}} = 10^{-3}$. For the primary controlled grid, learning rate, width, and depth follow fixed pilot-study defaults shared across conditions and datasets. Every 50 steps, during Phase 2 for two-phase conditions, we score validation accuracy and restore the best checkpoint at the end of training. G-ONLY has no comprehension head and is selected by the gen-MMD of 200 samples against the first 200 validation graphs, with the prior fitted

on the pool. The primary controlled grid uses 10 paired seeds. The same 10-seed pairing is used for the in-protocol Erdős-Rényi and GraphVAE baselines in Appendix D.2, Pareto computations in Appendix D.5, the update-matched controls in Appendix D.6, the additional controls and per-graph checks in Appendix D.9, and the ogbn-arxiv study in Appendix D.7. The GAT ablation in Appendix D.1 and the λ sweep in Appendix D.4 use 5 seeds, the latter seeds 0 to 4 of the primary grid. The single-GPU DiGress run in Appendix D.3 is a reference row.

Update accounting. Table A.14 reports the measured optimizer-update counts. Losses use mean reduction within a mini-batch (cross-entropy over the graphs of the batch, and $\mathcal{L}_{\text{gen}}$ over real nodes and real node pairs of the batch). The single-phase objective sums the per-mini-batch comprehension losses, so its per-update objective is $\sum_{b=1}^{B} \mathcal{L}_{\text{comp}}^{(b)} + \lambda \mathcal{L}_{\text{gen}}$ rather than an average over b. The loss-averaged control of Appendix D.6 divides the sum by B.

Node ordering. Equation (1) compares decoder output and input slot by slot, so the node order of every graph matters. Synthetic SBM graphs are generated with community-contiguous node indices (all nodes of community 0 precede those of community 1, and so on) and no permutation is applied, so slot index is informative about community membership. Citation ego-graphs place the seed node in slot 0, followed by its 1-hop neighbors and then its 2-hop neighbors in set-iteration order. Ego-graphs with more than 24 nodes keep the seed and a uniformly random subset of the remaining nodes, in sampling order. QM9 molecules keep the atom order of the PyG QM9 source files, in which heavy atoms precede hydrogens for 99.7% of molecules. The 2.2% of molecules with more than 24 atoms lose only hydrogens under truncation. No canonicalization, matching, or permutation augmentation is used during training or evaluation, and round-trip evaluation uses the same order as training. The encoder (sum-aggregation GIN with mean and max pooling) is permutation invariant, whereas the decoder reconstructs slots through a learned positional embedding, so the reconstruction loss is slot aligned rather than permutation invariant. This is a design limitation of the compact model: the decoder learns to reconstruct graphs in the order the dataset presents them. Appendix D.9 therefore reports per-graph reconstruction both in slot order and after Hungarian matching of decoded to input nodes.

Statistical protocol. For the primary controlled grid and the ogbn-arxiv study in Appendix D.7, descriptor MMD and C2ST use 400 generated graphs per condition and seed. The primary comparisons use 10 paired seeds that index identical partitions across conditions. Seed-aggregated result tables report mean $\pm$ standard deviation across seeds. Paired-bootstrap 95% CIs from 10,000 resamples are used for paired significance tests and Pareto-front computation. A contrast is significant when the CI for the paired difference excludes 0. Unless a caption states otherwise, seed-aggregated tables in this appendix use these 10 seeds, difference tables give the paired mean difference with this CI, and bold marks a CI that excludes zero. The Holm correction of Appendix D.8 covers the primary SEQ-REG contrasts. The paired CIs of the hyperparameter sweeps and of the supplementary controls are per-contrast intervals without a family-wise correction.

Compute. We use $1\times$ Quadro P5000 (16 GB), WSL2 + CUDA 12.1, and torch 2.5.1. Parameter count is $\sim$56 K for GIN and $\sim$44 K for GAT, measured via `sum(p.numel())`. Table A.6 reports the measured training and sampling cost on QM9.

Environment. All primary-grid tables are generated from a single run of the grid in one software environment (torch 2.5.1 with CUDA 12.1, numpy 2.1.3, PyG 2.7.0, one Quadro P5000). This run reproduces an earlier run of the grid bit for bit on Synthetic, Cora, CiteSeer, and QM9. PubMed values moved at the third decimal after a numpy update on the experiment machine and are reported from the final run throughout.

C Evaluation Protocol Details

This section specifies the metrics, diagnostic probes, and deployment procedures used by the main paper.

Descriptor MMD. Descriptor MMD averages degree MMD, clustering MMD, and spectral MMD against a held-out reference set (You et al., 2018; Liao et al., 2019). Each graph is summarized by three 12-bin normalized histograms: node degrees on $[0, 24]$, local clustering coefficients on $[0, 1]$, and all eigenvalues of the symmetric normalized Laplacian $I - D^{-1/2} A D^{-1/2}$ on $[0, 2]$. With $N{=}24$

Table A.2: Descriptor MMD with the kernel bandwidth fixed by the reference set. For each domain, seed, and descriptor, γ is the inverse median of the non-zero pairwise squared distances among the 400 reference histograms alone and is shared by every generated and decoded set of that seed, in place of the pooled rule of Appendix C. Rows repeat the conditions of Tables 2, A.4, and 5 on retrained checkpoints that reproduce the stored grid except on PubMed seed 1 (every condition) and PubMed seed 7 (GraphVAE).

	Synthetic SBM	CiteSeer	Cora	PubMed	QM9
Generation MMD, refit prior					
G-ONLY	0.004 ± 0.002	0.012 ± 0.005	0.039 ± 0.006	0.077 ± 0.014	0.234 ± 0.012
SEQ-REG	0.006 ± 0.003	0.012 ± 0.005	0.036 ± 0.006	0.073 ± 0.015	0.258 ± 0.017
JOINT-RETAINED	0.012 ± 0.009	0.021 ± 0.010	0.057 ± 0.014	0.103 ± 0.015	0.274 ± 0.028
NAIVE-DECOUPLED	0.904 ± 0.037	0.291 ± 0.101	0.420 ± 0.104	0.284 ± 0.107	0.448 ± 0.021
SEQ-FREEZE	0.334 ± 0.238	0.441 ± 0.165	0.447 ± 0.190	0.432 ± 0.099	0.463 ± 0.140
SEQ-ANCHOR	0.256 ± 0.204	0.379 ± 0.130	0.438 ± 0.150	0.359 ± 0.087	0.426 ± 0.136
Generation MMD, cached prior					
SEQ-FREEZE, SEQ-ANCHOR	0.009 ± 0.005	0.018 ± 0.010	0.044 ± 0.009	0.105 ± 0.025	0.278 ± 0.027
Generation MMD, references					
Erdős-Rényi null	0.146 ± 0.010	0.081 ± 0.005	0.085 ± 0.008	0.238 ± 0.011	0.211 ± 0.007
GraphVAE	0.024 ± 0.006	0.008 ± 0.004	0.057 ± 0.015	0.167 ± 0.103	0.267 ± 0.020
Round-trip MMD					
SEQ-REG	0.003 ± 0.004	0.017 ± 0.011	0.034 ± 0.006	0.071 ± 0.034	0.262 ± 0.019
JOINT-RETAINED	0.005 ± 0.004	0.015 ± 0.007	0.053 ± 0.013	0.098 ± 0.012	0.270 ± 0.023
SEQ-FREEZE-ACTIVE	0.008 ± 0.004	0.015 ± 0.006	0.032 ± 0.006	0.076 ± 0.020	0.268 ± 0.019
SEQ-FREEZE	0.314 ± 0.202	0.353 ± 0.187	0.350 ± 0.172	0.398 ± 0.106	0.458 ± 0.144
SEQ-ANCHOR	0.246 ± 0.188	0.291 ± 0.138	0.335 ± 0.152	0.331 ± 0.080	0.417 ± 0.131

no degree in the primary grid exceeds 23. On the $N{=}64$ ogbn-arxiv graphs of Appendix D.7, where 3.8% of reference nodes, in 48% of reference graphs, have degree above 24, the degree histogram instead spans $[0, 64]$ with 32 bins of the same width. For each statistic we compute the squared MMD with a single-bandwidth RBF kernel whose γ is the inverse median of the non-zero pairwise squared distances over the pooled reference and generated histograms, using the unbiased estimator (diagonal terms removed from the within-sample means), clipped at zero. Because this γ depends on the generated set, Table A.2 rescores the generated, cached, and decoded sets of Tables 2, A.4, and 5 with γ fixed per domain, seed, and descriptor from the reference histograms alone. Every paired contrast of those tables that involves SEQ-REG, the refit-versus-cached contrast of each frozen condition, and the round-trip contrast of SEQ-FREEZE-ACTIVE against SEQ-FREEZE keep their sign and significance under the fixed kernel. The reported gen-MMD is the arithmetic mean of the three values. The reference set is the first 400 held-out test graphs of the seed, and the generated set contains 400 graphs whose node counts are drawn from the reference size distribution. We retain descriptor MMD as the standard structural score and use its three component MMDs as a diagnostic decomposition.

Auxiliary generation diagnostics. We use the following diagnostics to interpret descriptor MMD and C2ST. They are diagnostic checks for duplication, memorization, and specific structural deviations from the reference distribution.

- **Uniqueness** measures the fraction of generated graphs that are pairwise distinct under a signature based on node count, edge count, and sorted degree sequence. Because the signature is a coarsening of graph isomorphism, distinct graphs can collide and the reported uniqueness is a lower bound on the true value.
- **Novelty-NN** is the generated-to-training nearest-neighbor distance, normalized by the held-out real-to-real distance. Values near 1.0 indicate novelty comparable to the real-to-real spread, whereas values $\ll 1.0$ would indicate memorization. Values $\gg 1.0$ place generated graphs, on average, farther from the training graphs than held-out real graphs are, and require joint interpretation with MMD because they can also indicate off-distribution generation. Novelty-NN is used as a memorization diagnostic.

- **Per-statistic MMD** decomposes descriptor MMD into degree, clustering, and spectral MMD.

In the primary controlled grid, SEQ-REG uniqueness across the five domains ranges from 0.63 on CiteSeer to 0.999 on Synthetic, and novelty-NN ranges from 0.98 to 3.35. The smallest novelty-NN value is consistent with real-to-real spread. The QM9 value, 3.35, places generated molecules, on average, far from the training graphs and, together with the larger QM9 MMD discussed in Appendix D.2, indicates off-distribution molecular structure. SEQ-FREEZE's uniqueness drops to 0.21 to 0.37 on the citation domains, consistent with the encoder-decoder drift discussed in the main text.

Classifier two-sample test (C2ST). For each condition, domain, and seed we draw the same 400 generated graphs used for descriptor MMD and pair them with the first 400 held-out real graphs, labeling real as 1 and generated as 0. Node attributes are withheld. Each node receives three structural inputs, a node-present indicator, its degree, and the sum of its neighbors' degrees, so the score measures structural realism rather than feature reconstruction. The discriminator is a fresh 3-layer masked dense GIN of hidden width 32 with mean-plus-max pooling and a two-layer MLP head, with no weight sharing with the model under test, trained with Adam (learning rate 2×10^{-3}, weight decay 10^{-5}, batch size 64) for 80 epochs without learning-rate schedule or early stopping, and the last-epoch model is evaluated. Real graph i and generated graph i are assigned to the same fold of a 70/30 split, and we report balanced accuracy $\frac{1}{2}(\mathrm{TPR} + \mathrm{TNR})$ at threshold 0.5 on the 30% fold (120 graphs per side), which prevents residual class imbalance from inflating the score. A value of 0.5 corresponds to chance-level discrimination, the best attainable result for a generator, and 1.0 corresponds to perfect separation. The discriminator seed equals the paired experiment seed of the model under test. Single sample sets (DiGress) are scored with discriminator seeds 0 to 4. The test is repeated independently per seed and aggregated with the paired-bootstrap protocol used for the primary comparisons. All C2ST values in this paper use this one discriminator configuration.

Density-matched null. The null is an Erdős-Rényi $G(n, p)$ sampler whose n is drawn from the empirical size distribution of the unlabeled pool and whose p is the maximum-likelihood edge density of the masked pool graphs (Appendix D.2). Pool and reference graphs are iid splits of one distribution, so the null matches the reference in expectation on node count and pooled edge density and carries no other structure. It does not match the per-graph density distribution or the dependence of density on graph size, so a generator that only reproduces how density varies across graphs can improve on it. Improving on the null therefore shows structure beyond graph size and one pooled edge density, not beyond per-graph density. We pass it through the identical C2ST and descriptor MMD pipelines. The decision rule is as follows. A condition improves on the null on a domain when the paired-bootstrap 95% CI of its C2ST minus the null's C2ST over the same seeds lies below zero. Its paired difference to the real-versus-real floor is computed the same way, and a CI that includes zero there means only that the difference is not resolved, not that the condition is indistinguishable from real graphs. Improving on the null is necessary for a usable generator and does not establish distributional equality. The null gives the score an explicit rejection criterion, and on QM9 it shows that descriptor MMD understates discrepancies. Table A.18 reports the paired CIs against the null together with a real-versus-real floor, obtained by scoring two disjoint halves of the held-out real graphs against each other, whose expected C2ST is 0.5 at any sample size, and a label-shuffle check that pools real and null graphs and re-splits them at random.

Prior deployment. When the deployed decoder remains tied to Phase-1 latents, we evaluate two deployments of the same trained weights. The refit deployment encodes the training pool with the Phase-2 encoder and fits $\mathcal{N}(\mu, \Sigma)$ to those latents. The cached deployment reuses the (μ, Σ) snapshotted at the end of Phase 1, before comprehension fine-tuning moved the encoder. The weights, decoder, sampling procedure, and thresholding procedure are otherwise unchanged. Where the distinction matters we write SEQ-FREEZE^c and SEQ-ANCHOR^c for the cached-prior variants. The cache stores a mean vector and a Cholesky factor, under 50 floats at d_z=8, alongside the model. This distinction matters for the frozen-decoder conditions. It is absent for MAE-DISCARD, which retains no decoder, and SEQ-REG uses the refit deployment because its encoder and decoder co-adapt during Phase 2.

Linear CKA. Linear CKA (Kornblith et al., 2019) is computed between pooled test-set encoder representations at the end of Phase 1 and at the end of Phase 2. For single-phase conditions with no Phase 1, the reference representation is captured at initialization. The resulting score is inter-

preted as encoder drift for JOINT-RETAINED and pretrained-structure preservation for the two-phase conditions.

Round-trip reconstruction. Round-trip complements unconditional sampling by measuring how compatible the current encoder and the deployed decoder are. We encode the held-out test graphs with the Phase-2 encoder, decode the resulting latents with the deployed decoder using the same edge thresholding as unconditional generation, and compute the same three-statistic structural MMD against the input graphs. Lower round-trip MMD means that the decoded set is closer to the input set on the three descriptors. Like generation MMD, it does not certify distributional equality, and we read it, with round-trip C2ST, as a compatibility score relative to the frozen-decoder controls. The score is set-level: permuting decoded graphs across inputs leaves it unchanged, so it does not by itself certify per-input recovery. Cached-prior deployment supplies a sampling distribution, while round-trip reconstruction measures, at the distribution level, how well the deployed decoder accepts the latents of the current encoder. Round-trip C2ST applies the discriminator of the generation C2ST to the first $\min(400, n_{\text{test}})$ test graphs against their own decodings. Appendix D.9 adds the per-graph protocol, which decodes each test graph's own latent against a size-matched foreign latent.

Generator families. The three generative mechanisms compared in the main paper share the encoder and decoder architectures and the data. The default is the Gaussian prior fitted after training, as described above. The latent-diffusion prior standardizes the encoded pool latents (d_z=8) per dimension and trains a DDPM with T=200 steps and a linear β schedule from 10^{-4} to 0.02 to predict the noise with a 3-layer MLP denoiser (hidden width 128, SiLU activations, 32-dimensional sinusoidal time embedding) for 1500 Adam updates (learning rate 2×10^{-3}, batch size 128), after the autoencoder is trained. Samples use 200-step ancestral sampling from $\mathcal{N}(0, I)$, are de-standardized, and are decoded by the unchanged decoder. It differs from the default only in how latents are drawn. The variational condition instead keeps the encoder of the deterministic model and maps its pooled output by two linear layers to a diagonal Gaussian posterior $(\mu, \log \sigma^2)$ over d_z=8. The classifier reads μ and the decoder is unchanged. The negative ELBO is the reconstruction loss of Eq. (1) on a reparameterized sample plus β times the mean KL to $\mathcal{N}(0, I)$ with $\beta = 10^{-3}$. Training follows the two-phase SEQ-REG schedule: S ELBO updates on pool mini-batches of 64, then S Phase-2 outer iterations with one update per labeled mini-batch of 64 and a fresh comprehension head. Each Phase-2 update of VAE-reg minimizes $\mathcal{L}_{\text{comp}} + \lambda \,\text{ELBO}$ with $\lambda = 1$ and the ELBO on one pool mini-batch, and VAE-freeze (Appendix D.8) freezes the decoder and minimizes $\mathcal{L}_{\text{comp}}$ alone. The best validation-accuracy snapshot is kept. Generation samples $\mathbf{z} \sim \mathcal{N}(0, I)$ and decodes, with no post-hoc prior. For the refit column of Table A.21, a Gaussian is instead fitted to the mean and covariance of the Phase-2 posterior means of the pool graphs. This refit omits the average posterior variance, so it is narrower than the aggregate posterior. This condition changes the latent itself rather than the prior over it and tests how a variational bottleneck affects the shared classifier.

D ADDITIONAL EXPERIMENT ANALYSES

D.1 GAT BACKBONE ABLATION

As a backbone check, we re-run the synthetic domain with a single-head dense GAT (Veličković et al., 2018) encoder of comparable parameter count (3 layers, hidden 64, $d_z = 8$, same step budgets, $\lambda = 1.0$, 5 seeds).

Table A.3 reports the result under the refit deployment. Absolute accuracy shifts under the GAT encoder, with C-ONLY dropping from 0.794 to 0.542. SEQ-REG reaches gen-MMD 0.004, matching the G-ONLY reference of 0.005 and far below SEQ-FREEZE's 0.459 and SEQ-ANCHOR's 0.302, and is second in accuracy only to NAIVE-DECOUPLED, whose generator is the weakest. The generation pattern of the primary grid therefore transfers to the GAT backbone.

D.2 EXTERNAL GRAPH-GENERATION BASELINES

We evaluate two in-protocol baselines with the same MMD pipeline as the main conditions.

Erdős-Rényi $G(n, p)$. Edge probability p is fitted by maximum likelihood on the masked pool, and n is drawn from the empirical pool-size distribution. The learned generators mask samples to sizes drawn from the held-out reference-size distribution. Pool and reference sizes are iid splits of the

Table A.3: GAT-backbone ablation on Synthetic SBM under the refit deployment, mean ± standard deviation over 5 seeds. N/A marks a missing task head.

Condition	Test accuracy	gen-MMD
C-ONLY	0.542 ± 0.021	N/A
G-ONLY	N/A	0.005 ± 0.004
MAE-DISCARD	0.563 ± 0.025	N/A
SEQ-FREEZE	0.563 ± 0.025	0.459 ± 0.142
SEQ-ANCHOR	0.596 ± 0.028	0.302 ± 0.165
SEQ-REG	0.620 ± 0.047	0.004 ± 0.002
NAIVE-DECOUPLED	0.637 ± 0.023	1.290 ± 0.016
JOINT-RETAINED	0.577 ± 0.042	0.015 ± 0.013

Table A.4: Descriptor generation MMD of the density-matched Erdős-Rényi null, a GraphVAE trained on the same unlabeled pool at the single-phase budget of G-ONLY, and SEQ-REG, scored through one pipeline on the same 10 paired seeds and 400 generated graphs per seed. The upper panel averages the degree, clustering, and spectral MMDs, and its SEQ-REG column is that of Table 2. The lower panel splits the QM9 row by descriptor.

Dataset	Erdős-Rényi null	GraphVAE	SEQ-REG
Mean over the three descriptors			
Synthetic SBM	0.149 ± 0.010	0.024 ± 0.006	0.006 ± 0.003
CiteSeer	0.078 ± 0.005	0.008 ± 0.005	0.012 ± 0.005
Cora	0.087 ± 0.008	0.059 ± 0.015	0.036 ± 0.006
PubMed	0.242 ± 0.012	0.172 ± 0.108	0.075 ± 0.014
QM9	0.245 ± 0.012	0.280 ± 0.023	0.269 ± 0.021
QM9 by descriptor			
Degree	0.552 ± 0.027	0.395 ± 0.046	0.356 ± 0.051
Clustering	0.133 ± 0.012	0.295 ± 0.077	0.337 ± 0.028
Spectral	0.050 ± 0.005	0.150 ± 0.020	0.115 ± 0.008

same ego-graph size distribution. The two choices therefore coincide in expectation. This sampler is the null random-graph baseline for learned generators.

Whole-graph VAE. A compact whole-graph VAE based on GraphVAE (Simonovsky & Komodakis, 2018), sharing the same encoder and decoder families as G-ONLY, with a $(\mu, \log \sigma^2)$ posterior and KL term with $\beta = 10^{-3}$. It is trained from scratch on the pool for the same number of updates as G-ONLY. At inference, we sample $\mathbf{z} \sim \mathcal{N}(0, I)$ and decode a whole padded graph. This is the graph-set-generating GraphVAE of Simonovsky and Komodakis, distinct from the link-prediction VGAE of Kipf and Welling (Kipf & Welling, 2016).

Table A.4 reports the cross-domain comparison among the density-matched random-graph null, the GraphVAE baseline at the G-ONLY budget, and SEQ-REG. SEQ-REG is lowest on Synthetic SBM, Cora, and PubMed. CiteSeer favors the compact GraphVAE baseline, while QM9 favors the density-matched null under descriptor MMD. The lower panel of Table A.4 splits the QM9 row by descriptor: the null is worst on degree but closest to the reference on the clustering and spectral statistics, which is the weakness of descriptor MMD that RQ1 addresses.

Table A.5 re-scores the single-GPU DiGress run described in Appendix D.3 after matching the graph representation, and Table A.6 gives the measured compute of every generator compared in this paper.

D.3 SINGLE-GPU DIGRESS REFERENCE

Setup and scoring. We train DiGress (Vignac et al., 2023) from scratch on QM9 on the same Quadro P5000 for 30 epochs (62 minutes, 11,730 optimizer updates, batch size 256) with its `remove_h` configuration, which generates heavy-atom molecules of at most 9 nodes, whereas the reference graphs of this paper keep hydrogens. Table A.5 therefore matches the representation in both directions, by re-adding hydrogens to the samples with RDKit and by removing them from the

Table A.5: Re-scored single-GPU DiGress on QM9. DiGress generates heavy-atom molecules of at most 9 nodes under its `remove_h` configuration, whereas the reference graphs of this paper keep hydrogens (mean 18 nodes, truncated to 24), so the final samples of its epoch-29 checkpoint are scored against the seed-0 test set either after re-adding hydrogens with RDKit or with hydrogens removed from the test set. n is the number of scored graphs, and C2ST is the mean $\pm$ standard deviation over 5 discriminator seeds. DiGress rows compare 400 generated with 400 reference graphs, and the real-versus-real rows split the reference set into halves of 200.

			descriptor MMD ↓				
Samples	n	Nodes	Degree	Clust.	Spectral	Mean	C2ST ↓
All-atom reference (hydrogens kept, as elsewhere in the paper)							
DiGress, hydrogens re-added	400	18.7	0.017	0.000	0.001	0.006	0.502 ± 0.023
Real vs. real	200	18.0	0.000	0.005	0.000	0.002	0.490 ± 0.036
Heavy-atom reference (hydrogens removed from the test set)							
DiGress, heavy atoms only	400	8.8	0.000	0.000	0.000	0.000	0.533 ± 0.048
Real vs. real	200	8.8	0.014	0.004	0.000	0.006	0.510 ± 0.022

Table A.6: Compute accounting on QM9 on one Quadro P5000 (16 GB): optimizer updates, training wall-clock, peak GPU memory, time to draw 400 samples, and parameter count. Peak memory is `max_memory_allocated` for our models and GraphVAE and the nvidia-smi increment over the idle baseline during sampling for DiGress. The DiGress sampling time is the wall-clock difference between two `test_only` runs with 400 and 0 samples and includes its RDKit validity pass. The external rows share the GPU but do not equalize updates, wall-clock, or parameters.

Model	Updates	Train time	Peak mem. (MB)	400 samples (s)	Params
SEQ-REG, QM9 (N=24)	900	15 s	655	0.01	56 K
G-ONLY, QM9 (N=24)	300	6 s	650	0.01	56 K
GraphVAE, QM9 (N=24)	300	3 s	643	0.00	42 K
DiGress (`remove_h`, 30 epochs)	11730	62 min	877	107.40	N/A

reference. Once the representation is matched, DiGress reaches gen-MMD 0.006 and C2ST 0.502 against 0.490 for real versus real with hydrogens re-added, and its C2ST lies within 0.025 of the floor in either direction, so a dedicated molecular generator trained for 62 minutes on one GPU comes close to the real-versus-real floor on QM9 under our evaluator. Published DiGress results use a substantially larger multi-GPU training regime.

Compute. DiGress and GraphVAE were trained from scratch on the same Quadro P5000, but the runs do not equalize optimizer updates, wall-clock, or parameters. Table A.6 gives the measured values: training takes 15 seconds for SEQ-REG on QM9, 3 seconds for GraphVAE, and 62 minutes for DiGress. The external rows are therefore references at single-GPU scale, not budget-matched competitors, and they support no claim about which architecture is better at equal cost.

Latent Graph Diffusion. We also examined the official code of Latent Graph Diffusion (Zhou et al., 2024) for a reference row. Its shipped configurations cover unconditional molecular generation on QM9, decoded into RDKit molecules, and node-level classification. No shipped configuration performs unconditional generation of generic graphs or graph-level classification. Running it on the splits of this paper would require a new dataset class and new encoder, diffusion, and prediction-head configurations with no official default, which the reference-row condition of a shipped configuration without tuning excludes, so no LGD row is reported.

D.4 HYPERPARAMETER SENSITIVITY: λ AND ANCHOR-WEIGHT SWEEPS

Table A.7 sweeps $\lambda \in \{0.1, 0.3, 1, 3, 10\}$ for SEQ-REG on Cora with seeds 0 to 4 of the primary grid. Over two orders of magnitude accuracy varies by under four points, and both accuracy and gen-MMD improve as λ grows, so the default $\lambda = 1$ is not tuned in favor of SEQ-REG. Fig. A.1 varies λ, d_z, and the label budget on Synthetic SBM for JOINT-RETAINED, which optimizes the same $\mathcal{L}_{\text{comp}} + \lambda \mathcal{L}_{\text{gen}}$ objective as SEQ-REG. Its accuracy is stable for λ from 0.5 to 2 and across

Table A.7: Sensitivity of SEQ-REG to the generative-loss weight λ on Cora, with seeds 0 to 4 of the primary grid, so the $\lambda = 1$ row is the primary SEQ-REG run on those seeds. Mean $\pm$ standard deviation over these five seeds.

λ	Accuracy	Gen-MMD
0.1	0.677 ± 0.021	0.0456 ± 0.0116
0.3	0.677 ± 0.022	0.0408 ± 0.0107
1.0	0.689 ± 0.040	0.0359 ± 0.0075
3.0	0.713 ± 0.040	0.0346 ± 0.0082
10.0	0.716 ± 0.030	0.0265 ± 0.0053

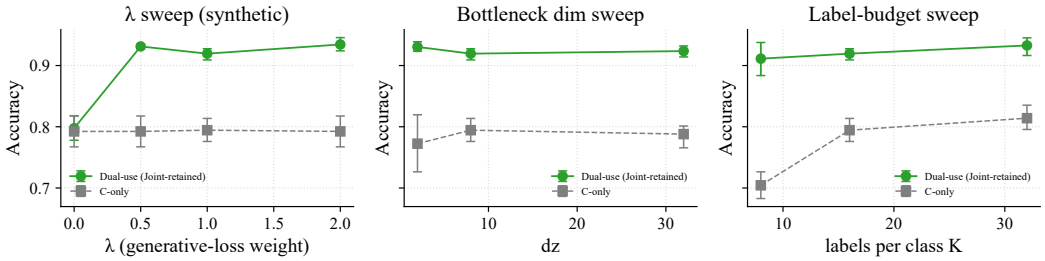


Figure A.1: Ablations on synthetic SBM with 95% CIs. The dual-use representative is JOINT-RETAINED because it uses the same $\mathcal{L}_{\text{comp}} + \lambda\mathcal{L}_{\text{gen}}$ objective as SEQ-REG. Left panel varies the generative-loss weight. Center panel varies bottleneck width d_z from 2 to 32. Right panel varies the label budget, where $\Delta C(\text{JOINT-RETAINED} - \text{C-ONLY})$ shrinks from $+0.207^*$ at $K = 8$ to $+0.119^*$ at $K = 32$. The * marker denotes paired-bootstrap 95% CI excluding 0.

the tested bottleneck widths. Table A.11 sweeps the bottleneck width of SEQ-REG itself over $d_z \in \{2, 4, 8, 16, 32\}$ on all five domains with the 10 paired seeds and the other hyperparameters of the grid. Narrowing the bottleneck hurts: d_z=2 lowers accuracy significantly on Cora, CiteSeer, PubMed, and QM9, and $d_z \in \{2, 4\}$ raise C2ST significantly on CiteSeer and PubMed, d_z=2 also on Cora. Widening to 16 or 32 raises accuracy significantly only on PubMed, by 0.03 and 0.04, and changes C2ST significantly only at d_z=32, where it falls by 0.025 on CiteSeer and rises by 0.029 on Synthetic. At every width, C2ST stays above 0.69 on the citation domains and above 0.91 on QM9.

Table A.8 sweeps the L2-SP weight of SEQ-ANCHOR over $\lambda_{\text{anc}} \in \{10^{-4}, \ldots, 10\}$ with 10 paired seeds. The weight $\lambda_{\text{anc}} = 1$ matches or exceeds SEQ-REG on both accuracy and CKA on Synthetic, Cora, and PubMed. On CiteSeer the anchored encoder stays significantly less accurate at every weight, and on QM9 accuracy falls monotonically as the anchor tightens, from 0.840 at 10^{-2} to 0.704 at 10, while CKA rises to 0.987. Tight anchoring also restores refit generation for the frozen decoder, because the encoder barely moves, but at the cost of accuracy on the molecular domain. Table A.9 gives the paired differences. Against SEQ-REG, the anchored encoder is significantly more accurate on Cora at $\lambda_{\text{anc}} \in \{1, 10\}$, by 0.026 and 0.014, and significantly less accurate on CiteSeer at every weight and on QM9 at $\lambda_{\text{anc}} = 10^{-3}$ and at $\lambda_{\text{anc}} \geq 0.1$. Against MAE-DISCARD it is significantly more accurate at every weight on Synthetic and at $\lambda_{\text{anc}} \in \{1, 10\}$ on Cora, and significantly less accurate on PubMed at $\lambda_{\text{anc}} \in \{0.1, 10\}$ and on QM9 at $\lambda_{\text{anc}} \geq 0.1$. At $\lambda_{\text{anc}} = 10$ the refit generation MMD remains significantly above SEQ-REG on the four non-molecular domains, by 0.003 to 0.034.

Table A.10 tracks the refit penalty of the frozen decoder, the refit minus the cached deployment, across the sweep. Without an anchor, SEQ-FREEZE pays 0.035 to 0.287 in C2ST, significantly on every domain, and the penalty stays significant on every domain up to $\lambda_{\text{anc}} = 10^{-3}$. It shrinks as the anchor tightens and CKA rises (Table A.8), and at $\lambda_{\text{anc}} \geq 1$ it is at most 0.020 in C2ST and significant only on CiteSeer. The refit penalty therefore scales with encoder drift. Tables 1, 2, and 3 report the $\lambda_{\text{anc}} = 1$ anchor as SEQ-ANCHOR $_1$, the smallest tested weight at which the refit and cached priors agree within 0.02 in C2ST on every domain. They also agree within 0.02 in descriptor MMD: the refit is significantly worse only on CiteSeer, by 0.020 in C2ST and 0.008 in MMD, and significantly better in MMD on PubMed, by 0.014. Against SEQ-REG, SEQ-ANCHOR $_1$ has significantly higher refit MMD on the four non-molecular domains, by 0.004 to 0.020, and

Table A.8: L2-SP anchor-weight sweep for SEQ-ANCHOR, with SEQ-REG for reference, where $\lambda_{\text{anc}} = 10^{-3}$ is the primary-grid setting. Generation and round-trip MMD use the refit deployment. The cached-prior deployment pairs the Phase-1 prior with the unchanged Phase-1 decoder, so it does not depend on λ_{anc}, and its generation MMD is the cached value of Table 2 at every weight. Table A.9 gives the paired differences.

λ_{anc}	Acc.	CKA	Gen-MMD	Round-trip MMD
Synthetic SBM				
10^{-4}	0.886 ± 0.028	0.705 ± 0.060	0.395 ± 0.328	0.382 ± 0.279
10^{-3}	0.915 ± 0.020	0.865 ± 0.066	0.301 ± 0.260	0.288 ± 0.237
10^{-2}	0.923 ± 0.015	0.940 ± 0.054	0.106 ± 0.148	0.100 ± 0.141
10^{-1}	0.935 ± 0.009	0.974 ± 0.017	0.017 ± 0.011	0.013 ± 0.012
1	0.935 ± 0.008	0.980 ± 0.013	0.010 ± 0.006	0.007 ± 0.008
10	0.937 ± 0.008	0.980 ± 0.013	0.009 ± 0.006	0.006 ± 0.008
SEQ-REG	0.924 ± 0.028	0.863 ± 0.031	0.006 ± 0.003	0.003 ± 0.004
CiteSeer				
10^{-4}	0.563 ± 0.035	0.321 ± 0.108	0.382 ± 0.137	0.307 ± 0.160
10^{-3}	0.574 ± 0.034	0.338 ± 0.074	0.353 ± 0.124	0.272 ± 0.132
10^{-2}	0.572 ± 0.032	0.388 ± 0.089	0.233 ± 0.100	0.184 ± 0.113
10^{-1}	0.581 ± 0.033	0.579 ± 0.086	0.084 ± 0.072	0.053 ± 0.037
1	0.592 ± 0.029	0.812 ± 0.038	0.026 ± 0.017	0.025 ± 0.021
10	0.595 ± 0.028	0.972 ± 0.019	0.019 ± 0.013	0.020 ± 0.015
SEQ-REG	0.616 ± 0.024	0.518 ± 0.054	0.012 ± 0.005	0.017 ± 0.011
Cora				
10^{-4}	0.681 ± 0.027	0.378 ± 0.078	0.409 ± 0.156	0.300 ± 0.142
10^{-3}	0.683 ± 0.029	0.425 ± 0.071	0.418 ± 0.148	0.325 ± 0.153
10^{-2}	0.690 ± 0.019	0.419 ± 0.083	0.319 ± 0.131	0.236 ± 0.114
10^{-1}	0.684 ± 0.025	0.519 ± 0.064	0.086 ± 0.066	0.069 ± 0.059
1	0.709 ± 0.030	0.826 ± 0.033	0.043 ± 0.011	0.035 ± 0.008
10	0.697 ± 0.025	0.954 ± 0.021	0.041 ± 0.008	0.032 ± 0.006
SEQ-REG	0.683 ± 0.033	0.492 ± 0.075	0.036 ± 0.006	0.034 ± 0.006
PubMed				
10^{-4}	0.687 ± 0.027	0.265 ± 0.059	0.416 ± 0.101	0.382 ± 0.102
10^{-3}	0.692 ± 0.025	0.347 ± 0.077	0.364 ± 0.086	0.338 ± 0.077
10^{-2}	0.683 ± 0.020	0.454 ± 0.098	0.234 ± 0.100	0.206 ± 0.098
10^{-1}	0.667 ± 0.042	0.713 ± 0.105	0.085 ± 0.028	0.098 ± 0.053
1	0.675 ± 0.038	0.838 ± 0.071	0.095 ± 0.023	0.093 ± 0.046
10	0.657 ± 0.040	0.932 ± 0.038	0.109 ± 0.032	0.118 ± 0.058
SEQ-REG	0.671 ± 0.047	0.650 ± 0.081	0.075 ± 0.014	0.074 ± 0.036
QM9				
10^{-4}	0.827 ± 0.045	0.332 ± 0.112	0.614 ± 0.228	0.609 ± 0.230
10^{-3}	0.828 ± 0.040	0.360 ± 0.123	0.519 ± 0.187	0.506 ± 0.187
10^{-2}	0.840 ± 0.029	0.476 ± 0.182	0.339 ± 0.084	0.333 ± 0.089
10^{-1}	0.819 ± 0.032	0.708 ± 0.144	0.297 ± 0.062	0.283 ± 0.054
1	0.771 ± 0.048	0.921 ± 0.037	0.287 ± 0.035	0.283 ± 0.034
10	0.704 ± 0.030	0.987 ± 0.006	0.286 ± 0.029	0.289 ± 0.031
SEQ-REG	0.840 ± 0.034	0.708 ± 0.087	0.269 ± 0.021	0.273 ± 0.020

higher C2ST only on CiteSeer, by 0.040. It is significantly less separable than the null on every domain except QM9 (Table A.18), and its refit MMD is significantly higher than that of G-ONLY ^{m}, by 0.006 to 0.051.

D.5 UNCERTAINTY-AWARE PARETO-FRONT COMPUTATION

For each domain we form per-method per-task seed lists, with 10 seeds each for accuracy and a generation metric, gen-MMD or C2ST, then compute paired-bootstrap 95% CIs on every pairwise difference $\Delta = m_d - m_c$ using 10,000 resamples. Method d significantly dominates method c if d is significantly better on at least one task (CI excludes 0) and not significantly worse on the other. This

Table A.9: Paired differences for the L2-SP anchor-weight sweep of Table A.8. Each cell gives SEQ-ANCHOR at the stated weight minus the reference condition, and bold marks a CI that excludes zero. SEQ-REG and MAE-DISCARD are the primary-grid rows of Tables 3 and 2. Positive is better for accuracy and negative for MMD.

λ_{anc}	Synthetic SBM	CiteSeer	Cora	PubMed	QM9
Accuracy, SEQ-ANCHOR *minus* SEQ-REG*, paired* Δ *[95% CI]*					
10^{-4}	**−0.038** [−0.047, −0.029]	**−0.054** [−0.070, −0.039]	−0.002 [−0.017, +0.014]	+0.016 [−0.006, +0.041]	−0.013 [−0.027, +0.001]
10^{-3}	−0.009 [−0.026, +0.009]	**−0.043** [−0.061, −0.025]	−0.000 [−0.021, +0.020]	+0.021 [−0.002, +0.046]	**−0.013** [−0.024, −0.001]
10^{-2}	−0.001 [−0.011, +0.010]	**−0.044** [−0.056, −0.034]	+0.008 [−0.005, +0.022]	+0.012 [−0.014, +0.040]	−0.001 [−0.011, +0.011]
10^{-1}	+0.011 [−0.004, +0.032]	**−0.036** [−0.053, −0.019]	+0.001 [−0.018, +0.021]	−0.004 [−0.019, +0.012]	**−0.021** [−0.028, −0.014]
1	+0.012 [−0.004, +0.032]	**−0.024** [−0.041, −0.009]	**+0.026** [+0.006, +0.044]	+0.004 [−0.020, +0.030]	**−0.069** [−0.104, −0.036]
10	+0.013 [−0.002, +0.032]	**−0.022** [−0.039, −0.005]	**+0.014** [+0.004, +0.026]	−0.014 [−0.048, +0.023]	**−0.136** [−0.160, −0.112]
Accuracy, SEQ-ANCHOR *minus* MAE-DISCARD					
10^{-4}	**+0.043** [+0.031, +0.054]	−0.008 [−0.022, +0.007]	−0.001 [−0.020, +0.018]	−0.003 [−0.007, +0.002]	−0.008 [−0.020, +0.004]
10^{-3}	**+0.071** [+0.051, +0.094]	+0.003 [−0.010, +0.018]	+0.001 [−0.018, +0.018]	+0.002 [−0.006, +0.011]	−0.007 [−0.017, +0.003]
10^{-2}	**+0.079** [+0.063, +0.097]	+0.002 [−0.018, +0.023]	+0.008 [−0.007, +0.023]	−0.007 [−0.024, +0.011]	+0.004 [−0.008, +0.018]
10^{-1}	**+0.092** [+0.072, +0.117]	+0.010 [−0.019, +0.039]	+0.002 [−0.010, +0.015]	**−0.022** [−0.042, −0.001]	**−0.016** [−0.029, −0.004]
1	**+0.092** [+0.073, +0.116]	+0.021 [−0.005, +0.049]	**+0.027** [+0.012, +0.039]	−0.014 [−0.040, +0.013]	**−0.064** [−0.099, −0.029]
10	**+0.093** [+0.075, +0.117]	+0.024 [−0.004, +0.051]	**+0.014** [+0.006, +0.022]	**−0.032** [−0.061, −0.004]	**−0.131** [−0.157, −0.104]
Refit generation MMD, SEQ-ANCHOR *minus* SEQ-REG					
10^{-4}	**+0.389** [+0.211, +0.589]	**+0.370** [+0.289, +0.452]	**+0.374** [+0.277, +0.461]	**+0.341** [+0.279, +0.399]	**+0.344** [+0.213, +0.480]
10^{-3}	**+0.295** [+0.156, +0.455]	**+0.341** [+0.265, +0.413]	**+0.383** [+0.291, +0.467]	**+0.289** [+0.238, +0.338]	**+0.250** [+0.145, +0.369]
10^{-2}	**+0.100** [+0.026, +0.194]	**+0.221** [+0.163, +0.283]	**+0.284** [+0.207, +0.358]	**+0.159** [+0.104, +0.213]	**+0.069** [+0.027, +0.126]
10^{-1}	**+0.011** [+0.006, +0.017]	**+0.071** [+0.039, +0.120]	**+0.051** [+0.021, +0.093]	+0.011 [−0.008, +0.028]	+0.028 [−0.002, +0.063]
1	**+0.004** [+0.002, +0.006]	**+0.014** [+0.007, +0.024]	**+0.007** [+0.000, +0.016]	**+0.020** [+0.002, +0.039]	+0.018 [−0.009, +0.045]
10	**+0.003** [+0.001, +0.005]	**+0.007** [+0.002, +0.014]	**+0.006** [+0.001, +0.011]	**+0.034** [+0.012, +0.059]	+0.016 [−0.009, +0.042]

is a convention for declaring front membership under paired uncertainty. Not being significantly worse is not a non-inferiority guarantee, so membership means not significantly dominated rather than certified Pareto optimality. A method without a generator is assigned a generation MMD of $+\infty$. Method c is on the Pareto front if no other single-model method significantly dominates it. Table A.12 reports the resulting fronts for both generation metrics in four comparison sets: the grid under the refit deployment, the grid plus SEQ-ANCHOR $_1$, the grid under the cached prior, and the grid plus the update-matched controls. The cached-prior frozen conditions SEQ-FREEZE^c and SEQ-ANCHOR^c are included as single-model methods, deploying the same trained models with the Phase-1 prior, although their generator then bypasses the Phase-2 encoder (Sec. 4.4), while SEQ-REG keeps the refit prior. Within the eight-condition grid, SEQ-REG is on every front: under gen-MMD it is alone except on QM9, where JOINT-RETAINED and, under the cached prior, SEQ-FREEZE^c join it, and under C2ST JOINT-RETAINED joins it on every domain except Cora, while the cached frozen conditions join on Synthetic, Cora, and QM9. Adding SEQ-ANCHOR $_1$, SEQ-REG keeps every front except C2ST on Cora, where SEQ-ANCHOR $_1$ is significantly more accurate with no

Table A.10: Refit penalty of the frozen decoder across the L2-SP anchor sweep of Table A.8: the refit deployment minus the cached Phase-1 prior on the same seeds, for SEQ-FREEZE, which has no anchor, and for SEQ-ANCHOR at each weight. The cached deployment is the unchanged Phase-1 generator at every weight. Bold marks a CI that excludes zero, and Table A.8 gives the CKA of each weight.

λ_{anc}	Synthetic SBM	CiteSeer	Cora	PubMed	QM9
Refit minus cached C2ST, paired Δ [95% CI]					
none (SEQ-FREEZE)	**+0.287** [+0.203, +0.362]	**+0.170** [+0.140, +0.200]	**+0.070** [+0.044, +0.094]	**+0.065** [+0.023, +0.104]	**+0.035** [+0.013, +0.063]
10^{-4}	**+0.267** [+0.173, +0.352]	**+0.163** [+0.134, +0.195]	**+0.066** [+0.051, +0.080]	**+0.055** [+0.008, +0.096]	**+0.035** [+0.012, +0.063]
10^{-3}	**+0.252** [+0.159, +0.341]	**+0.155** [+0.123, +0.189]	**+0.080** [+0.059, +0.100]	**+0.040** [+0.005, +0.077]	**+0.035** [+0.013, +0.063]
10^{-2}	**+0.139** [+0.051, +0.230]	**+0.124** [+0.108, +0.142]	**+0.045** [+0.032, +0.058]	+0.009 [−0.017, +0.039]	**+0.025** [+0.005, +0.046]
10^{-1}	**+0.055** [+0.019, +0.085]	**+0.063** [+0.043, +0.086]	−0.004 [−0.032, +0.023]	−0.013 [−0.029, +0.008]	−0.015 [−0.079, +0.037]
1	+0.019 [−0.011, +0.050]	**+0.020** [+0.006, +0.035]	−0.012 [−0.035, +0.013]	−0.019 [−0.051, +0.012]	+0.011 [−0.025, +0.045]
10	+0.006 [−0.030, +0.043]	**+0.017** [+0.002, +0.033]	+0.003 [−0.013, +0.022]	−0.005 [−0.019, +0.008]	+0.012 [−0.018, +0.048]
Refit minus cached gen-MMD, paired Δ [95% CI]					
none (SEQ-FREEZE)	**+0.406** [+0.232, +0.604]	**+0.388** [+0.295, +0.476]	**+0.382** [+0.271, +0.483]	**+0.317** [+0.254, +0.375]	**+0.282** [+0.175, +0.397]
10^{-4}	**+0.385** [+0.209, +0.586]	**+0.364** [+0.281, +0.448]	**+0.367** [+0.272, +0.451]	**+0.307** [+0.242, +0.368]	**+0.324** [+0.195, +0.454]
10^{-3}	**+0.292** [+0.153, +0.452]	**+0.335** [+0.256, +0.408]	**+0.376** [+0.286, +0.459]	**+0.255** [+0.196, +0.310]	**+0.229** [+0.125, +0.344]
10^{-2}	**+0.097** [+0.023, +0.192]	**+0.215** [+0.155, +0.278]	**+0.277** [+0.201, +0.349]	**+0.125** [+0.061, +0.188]	**+0.049** [+0.009, +0.101]
10^{-1}	**+0.008** [+0.002, +0.014]	**+0.066** [+0.034, +0.114]	**+0.044** [+0.016, +0.085]	**−0.023** [−0.045, −0.004]	+0.008 [−0.018, +0.036]
1	+0.001 [−0.004, +0.005]	**+0.008** [+0.003, +0.014]	+0.000 [−0.003, +0.004]	**−0.014** [−0.023, −0.004]	−0.003 [−0.015, +0.010]
10	−0.000 [−0.005, +0.004]	+0.001 [−0.001, +0.004]	−0.001 [−0.003, +0.001]	+0.000 [−0.007, +0.008]	−0.004 [−0.015, +0.009]

significant C2ST difference and is the only member. SEQ-ANCHOR $_1$ also joins the Cora gen-MMD front and the Synthetic and PubMed C2ST fronts, where it displaces JOINT-RETAINED on Synthetic. The last comparison set adds the update-matched controls of Appendix D.6. A single-phase JOINT-RETAINED variant is then on every front, and on CiteSeer under gen-MMD the loss-averaged matched JOINT-RETAINED, which is significantly more accurate than SEQ-REG, dominates SEQ-REG and is the only front member. Front membership therefore depends on the metric and the comparison set: SEQ-REG is on every front except the matched gen-MMD front on CiteSeer and the SEQ-ANCHOR $_1$ C2ST front on Cora, but it rarely holds a front alone once C2ST, SEQ-ANCHOR $_1$, or the matched controls are considered. Table A.13 aggregates SEQ-REG against each single-model condition under the refit deployment.

D.6 UPDATE-MATCHED SINGLE-PHASE CONTROLS

Table A.14 reports the measured optimizer-update counts. Because Phase 2 takes one update per labeled mini-batch, a two-phase condition takes $S(1 + B)$ updates against S for a single-phase condition, so the standard grid is not update-matched. We therefore add update-matched controls that run the single-phase schedule for $S(1 + B)$ outer iterations, which equalizes both the update count and the pool exposure with SEQ-REG on every domain, while the labeled exposure becomes $1 + B$ times larger. The runs cover all five domains with the same 10 paired seeds and the same evaluation pipeline. Table A.15 reports JOINT-RETAINED at the standard and matched budgets, with the loss-averaged control described below, and Table A.16 the matched single-task references.

Table A.11: Bottleneck-width sweep for SEQ-REG over $d_z \in \{2, 4, 8, 16, 32\}$, with every other hyperparameter of the primary grid and the same 10 paired seeds. d_z=8 is the stored primary grid, which a reproduction run on seed 0 reproduces on every domain. Generation MMD and C2ST use the refit deployment. The lower panels give each width minus d_z=8, and bold marks a CI that excludes zero. Higher accuracy and lower MMD and C2ST are better.

	Synthetic SBM	CiteSeer	Cora	PubMed	QM9
Accuracy, mean ± standard deviation					
2	0.879 ± 0.127	0.579 ± 0.030	0.660 ± 0.030	0.601 ± 0.061	0.826 ± 0.038
4	0.932 ± 0.012	0.607 ± 0.025	0.681 ± 0.030	0.654 ± 0.037	0.834 ± 0.030
8 (primary grid)	0.924 ± 0.028	0.616 ± 0.024	0.683 ± 0.033	0.671 ± 0.047	0.840 ± 0.034
16	0.930 ± 0.013	0.615 ± 0.022	0.694 ± 0.024	0.703 ± 0.026	0.841 ± 0.032
32	0.922 ± 0.019	0.611 ± 0.021	0.697 ± 0.021	0.712 ± 0.020	0.843 ± 0.020
Generation MMD, refit prior, mean ± standard deviation					
2	0.007 ± 0.006	0.016 ± 0.007	0.055 ± 0.013	0.139 ± 0.051	0.268 ± 0.019
4	0.008 ± 0.006	0.017 ± 0.010	0.039 ± 0.007	0.106 ± 0.031	0.276 ± 0.016
8 (primary grid)	0.006 ± 0.003	0.012 ± 0.005	0.036 ± 0.006	0.075 ± 0.014	0.269 ± 0.021
16	0.007 ± 0.008	0.011 ± 0.006	0.031 ± 0.005	0.060 ± 0.011	0.285 ± 0.023
32	0.008 ± 0.004	0.013 ± 0.005	0.037 ± 0.015	0.091 ± 0.025	0.274 ± 0.017
Generation C2ST, refit prior, mean ± standard deviation					
2	0.551 ± 0.037	0.753 ± 0.022	0.862 ± 0.023	0.859 ± 0.045	0.957 ± 0.051
4	0.557 ± 0.046	0.753 ± 0.030	0.829 ± 0.020	0.844 ± 0.016	0.975 ± 0.020
8 (primary grid)	0.538 ± 0.041	0.716 ± 0.038	0.816 ± 0.026	0.806 ± 0.025	0.963 ± 0.023
16	0.567 ± 0.050	0.705 ± 0.040	0.814 ± 0.034	0.819 ± 0.033	0.915 ± 0.158
32	0.568 ± 0.041	0.691 ± 0.029	0.789 ± 0.036	0.809 ± 0.023	0.947 ± 0.092
Round-trip MMD, mean ± standard deviation					
2	0.003 ± 0.004	0.017 ± 0.007	0.061 ± 0.012	0.131 ± 0.038	0.277 ± 0.018
4	0.002 ± 0.002	0.021 ± 0.008	0.040 ± 0.013	0.106 ± 0.031	0.267 ± 0.020
8 (primary grid)	0.003 ± 0.004	0.017 ± 0.011	0.034 ± 0.006	0.074 ± 0.036	0.273 ± 0.020
16	0.004 ± 0.005	0.011 ± 0.005	0.033 ± 0.005	0.054 ± 0.017	0.269 ± 0.028
32	0.004 ± 0.004	0.011 ± 0.010	0.032 ± 0.006	0.096 ± 0.047	0.258 ± 0.012
Accuracy, d_z minus d_z=8, paired Δ [95% CI]					
2	-0.045 [-0.129, $+0.009$]	**-0.038** [-0.060, -0.016]	**-0.023** [-0.038, -0.010]	**-0.070** [-0.100, -0.039]	**-0.014** [-0.027, -0.003]
4	$+0.008$ [-0.007, $+0.028$]	-0.010 [-0.027, $+0.006$]	-0.002 [-0.026, $+0.024$]	-0.017 [-0.051, $+0.025$]	-0.006 [-0.020, $+0.007$]
16	$+0.006$ [-0.006, $+0.021$]	-0.002 [-0.014, $+0.011$]	$+0.011$ [-0.013, $+0.035$]	**$+0.031$** [$+0.001$, $+0.071$]	$+0.001$ [-0.009, $+0.010$]
32	-0.001 [-0.014, $+0.011$]	-0.006 [-0.025, $+0.013$]	$+0.014$ [-0.002, $+0.032$]	**$+0.041$** [$+0.014$, $+0.072$]	$+0.003$ [-0.010, $+0.015$]
Generation C2ST, d_z minus d_z=8, paired Δ [95% CI]					
2	$+0.012$ [-0.022, $+0.052$]	**$+0.036$** [$+0.015$, $+0.058$]	**$+0.045$** [$+0.025$, $+0.064$]	**$+0.053$** [$+0.020$, $+0.085$]	-0.006 [-0.047, $+0.023$]
4	$+0.018$ [-0.015, $+0.048$]	**$+0.036$** [$+0.015$, $+0.057$]	$+0.013$ [-0.005, $+0.031$]	**$+0.038$** [$+0.025$, $+0.053$]	$+0.013$ [-0.003, $+0.027$]
16	$+0.028$ [-0.001, $+0.056$]	-0.011 [-0.032, $+0.011$]	-0.002 [-0.025, $+0.021$]	$+0.013$ [-0.006, $+0.030$]	-0.047 [-0.162, $+0.031$]
32	**$+0.029$** [$+0.007$, $+0.050$]	**-0.025** [-0.045, -0.005]	-0.027 [-0.059, $+0.006$]	$+0.003$ [-0.013, $+0.019$]	-0.016 [-0.085, $+0.025$]

With the update count equalized, SEQ-REG's accuracy advantage over single-phase joint training is no longer resolved on any of the five domains, and neither is its descriptor-MMD advantage. Its advantage on the learned metric survives: the paired C2ST difference remains significantly below zero on Synthetic, CiteSeer, Cora, and PubMed (-0.049 [-0.079, -0.018], -0.016 [-0.031, -0.002], -0.015 [-0.027, -0.003], and -0.028 [-0.048, -0.009]) and is a tie on QM9. The update-count confound therefore accounts for the comprehension and descriptor-level gaps between the two-phase and joint schedules, but not for the whole discriminator-level gap. Against the matched single-task references, SEQ-REG keeps a significant accuracy advantage over C-ONLY on Synthetic, CiteSeer, and QM9, while G-ONLY at the matched budget reaches significantly lower descriptor MMD than

Table A.12: Per-dataset Pareto-front membership over accuracy and either generation metric, under the dominance rule of Appendix D.5. The first and third rows of each panel cover the eight-condition grid under the refit and cached-prior deployments, where SEQ-REG keeps the refit prior in both and ^{c} marks the cached-prior deployment of a frozen-decoder condition. The second row adds SEQ-ANCHOR$_1$, SEQ-ANCHOR at λ_{anc}=1 (Appendix D.4), to the refit grid, and Fig. 2 outlines this front. The fourth row adds the update-matched controls of Tables A.15 and A.16 under the refit deployment: JOINT-RETAINED^{m} is JOINT-RETAINED at $S(1+B)$ updates with the summed comprehension loss, JOINT-RETAINED^{a} the same with the loss averaged over the B mini-batches (domains with $B>1$), and C-ONLY^{m} is C-ONLY at $S(1+B)$ updates.

	Synthetic SBM	CiteSeer	Cora	PubMed	QM9
Accuracy and gen-MMD					
Grid, refit	SEQ-REG	SEQ-REG	SEQ-REG	SEQ-REG	SEQ-REG JOINT-RETAINED
Grid and SEQ-ANCHOR $_1$, refit	SEQ-REG	SEQ-REG	SEQ-REG SEQ-ANCHOR$_1$	SEQ-REG	SEQ-REG JOINT-RETAINED
Grid, cached prior	SEQ-REG	SEQ-REG	SEQ-REG	SEQ-REG	SEQ-REG JOINT-RETAINED SEQ-FREEZE^{c}
With matched controls	SEQ-REG JOINT-RETAINED^{m}	JOINT-RETAINED^{a}	SEQ-REG JOINT-RETAINED^{a}	SEQ-REG JOINT-RETAINED^{m}	SEQ-REG JOINT-RETAINED^{m} JOINT-RETAINED^{a}
Accuracy and C2ST					
Grid, refit	SEQ-REG JOINT-RETAINED	SEQ-REG JOINT-RETAINED	SEQ-REG	SEQ-REG JOINT-RETAINED	SEQ-REG JOINT-RETAINED
Grid and SEQ-ANCHOR $_1$, refit	SEQ-REG SEQ-ANCHOR$_1$	SEQ-REG JOINT-RETAINED	SEQ-ANCHOR$_1$	SEQ-REG JOINT-RETAINED SEQ-ANCHOR$_1$	SEQ-REG JOINT-RETAINED
Grid, cached prior	SEQ-REG JOINT-RETAINED SEQ-ANCHOR^{c}	SEQ-REG JOINT-RETAINED	SEQ-REG SEQ-FREEZE^{c} SEQ-ANCHOR^{c}	SEQ-REG JOINT-RETAINED	SEQ-REG JOINT-RETAINED SEQ-FREEZE^{c}
With matched controls	SEQ-REG JOINT-RETAINED	SEQ-REG JOINT-RETAINED^{a}	SEQ-REG JOINT-RETAINED^{a}	SEQ-REG JOINT-RETAINED	SEQ-REG JOINT-RETAINED^{m} JOINT-RETAINED^{a}

Table A.13: Aggregate SEQ-REG against each single-model condition under refit deployment. W/T/L counts datasets where the paired-bootstrap 95% CI favors SEQ-REG, overlaps 0, or favors the opponent. † marks no-generator conditions placed at $+\infty$.

Opponent	Comprehension W/T/L	Generation W/T/L
C-ONLY	3/2/0	5/0/0†
MAE-DISCARD	2/3/0	5/0/0†
SEQ-FREEZE	2/3/0	5/0/0
SEQ-ANCHOR	2/3/0	5/0/0
NAIVE-DECOUPLED	2/3/0	5/0/0
JOINT-RETAINED	1/4/0	4/1/0

SEQ-REG on CiteSeer, Cora, and QM9, by 0.003, 0.016, and 0.033. Tables 1 and 2 report G-ONLY at this matched budget as G-ONLY ^{m}, and Table A.18 adds its paired difference to the null. In C2ST, SEQ-REG minus G-ONLY ^{m} is significant only on Cora, $+0.064$ $[+0.044, +0.085]$. Under the cached prior, the frozen conditions exceed G-ONLY ^{m} in descriptor MMD on every domain, by 0.005 on Synthetic to 0.054 on QM9, and in C2ST on Cora and PubMed, by 0.077 and 0.025.

Per-update loss weighting. Matching the update count does not match the per-update objective when $B>1$. A single-phase update sums the B labeled mini-batch losses before adding $\lambda\mathcal{L}_{\text{gen}}$ (Appendix B), whereas a Phase-2 update of SEQ-REG uses one mini-batch, so on Cora, CiteSeer, and QM9 the matched control weights the comprehension loss B times more heavily relative to $\mathcal{L}_{\text{gen}}$. We therefore rerun the matched control on these three domains with the B comprehension losses averaged rather than summed, keeping the batch order and random-number use of the summed run. Where $B=1$ the two runs coincide, which we verified field by field on two Synthetic seeds. Against this control (Table A.15), the C2ST advantage of SEQ-REG remains on

Table A.14: Optimizer-update counts read from instrumented training loops. n_{lab} is the labeled-set size, $B = \lceil n_{\text{lab}}/64 \rceil$ the number of labeled mini-batches per Phase-2 outer iteration (the last partial batch is kept), and S the per-phase budget. Two-phase conditions take S Phase-1 and SB Phase-2 updates, single-phase conditions take S, and the update-matched controls of Table A.15 take $S(1+B)$. The exposure columns count the pool and labeled graphs drawn during training, and labeled exposure is the same for SEQ-REG and JOINT-RETAINED.

				two-phase updates		pool graphs		labeled
Dataset	n_{lab}	B	S	Phase 2	total	SEQ-REG	JOINT-RETAINED	graphs
Synthetic SBM	32	1	500	500	1,000	64,000	32,000	16,000
CiteSeer	120	2	250	500	750	48,000	16,000	30,000
Cora	140	3	250	750	1,000	64,000	16,000	35,000
PubMed	60	1	250	250	500	32,000	16,000	15,000
QM9	128	2	300	600	900	57,600	19,200	38,400

Table A.15: Update-matched single-phase control: JOINT-RETAINED at the standard budget of S updates and at $S(1+B)$, the two-phase total of SEQ-REG on every domain (Table A.14). On the domains with $B > 1$, the averaged row divides the summed comprehension loss by B, so that each update weights $\mathcal{L}_{\text{comp}}$ against $\lambda\mathcal{L}_{\text{gen}}$ as a Phase-2 update of SEQ-REG does. Generation MMD and C2ST use the refit deployment.

Condition	Updates	Acc.	Gen-MMD	C2ST
Synthetic SBM				
JOINT-RETAINED	S	0.919 ± 0.016	0.012 ± 0.009	0.569 ± 0.046
JOINT-RETAINED	$S(1+B)$	0.922 ± 0.015	0.009 ± 0.008	0.588 ± 0.038
SEQ-REG	$S(1+B)$	0.924 ± 0.028	0.006 ± 0.003	0.538 ± 0.041
CiteSeer				
JOINT-RETAINED	S	0.602 ± 0.030	0.020 ± 0.010	0.742 ± 0.038
JOINT-RETAINED	$S(1+B)$	0.613 ± 0.024	0.014 ± 0.006	0.733 ± 0.023
JOINT-RETAINED, averaged	$S(1+B)$	0.630 ± 0.022	0.014 ± 0.008	0.746 ± 0.025
SEQ-REG	$S(1+B)$	0.616 ± 0.024	0.012 ± 0.005	0.716 ± 0.038
Cora				
JOINT-RETAINED	S	0.656 ± 0.030	0.056 ± 0.014	0.849 ± 0.031
JOINT-RETAINED	$S(1+B)$	0.676 ± 0.027	0.044 ± 0.024	0.832 ± 0.030
JOINT-RETAINED, averaged	$S(1+B)$	0.698 ± 0.032	0.037 ± 0.011	0.823 ± 0.029
SEQ-REG	$S(1+B)$	0.683 ± 0.033	0.036 ± 0.006	0.816 ± 0.026
PubMed				
JOINT-RETAINED	S	0.689 ± 0.030	0.104 ± 0.018	0.828 ± 0.058
JOINT-RETAINED	$S(1+B)$	0.696 ± 0.028	0.083 ± 0.029	0.834 ± 0.031
SEQ-REG	$S(1+B)$	0.671 ± 0.047	0.075 ± 0.014	0.806 ± 0.025
QM9				
JOINT-RETAINED	S	0.830 ± 0.037	0.286 ± 0.031	0.967 ± 0.035
JOINT-RETAINED	$S(1+B)$	0.840 ± 0.033	0.277 ± 0.021	0.929 ± 0.145
JOINT-RETAINED, averaged	$S(1+B)$	0.835 ± 0.036	0.273 ± 0.014	0.949 ± 0.063
SEQ-REG	$S(1+B)$	0.840 ± 0.034	0.269 ± 0.021	0.963 ± 0.023

CiteSeer, -0.030 $[-0.054, -0.005]$, but not on Cora, -0.007 $[-0.027, +0.016]$, and QM9 stays a tie, $+0.014$ $[-0.020, +0.058]$, so on Cora the residual gap of the summed control reflects the loss weighting. Averaging also raises the accuracy of the single-phase control on Cora and CiteSeer, by 0.021 $[+0.003, +0.039]$ and 0.018 $[+0.004, +0.031]$, and on CiteSeer the averaged control is more accurate than SEQ-REG, with SEQ-REG minus control -0.014 $[-0.029, -0.0002]$. Descriptor MMD stays a tie on all three domains. With the update count and the per-update loss weight both matched, SEQ-REG therefore keeps a discriminator-level advantage over single-phase joint training on Synthetic, CiteSeer, and PubMed and no comprehension or descriptor-level advantage on any domain. The two-phase schedule also yields a Phase-1 encoder and decoder before any labeled set is seen, which no single-phase budget provides.

Table A.16: Update-matched single-task references: C-ONLY and G-ONLY at the standard S updates and at the two-phase total $S(1+B)$. The SEQ-REG values are those of Table A.15. Generation MMD and C2ST use the refit deployment.

	C-ONLY accuracy		G-ONLY gen-MMD		G-ONLY C2ST	
Dataset	S	$S(1+B)$	S	$S(1+B)$	S	$S(1+B)$
Synthetic SBM	0.794 ± 0.033	0.796 ± 0.034	0.004 ± 0.002	0.004 ± 0.002	0.555 ± 0.041	0.560 ± 0.040
CiteSeer	0.524 ± 0.038	0.534 ± 0.031	0.012 ± 0.004	0.009 ± 0.003	0.723 ± 0.023	0.719 ± 0.038
Cora	0.659 ± 0.039	0.666 ± 0.041	0.038 ± 0.006	0.020 ± 0.003	0.838 ± 0.028	0.753 ± 0.031
PubMed	0.646 ± 0.028	0.647 ± 0.029	0.076 ± 0.016	0.068 ± 0.015	0.829 ± 0.022	0.811 ± 0.038
QM9	0.825 ± 0.043	0.823 ± 0.043	0.242 ± 0.014	0.236 ± 0.015	0.931 ± 0.096	0.972 ± 0.017

Table A.17: ogbn-arxiv scale study for representative conditions and the density-matched Erdős-Rényi null, with $N{=}64$ ego-graphs and 10 paired seeds. Gen. MMD uses a degree histogram over $[0, 64]$ with 32 bins, the bin width of the primary grid, and is computed on checkpoints retrained for this evaluation, whereas accuracy and C2ST are the archived values. N/A marks a condition that deploys no decoder (C-ONLY) or has no comprehension head (G-ONLY and the null). MAE-DISCARD shares the Phase-2 training path of SEQ-FREEZE and is not listed separately.

Condition	Acc. ↑	Gen. MMD ↓	Gen. C2ST ↓
SEQ-REG	0.310 ± 0.052	0.162 ± 0.017	0.857 ± 0.031
C-ONLY	0.275 ± 0.042	N/A	N/A
SEQ-FREEZE	0.286 ± 0.062	0.605 ± 0.128	0.929 ± 0.067
JOINT-RETAINED	0.284 ± 0.048	0.463 ± 0.089	0.928 ± 0.030
G-ONLY	N/A	0.158 ± 0.023	0.853 ± 0.046
Erdős-Rényi	N/A	0.312 ± 0.013	0.929 ± 0.012

D.7 OGBN-ARXIV COMPACT FEW-SHOT REGIME

This subsection extends the controlled comparison to an OGB-scale graph with the same encoder and decoder families and the same two-phase protocol. It is the single ogbn-arxiv study of the paper and uses the 10 paired seeds of the primary grid.

Setup. We extend to ogbn-arxiv (Hu et al., 2020a), a graph with 169K nodes, 1.17M edges, 128 raw features, and 40 classes.

- Each labeled node defines a 2-hop ego-graph subsampled to $N{=}64$ nodes. The 128-d features are SVD-projected to FDIM=32 with a projection fit, without labels, on 20,000 nodes drawn uniformly from the whole graph, and then standardized.
- $K{=}8$ labeled centers per class, a stricter per-class budget than the $K{=}20$ of the citation domains. With 40 classes the labeled set is already 320 ego-graphs at $K{=}8$.
- For the larger graphs the latent has 16 dimensions, and training uses learning rate 10^{-3}, batch size 32, and 40 Phase-1 and 30 Phase-2 epochs with best-validation checkpoints, for every condition.
- Labeled centers come from OGB's training split, and the unlabeled pool of 2000 centers from the rest of that split. The validation set consists of 1000 ego-graphs from OGB's validation split, and the test set of 1000 ego-graphs from OGB's official test split per seed. The run uses 400 generated graphs per condition and seed and the same descriptor-MMD, C2ST, and paired-bootstrap pipeline as the primary grid, with the degree histogram extended to $[0, 64]$ (Appendix C).

Results in Table A.17. The comprehension-generation pattern persists at this scale. SEQ-REG leads on accuracy and stays close to G-ONLY on both generation metrics, whereas the frozen and joint conditions carry about three to four times its gen-MMD and score C2ST 0.929 and 0.928, against 0.929 for the density-matched null. SEQ-REG improves on the null by 0.07 in mean C2ST (-0.071 with CI $[-0.090, -0.054]$), comparable to the citation domains, and remains far above the chance level of 0.5. On comprehension, the paired CI of its margin over JOINT-RETAINED includes zero ($+0.026$, $[-0.001, +0.054]$), whereas its margins over SEQ-FREEZE ($+0.025$, $[+0.007, +0.044]$)

and C-ONLY ($+0.036$, $[+0.019, +0.052]$) exclude zero. Every paired contrast in this subsection keeps its sign and significance when restricted to the nine seeds on which the retrained checkpoints of the Gen. MMD column reproduce the archived accuracy and C2ST. The scale study therefore supports the generation-side result more strongly than the comprehension-side one.

At this scale, a comprehension gain over C-ONLY appears only in the two-phase condition that keeps the generative loss active. SEQ-FREEZE, whose Phase-2 training path is also that of MAE-DISCARD, is not significantly above C-ONLY, $+0.011$ with CI $[-0.014, +0.037]$, and neither is JOINT-RETAINED, $+0.009$ with CI $[-0.013, +0.030]$. On Synthetic SBM the gain of JOINT-RETAINED over C-ONLY shrinks as the label budget grows from $K{=}8$ to 32 (Fig. A.1), so the comprehension gain that survives at the few-shot budget of ogbn-arxiv belongs to the two-phase condition with an active generative loss.

D.8 AUXILIARY ROBUSTNESS CHECKS

Decoupled readouts. NAIVE-DECOUPLED transposes the encoder-side pathway decoupling of Janus (Wu et al., 2025) to the readout side, using i.i.d. node sampling at inference. It is comprehension-competitive, with W/T/L 2/3/0 in SEQ-REG's favor. The largest accuracy gap is on Synthetic, where SEQ-REG reaches 0.924 and NAIVE-DECOUPLED reaches 0.772. On generation, NAIVE-DECOUPLED has substantially higher MMD, 0.29 to 1.19. Both gaps are consistent with a role for the shared $\mathbf{z}$ and for joint node-set sampling, but NAIVE-DECOUPLED changes the bottleneck, the decoder inputs, and the node sampler together, so it does not attribute either gap to one of them.

Learned metric against the null. Table A.18 reports, per domain, the C2ST of the density-matched null, the real-versus-real floor, a label-shuffle check, and the paired difference of every grid condition against the null. The active-loss conditions and G-ONLY are significantly less separable than the null on Synthetic and the three citation domains. The frozen and decoupled conditions clear it only on Cora and PubMed, and on QM9 only JOINT-RETAINED clears it, by less than 0.01. The floor lies between 0.48 and 0.52 on every domain. Against it, every condition outside Synthetic lies at least 0.21 above it. Table A.19 repeats the comparison with each condition scored at the floor's sample size of 200 graphs per side. Every condition outside Synthetic remains significantly above the floor, by at least 0.199, and halving the sample size changes a condition's mean C2ST by at most 0.041.

Cached prior under the learned metric. Table A.20 scores the cached-prior deployment of Table 2 with the discriminator. Reusing the Phase-1 prior lowers C2ST significantly on all five domains for both frozen conditions, by 0.29 on Synthetic down to 0.04 on QM9 for SEQ-FREEZE. The deployment result of RQ2 is therefore not an artifact of descriptor MMD. The cached prior is a rule for a decoder frozen at Phase 1: for SEQ-REG, whose decoder co-adapts in Phase 2, the cached prior raises descriptor MMD over the refit prior on every domain, significantly on Synthetic, $+0.029$ $[+0.013, +0.047]$, and CiteSeer, $+0.010$ $[+0.002, +0.019]$, and raises C2ST significantly on Synthetic, $+0.067$ $[+0.030, +0.102]$.

Generator family on a matched schedule. Table A.21 trains the VAE on the two-phase schedule of SEQ-REG on all five domains, with the decoder frozen (VAE-freeze) or trained under $\mathcal{L}_{\text{comp}} + \lambda\,\text{ELBO}$ (VAE-reg, the VAE of Table 6), so its differences from SEQ-REG do not mix the variational objective with the schedule. Under its own fixed prior, the C2ST of VAE-reg differs from SEQ-REG (Table A.22) by $+0.100$ $[+0.065, +0.137]$ on Synthetic, -0.035 $[-0.068, -0.005]$ on CiteSeer, and $+0.025$ $[+0.011, +0.041]$ on QM9. The last two columns of Table A.22 compare these family effects with the prior-deployment effect of the same domain, the refit-minus-cached C2ST of SEQ-FREEZE. The deployment effect is significantly larger than the VAE effect on Synthetic, CiteSeer, and Cora, by 0.186, 0.134, and 0.060, and the same pattern holds for the diffusion prior. Testing the deployment effect minus and plus the family effect separately gives the same significant domains. Inside the VAE family, refitting the prior of VAE-freeze raises its C2ST significantly on all five domains, by 0.045 on Cora to 0.253 on Synthetic. This matches the direction of the prior-deployment effect but is not solely attributable to encoder drift, because the refit is a Gaussian on the posterior means that omits the average posterior variance (Table A.21). For VAE-reg, whose decoder is trained in Phase 2, neither prior is uniformly better: refit minus fixed C2ST is -0.092 $[-0.125, -0.056]$ on Synthetic and -0.013 $[-0.020, -0.006]$ on QM9, $+0.030$ $[+0.000, +0.061]$ on CiteSeer, where descriptor MMD favors the refit instead. The cached-versus-refit rule of the main

Table A.18: Generation C2ST against the density-matched null. The upper panel gives the C2ST of the Erdős-Rényi null, of the real-versus-real floor, which scores two disjoint halves of 200 reference graphs against each other, and of a label-shuffle check that pools real and null graphs and re-splits them at random. Each condition is scored with 400 graphs per side, and Table A.19 repeats the floor comparison at 200. G-ONLY ^{m} is G-ONLY at the update-matched budget (Table A.16), and SEQ-ANCHOR $_{1}$ is SEQ-ANCHOR at λ_{anc}=1 (Table A.8). The middle panel gives each condition's paired difference to the null, and bold marks a CI entirely below zero, that is, a condition significantly less separable from real graphs than the null. The bottom panel gives the paired difference to the floor on the same seeds, where a CI that excludes zero means the condition is significantly more separable than two halves of the real set.

	Synthetic SBM	CiteSeer	Cora	PubMed	QM9
C2ST of the reference sets					
Erdős-Rényi null	0.772 ± 0.038	0.908 ± 0.029	0.938 ± 0.027	0.937 ± 0.014	0.975 ± 0.040
Real vs. real	0.516 ± 0.044	0.502 ± 0.032	0.483 ± 0.039	0.486 ± 0.041	0.500 ± 0.044
Label shuffle	0.517 ± 0.022	0.494 ± 0.053	0.520 ± 0.060	0.501 ± 0.047	0.477 ± 0.039
Condition minus null, paired Δ C2ST [95% CI]					
G-ONLY	**−0.217** [−0.242, −0.192]	**−0.185** [−0.207, −0.164]	**−0.100** [−0.123, −0.081]	**−0.108** [−0.123, −0.089]	−0.043 [−0.114, +0.015]
G-ONLY ^{m}	**−0.212** [−0.253, −0.177]	**−0.189** [−0.218, −0.160]	**−0.185** [−0.212, −0.155]	**−0.126** [−0.149, −0.103]	−0.003 [−0.024, +0.027]
SEQ-REG	**−0.234** [−0.261, −0.206]	**−0.192** [−0.216, −0.171]	**−0.122** [−0.144, −0.099]	**−0.131** [−0.147, −0.116]	−0.012 [−0.036, +0.020]
JOINT-RETAINED	**−0.203** [−0.247, −0.160]	**−0.166** [−0.198, −0.132]	**−0.089** [−0.114, −0.064]	**−0.110** [−0.145, −0.085]	**−0.008** [−0.015, −0.001]
SEQ-ANCHOR	+0.030 [−0.053, +0.110]	−0.017 [−0.042, +0.005]	**−0.029** [−0.053, −0.004]	**−0.061** [−0.089, −0.033]	+0.024 [+0.008, +0.051]
SEQ-ANCHOR $_{1}$	**−0.203** [−0.243, −0.162]	**−0.152** [−0.170, −0.132]	**−0.120** [−0.150, −0.086]	**−0.120** [−0.148, −0.092]	−0.000 [−0.008, +0.008]
SEQ-FREEZE	+0.065 [+0.002, +0.128]	−0.002 [−0.030, +0.023]	**−0.039** [−0.071, −0.010]	**−0.036** [−0.072, −0.004]	+0.024 [+0.008, +0.051]
NAIVE-DECOUPLED	+0.209 [+0.166, +0.244]	+0.034 [+0.019, +0.053]	**−0.032** [−0.056, −0.005]	**−0.053** [−0.082, −0.022]	+0.025 [+0.010, +0.049]
Condition minus real-versus-real floor, paired Δ C2ST [95% CI]					
G-ONLY	+0.039 [+0.001, +0.073]	+0.222 [+0.192, +0.247]	+0.354 [+0.322, +0.381]	+0.343 [+0.320, +0.365]	+0.431 [+0.369, +0.478]
SEQ-REG	+0.022 [−0.013, +0.065]	+0.215 [+0.178, +0.247]	+0.333 [+0.315, +0.350]	+0.320 [+0.298, +0.342]	+0.463 [+0.429, +0.497]
JOINT-RETAINED	+0.053 [+0.015, +0.089]	+0.240 [+0.207, +0.271]	+0.365 [+0.330, +0.396]	+0.342 [+0.307, +0.372]	+0.467 [+0.431, +0.500]
SEQ-ANCHOR	+0.287 [+0.196, +0.369]	+0.390 [+0.363, +0.417]	+0.426 [+0.395, +0.456]	+0.390 [+0.360, +0.422]	+0.499 [+0.473, +0.522]
SEQ-FREEZE	+0.321 [+0.247, +0.389]	+0.404 [+0.375, +0.433]	+0.416 [+0.375, +0.452]	+0.415 [+0.379, +0.450]	+0.499 [+0.472, +0.523]
NAIVE-DECOUPLED	+0.465 [+0.424, +0.503]	+0.440 [+0.425, +0.455]	+0.422 [+0.398, +0.444]	+0.398 [+0.371, +0.426]	+0.499 [+0.473, +0.524]
G-ONLY ^{m}	+0.044 [+0.004, +0.087]	+0.217 [+0.183, +0.247]	+0.269 [+0.243, +0.297]	+0.325 [+0.293, +0.355]	+0.472 [+0.444, +0.498]
SEQ-ANCHOR $_{1}$	+0.053 [+0.022, +0.090]	+0.254 [+0.227, +0.280]	+0.335 [+0.298, +0.368]	+0.331 [+0.297, +0.368]	+0.475 [+0.440, +0.507]

text is therefore stated for the deterministic autoencoder, whose prior is fitted to its encoded latents. VAE-reg round-trips significantly better than VAE-freeze on every domain under both round-trip MMD and round-trip C2ST. Paired differences use the bootstrap of Appendix C on the same 10 seeds.

Single-GPU external generator. Table A.5 re-scores DiGress on QM9 after matching the graph representation. Trained for 62 minutes on one GPU, DiGress reaches C2ST 0.502 on QM9 against 0.490 for real versus real, whereas SEQ-REG reaches 0.963. The dual-use model therefore does not

Table A.19: Generation C2ST at the sample size of the real-versus-real floor. For each seed, the discriminator of Table A.18 separates the first 200 reference graphs, the first half of the floor, from the first 200 refit-deployment samples scored in Table 1. The upper panel gives the paired difference to the floor and the lower panel the difference from the same checkpoint's C2ST at 400 graphs per side, with bold marking a CI that excludes zero. Checkpoints were retrained for this evaluation and reproduce the stored grid except on PubMed seeds 1 and 7, without which every PubMed difference to the floor remains significant.

	Synthetic SBM	CiteSeer	Cora	PubMed	QM9
C2ST at 200 graphs per side minus the real-versus-real floor, paired Δ [95% CI]					
SEQ-REG	+0.025 [−0.026, +0.075]	**+0.199** [+0.165, +0.232]	**+0.323** [+0.306, +0.343]	**+0.312** [+0.281, +0.342]	**+0.447** [+0.400, +0.487]
JOINT-RETAINED	+0.042 [−0.003, +0.087]	**+0.242** [+0.217, +0.268]	**+0.368** [+0.324, +0.410]	**+0.361** [+0.332, +0.390]	**+0.428** [+0.323, +0.497]
SEQ-FREEZE	**+0.299** [+0.219, +0.376]	**+0.407** [+0.381, +0.436]	**+0.396** [+0.350, +0.444]	**+0.405** [+0.363, +0.442]	**+0.498** [+0.472, +0.522]
SEQ-ANCHOR	**+0.285** [+0.205, +0.358]	**+0.397** [+0.368, +0.424]	**+0.385** [+0.332, +0.438]	**+0.372** [+0.326, +0.418]	**+0.496** [+0.471, +0.520]
NAIVE-DECOUPLED	**+0.484** [+0.460, +0.510]	**+0.445** [+0.425, +0.464]	**+0.427** [+0.405, +0.453]	**+0.367** [+0.304, +0.419]	**+0.498** [+0.472, +0.523]
G-ONLY	+0.027 [−0.008, +0.058]	**+0.242** [+0.213, +0.277]	**+0.328** [+0.304, +0.352]	**+0.329** [+0.295, +0.360]	**+0.472** [+0.433, +0.507]
Erdős-Rényi null	**+0.208** [+0.142, +0.263]	**+0.412** [+0.397, +0.426]	**+0.457** [+0.436, +0.477]	**+0.467** [+0.435, +0.493]	**+0.474** [+0.442, +0.503]
C2ST at 200 minus C2ST at 400 graphs per side, same checkpoint, paired Δ [95% CI]					
SEQ-REG	+0.002 [−0.033, +0.044]	−0.015 [−0.045, +0.014]	−0.010 [−0.035, +0.013]	−0.012 [−0.036, +0.013]	−0.016 [−0.060, +0.017]
JOINT-RETAINED	−0.012 [−0.040, +0.017]	+0.002 [−0.026, +0.036]	+0.003 [−0.032, +0.035]	+0.019 [−0.010, +0.050]	−0.039 [−0.139, +0.019]
SEQ-FREEZE	−0.022 [−0.057, +0.013]	+0.004 [−0.009, +0.016]	−0.020 [−0.043, +0.002]	−0.009 [−0.028, +0.011]	−0.001 [−0.003, +0.000]
SEQ-ANCHOR	−0.002 [−0.087, +0.090]	+0.008 [−0.010, +0.028]	**−0.041** [−0.080, −0.009]	−0.017 [−0.055, +0.017]	−0.003 [−0.008, +0.001]
NAIVE-DECOUPLED	+0.019 [+0.000, +0.055]	+0.005 [−0.009, +0.021]	+0.005 [−0.018, +0.023]	−0.025 [−0.092, +0.020]	−0.001 [−0.002, +0.000]
G-ONLY	−0.013 [−0.037, +0.011]	+0.021 [−0.008, +0.046]	**−0.026** [−0.046, −0.005]	−0.017 [−0.036, +0.000]	+0.041 [−0.016, +0.110]

Table A.20: Generation C2ST of the two frozen-decoder conditions under the refit and cached-prior deployments, on the checkpoints and samples of Table 2, with the standard-budget G-ONLY as the generation-only reference. One column reports the cached prior for both conditions, whose cached generators coincide (Table 2).

	refit prior		cached prior	
Dataset	SEQ-FREEZE	SEQ-ANCHOR	SEQ-FREEZE, SEQ-ANCHOR	G-ONLY
Synthetic SBM	0.837 ± 0.123	0.802 ± 0.143	0.550 ± 0.028	0.555 ± 0.041
Cora	0.899 ± 0.041	0.909 ± 0.025	0.830 ± 0.027	0.838 ± 0.028
CiteSeer	0.905 ± 0.046	0.891 ± 0.047	0.736 ± 0.026	0.723 ± 0.023
PubMed	0.901 ± 0.059	0.876 ± 0.041	0.836 ± 0.035	0.829 ± 0.022
QM9	0.999 ± 0.003	0.999 ± 0.003	0.964 ± 0.042	0.931 ± 0.096

match a dedicated molecular generator on structure. What it provides instead is comprehension and generation from one checkpoint.

Family-wise statistics. The contrasts of SEQ-REG against the six other grid conditions across five datasets and both tasks form a family of 50 paired comparisons under the refit deployment, excluding the 10 generator-versus-no-generator cells. Under Holm-Bonferroni correction at family-wise α=0.05, 27 of 31 uncorrected wins remain significant on bootstrap p-values, and Holm correction on the more conservative paired t-test p-values retains 20 of 29. The four bootstrap Holm-sensitive

Table A.21: VAE family on the two-phase schedule of SEQ-REG (Appendix C), with the decoder frozen under $\mathcal{L}_{\text{comp}}$ alone (VAE-freeze) or trained under $\mathcal{L}_{\text{comp}} + \lambda$ ELBO (VAE-reg, the VAE of Table 6). Generation C2ST is given under the fixed $\mathcal{N}(0, I)$ prior and under a Gaussian refit to the Phase-2 posterior means of the pool graphs, the VAE counterpart of the refit deployment. Round-trip C2ST follows Table A.24. The SEQ-REG rows repeat Tables 3, 1, 5, and A.24.

		C2ST		Round-trip	
Condition	Acc.	fixed prior	refit prior	MMD	C2ST
Synthetic SBM					
VAE-freeze	0.840 ± 0.030	0.599 ± 0.042	0.852 ± 0.136	0.485 ± 0.294	0.861 ± 0.097
VAE-reg	0.894 ± 0.068	0.639 ± 0.059	0.547 ± 0.034	0.005 ± 0.003	0.532 ± 0.041
SEQ-REG	0.924 ± 0.028	N/A	0.538 ± 0.041	0.003 ± 0.004	0.534 ± 0.021
CiteSeer					
VAE-freeze	0.564 ± 0.036	0.720 ± 0.030	0.906 ± 0.034	0.295 ± 0.094	0.872 ± 0.046
VAE-reg	0.612 ± 0.030	0.681 ± 0.056	0.711 ± 0.032	0.015 ± 0.008	0.715 ± 0.019
SEQ-REG	0.616 ± 0.024	N/A	0.716 ± 0.038	0.017 ± 0.011	0.715 ± 0.027
Cora					
VAE-freeze	0.675 ± 0.017	0.843 ± 0.026	0.887 ± 0.040	0.317 ± 0.163	0.895 ± 0.030
VAE-reg	0.689 ± 0.033	0.825 ± 0.041	0.809 ± 0.035	0.035 ± 0.007	0.803 ± 0.020
SEQ-REG	0.683 ± 0.033	N/A	0.816 ± 0.026	0.034 ± 0.006	0.808 ± 0.028
PubMed					
VAE-freeze	0.672 ± 0.029	0.850 ± 0.068	0.912 ± 0.041	0.354 ± 0.147	0.892 ± 0.075
VAE-reg	0.665 ± 0.030	0.820 ± 0.049	0.805 ± 0.031	0.081 ± 0.062	0.832 ± 0.027
SEQ-REG	0.671 ± 0.047	N/A	0.806 ± 0.025	0.074 ± 0.036	0.832 ± 0.023
QM9					
VAE-freeze	0.836 ± 0.040	0.949 ± 0.117	0.997 ± 0.006	0.413 ± 0.138	0.996 ± 0.007
VAE-reg	0.839 ± 0.030	0.988 ± 0.005	0.975 ± 0.011	0.268 ± 0.017	0.978 ± 0.027
SEQ-REG	0.840 ± 0.034	N/A	0.963 ± 0.023	0.273 ± 0.020	0.969 ± 0.033

Table A.22: Paired differences of the latent-diffusion prior and of the VAE against the Gaussian prior on the same encoder (Table 6), and their comparison with the prior-deployment effect. The VAE is VAE-reg of Table A.21, sampled from its fixed prior. The last two columns give the prior-deployment effect of the same domain, the refit-minus-cached C2ST of SEQ-FREEZE (Table A.20), minus the absolute family effect, both recomputed in every bootstrap resample, so a positive CI means the deployment effect is significantly larger in magnitude. Bold marks a CI that excludes zero. Negative is better for C2ST and positive for accuracy.

	diffusion minus Gaussian	VAE minus Gaussian		deployment minus family	
Dataset	Δ C2ST	Δ Acc.	Δ C2ST	diffusion	VAE
Synthetic SBM	-0.005 [-0.042, $+0.030$]	-0.030 [-0.081, $+0.007$]	$\mathbf{+0.100}$ [$+0.065$, $+0.137$]	$\mathbf{+0.282}$ [$+0.176$, $+0.346$]	$\mathbf{+0.186}$ [$+0.120$, $+0.247$]
CiteSeer	$+0.011$ [-0.013, $+0.031$]	-0.005 [-0.019, $+0.010$]	$\mathbf{-0.035}$ [-0.068, -0.005]	$\mathbf{+0.159}$ [$+0.123$, $+0.189$]	$\mathbf{+0.134}$ [$+0.098$, $+0.180$]
Cora	$+0.010$ [-0.010, $+0.029$]	$+0.006$ [-0.015, $+0.023$]	$+0.009$ [-0.012, $+0.032$]	$\mathbf{+0.060}$ [$+0.031$, $+0.083$]	$\mathbf{+0.060}$ [$+0.033$, $+0.083$]
PubMed	$\mathbf{+0.030}$ [$+0.012$, $+0.047$]	-0.006 [-0.044, $+0.032$]	$+0.014$ [-0.008, $+0.039$]	$+0.035$ [-0.018, $+0.083$]	$+0.051$ [-0.003, $+0.096$]
QM9	-0.029 [-0.085, $+0.013$]	-0.001 [-0.014, $+0.014$]	$\mathbf{+0.025}$ [$+0.011$, $+0.041$]	$+0.006$ [-0.026, $+0.027$]	$+0.010$ [-0.008, $+0.030$]

comparisons have small effect sizes: Cora comprehension against C-ONLY at $+0.024$, QM9 comprehension against SEQ-ANCHOR at $+0.013$, and Synthetic and CiteSeer generation against JOINT-RETAINED at $+0.006$ and $+0.008$. SEQ-REG's generation advantage over every frozen, anchored, or decoupled variant stays Holm-significant on all five datasets under the refit deployment on bootstrap p-values, with Holm-adjusted $p = 0.005$. Three results do not depend on the deployment: the Synthetic comprehension gain is reproduced by an anchor with no generative loss, the active generative

Table A.23: Two additional controls on the primary grid, with SEQ-REG for reference: SEQ-FREEZE-ACTIVE, which freezes the decoder but keeps $\mathcal{L}_{\text{gen}}$ active, and LINEAR-PROBE, which freezes the whole Phase-1 model and trains only the comprehension head, so that its generator is the Phase-1 generator and its CKA is 1 by construction. Generation and round-trip MMD use the refit deployment.

Condition	Acc.	Gen-MMD	C2ST	Round-trip MMD	CKA
Synthetic SBM					
SEQ-FREEZE-ACTIVE	0.884 ± 0.049	0.013 ± 0.008	0.589 ± 0.037	0.008 ± 0.004	0.720 ± 0.103
LINEAR-PROBE	0.934 ± 0.008	0.009 ± 0.005	0.550 ± 0.028	0.004 ± 0.003	1.000 ± 0.000
SEQ-REG	0.924 ± 0.028	0.006 ± 0.003	0.538 ± 0.041	0.003 ± 0.004	0.863 ± 0.031
CiteSeer					
SEQ-FREEZE-ACTIVE	0.591 ± 0.024	0.025 ± 0.011	0.758 ± 0.025	0.015 ± 0.006	0.433 ± 0.136
LINEAR-PROBE	0.592 ± 0.033	0.018 ± 0.010	0.736 ± 0.026	0.016 ± 0.011	1.000 ± 0.000
SEQ-REG	0.616 ± 0.024	0.012 ± 0.005	0.716 ± 0.038	0.017 ± 0.011	0.518 ± 0.054
Cora					
SEQ-FREEZE-ACTIVE	0.694 ± 0.024	0.055 ± 0.014	0.825 ± 0.031	0.032 ± 0.006	0.323 ± 0.171
LINEAR-PROBE	0.697 ± 0.036	0.043 ± 0.008	0.830 ± 0.027	0.034 ± 0.008	1.000 ± 0.000
SEQ-REG	0.683 ± 0.033	0.036 ± 0.006	0.816 ± 0.026	0.034 ± 0.006	0.492 ± 0.075
PubMed					
SEQ-FREEZE-ACTIVE	0.619 ± 0.057	0.096 ± 0.017	0.837 ± 0.035	0.076 ± 0.023	0.596 ± 0.117
LINEAR-PROBE	0.647 ± 0.041	0.109 ± 0.030	0.836 ± 0.035	0.106 ± 0.043	1.000 ± 0.000
SEQ-REG	0.671 ± 0.047	0.075 ± 0.014	0.806 ± 0.025	0.074 ± 0.036	0.650 ± 0.081
QM9					
SEQ-FREEZE-ACTIVE	0.803 ± 0.049	0.276 ± 0.017	0.980 ± 0.010	0.279 ± 0.021	0.681 ± 0.236
LINEAR-PROBE	0.704 ± 0.033	0.290 ± 0.029	0.964 ± 0.042	0.288 ± 0.037	1.000 ± 0.000
SEQ-REG	0.840 ± 0.034	0.269 ± 0.021	0.963 ± 0.023	0.273 ± 0.020	0.708 ± 0.087

loss reduces encoder-decoder mismatch relative to frozen decoders, and under the learned metric no condition approaches indistinguishability outside the synthetic domain.

D.9 PER-GRAPH RECONSTRUCTION AND ADDITIONAL CONTROLS

Additional controls. SEQ-FREEZE-ACTIVE keeps $\mathcal{L}_{\text{gen}}$ active in Phase 2 but never updates the decoder, so the generative gradient reaches the encoder without decoder co-adaptation. LINEAR-PROBE freezes the whole Phase-1 model and trains only the comprehension head, so it preserves the Phase-1 generator exactly. Table A.23 reports both on the primary grid with the protocol of Appendix B. Without any decoder update, SEQ-FREEZE-ACTIVE reaches round-trip MMD 0.008 to 0.279, below both frozen conditions on every domain (Table 5), so the compatibility comes from the generative gradient rather than from decoder updates. Its accuracy is significantly below SEQ-REG on Synthetic, CiteSeer, PubMed, and QM9, by 0.03 to 0.05, so freezing the decoder under an active generative loss costs accuracy. LINEAR-PROBE is significantly less accurate than SEQ-REG on CiteSeer and QM9, by 0.02 and 0.14. Its generation and round-trip values are those of the Phase-1 generator.

Non-generative pool control. SEQ-SSL keeps the Phase-1 pretraining, the Phase-2 schedule, and the pool mini-batch sequence of SEQ-REG but replaces $\lambda\mathcal{L}_{\text{gen}}$ on each pool mini-batch by a graph-level contrastive loss with weight λ_{ssl} (two views per graph by dropping 20% of nodes and 20% of edges, a two-layer projection head, and NT-Xent at temperature 0.2), with the decoder frozen and outside every loss, so it matches the unlabeled exposure of SEQ-REG in Table A.14 through a non-generative objective (Table A.26). At both tested weights, SEQ-REG is significantly more accurate than SEQ-SSL on Synthetic, CiteSeer, and QM9, by 0.03 to 0.09, so equal unlabeled exposure through the contrastive objective reproduces neither the Synthetic nor the CiteSeer gain of decoder retention. On PubMed, where retention yields no gain over MAE-DISCARD, SEQ-SSL is more accurate than SEQ-REG by more than 0.04 at both weights and than MAE-DISCARD by 0.025 and 0.026.

Table A.24: Round-trip C2ST. For each seed, the discriminator of the generation C2ST separates the first $\min(400, n_{\text{test}})$ held-out test graphs from their own decodings through the deployed encoder and decoder, the draws that give the round-trip MMD of Table 5. *The null row repeats the generation C2ST of the density-matched null (Table 1) for reference. The lower panel gives the paired difference to SEQ-REG, where positive means that the condition's reconstructions are easier to tell from their inputs, and bold marks a CI that excludes zero. Checkpoints were retrained for this evaluation and reproduce the stored grid except on PubMed seed 1, without which only the PubMed differences of LINEAR-PROBE and of the $\lambda_{\text{anc}}=10$ anchor lose significance.

	Synthetic SBM	CiteSeer	Cora	PubMed	QM9
Round-trip C2ST, mean ± standard deviation					
SEQ-REG	0.534 ± 0.021	0.715 ± 0.027	0.808 ± 0.028	0.832 ± 0.023	0.969 ± 0.033
JOINT-RETAINED	0.537 ± 0.033	0.736 ± 0.029	0.862 ± 0.026	0.853 ± 0.022	0.977 ± 0.031
SEQ-FREEZE-ACTIVE	0.551 ± 0.040	0.738 ± 0.022	0.826 ± 0.022	0.844 ± 0.033	0.983 ± 0.011
SEQ-FREEZE	0.829 ± 0.112	0.874 ± 0.072	0.888 ± 0.034	0.903 ± 0.053	0.998 ± 0.004
SEQ-ANCHOR	0.751 ± 0.133	0.858 ± 0.047	0.866 ± 0.046	0.886 ± 0.026	0.949 ± 0.155
LINEAR-PROBE	0.540 ± 0.038	0.736 ± 0.025	0.809 ± 0.013	0.850 ± 0.024	0.987 ± 0.007
SEQ-ANCHOR, $\lambda_{\text{anc}}=10$	0.545 ± 0.039	0.735 ± 0.027	0.813 ± 0.020	0.850 ± 0.024	0.986 ± 0.010
G-ONLY	0.521 ± 0.023	0.720 ± 0.034	0.814 ± 0.042	0.840 ± 0.030	0.916 ± 0.117
Erdős-Rényi null*	0.772 ± 0.038	0.908 ± 0.029	0.938 ± 0.027	0.937 ± 0.014	0.975 ± 0.040
Condition minus SEQ-REG*, paired Δ round-trip C2ST [95% CI]*					
JOINT-RETAINED	$+0.003$ $[-0.021, +0.027]$	$+\mathbf{0.021}$ $[+0.000, +0.040]$	$+\mathbf{0.053}$ $[+0.032, +0.074]$	$+\mathbf{0.020}$ $[+0.002, +0.038]$	$+0.008$ $[-0.024, +0.038]$
SEQ-FREEZE-ACTIVE	$+0.018$ $[-0.005, +0.043]$	$+\mathbf{0.022}$ $[+0.008, +0.036]$	$+0.018$ $[-0.002, +0.036]$	$+0.012$ $[-0.010, +0.035]$	$+0.014$ $[-0.002, +0.035]$
SEQ-FREEZE	$+\mathbf{0.295}$ $[+0.237, +0.354]$	$+\mathbf{0.159}$ $[+0.112, +0.202]$	$+\mathbf{0.080}$ $[+0.046, +0.112]$	$+\mathbf{0.070}$ $[+0.032, +0.105]$	$+\mathbf{0.029}$ $[+0.013, +0.052]$
SEQ-ANCHOR	$+\mathbf{0.217}$ $[+0.143, +0.293]$	$+\mathbf{0.143}$ $[+0.108, +0.173]$	$+\mathbf{0.058}$ $[+0.023, +0.093]$	$+\mathbf{0.054}$ $[+0.029, +0.076]$	-0.020 $[-0.128, +0.044]$
LINEAR-PROBE	$+0.007$ $[-0.019, +0.033]$	$+\mathbf{0.021}$ $[+0.005, +0.038]$	$+0.001$ $[-0.016, +0.015]$	$+\mathbf{0.018}$ $[+0.001, +0.035]$	$+\mathbf{0.018}$ $[+0.002, +0.039]$
SEQ-ANCHOR, $\lambda_{\text{anc}}=10$	$+0.011$ $[-0.013, +0.033]$	$+\mathbf{0.019}$ $[+0.008, +0.030]$	$+0.005$ $[-0.009, +0.020]$	$+\mathbf{0.018}$ $[+0.000, +0.038]$	$+\mathbf{0.017}$ $[+0.000, +0.038]$
G-ONLY	$-\mathbf{0.012}$ $[-0.025, -0.001]$	$+0.004$ $[-0.010, +0.017]$	$+0.006$ $[-0.018, +0.030]$	$+0.008$ $[-0.017, +0.028]$	-0.053 $[-0.132, +0.015]$

Round-trip C2ST. Round-trip MMD compares three descriptors, so Table A.24 repeats the round-trip test with the discriminator of the generation C2ST. The learned metric narrows the result of Sec. 4.4 in two ways. First, reconstruction is no closer than sampling: the round-trip C2ST of SEQ-REG minus its generation C2ST on the same checkpoint is a tie on four domains and $+0.028$ $[+0.012, +0.043]$ on PubMed, and on QM9 no condition's round-trip is significantly less separable than the density-matched null, with -0.005 $[-0.035, +0.025]$ for SEQ-REG. Second, the contrast between active and frozen conditions holds outside QM9. SEQ-REG round-trips significantly better than SEQ-FREEZE on all five domains and than SEQ-ANCHOR on the four domains other than QM9, and so does SEQ-FREEZE-ACTIVE against both frozen conditions on those four. JOINT-RETAINED round-trips significantly better than both on Synthetic, CiteSeer, and PubMed. LINEAR-PROBE and the $\lambda_{\text{anc}}=10$ anchor, whose encoders barely move, stay within 0.021 of SEQ-REG but are significantly above it on CiteSeer, PubMed, and QM9.

Per-graph reconstruction. The set-level round-trip MMD of Appendix C cannot distinguish a decoder that reconstructs each input from one that returns the right distribution of graphs in the wrong order. We therefore decode each held-out test graph's own latent and compare it with the decoding of a size-matched foreign latent, drawn by a size-stratified random permutation of the test set, and report the paired matched-minus-shuffled gap over the 10 seeds. Because the reconstruction loss is slot aligned (Appendix B), we score edge structure both in the dataset node order and after Hungarian matching of decoded to input nodes on node features. Table A.25 reports the paired gap for edge AUROC in both node orders and for the feature cosine after matching. Node features round-trip per graph under the active loss: the matched-minus-shuffled gap in feature cosine excludes zero on all five domains for SEQ-REG, for example $+0.231$ $[+0.221, +0.242]$ on Cora, and the frozen

Table A.25: Per-graph reconstruction. Each held-out test graph G_i is compared with the decoding of its own latent $\mathbf{z}_i$ (matched) and with the decoding of $\mathbf{z}_{\pi(i)}$ for a size-stratified random permutation π (shuffled). Cells give the paired matched-minus-shuffled difference, and bold marks a CI that excludes zero in the matched direction. Edge AUROC is scored on the real node pairs in the dataset node order and again after Hungarian matching of decoded to input nodes on node features, and feature cosine similarity is scored after the same matching.

	Synthetic SBM	CiteSeer	Cora	PubMed	QM9
Edge AUROC, dataset node order					
SEQ-REG	**+0.064** [+0.060, +0.067]	**+0.004** [+0.000, +0.008]	**+0.004** [+0.001, +0.007]	**+0.016** [+0.012, +0.019]	**+0.040** [+0.036, +0.045]
JOINT-RETAINED	**+0.058** [+0.050, +0.065]	+0.002 [−0.003, +0.007]	+0.000 [−0.003, +0.003]	**+0.013** [+0.008, +0.016]	**+0.016** [+0.009, +0.022]
G-ONLY	**+0.060** [+0.055, +0.066]	+0.000 [−0.004, +0.004]	**+0.004** [+0.002, +0.007]	**+0.013** [+0.009, +0.017]	**+0.016** [+0.008, +0.025]
SEQ-FREEZE	**+0.039** [+0.034, +0.044]	+0.002 [−0.005, +0.008]	+0.003 [−0.002, +0.007]	+0.001 [−0.005, +0.006]	**+0.002** [+0.001, +0.003]
SEQ-ANCHOR	**+0.052** [+0.047, +0.058]	**+0.005** [+0.000, +0.010]	−0.000 [−0.004, +0.004]	+0.002 [−0.003, +0.007]	**+0.003** [+0.002, +0.004]
Edge AUROC, after Hungarian matching					
SEQ-REG	**+0.010** [+0.006, +0.015]	**+0.017** [+0.009, +0.026]	−0.004 [−0.011, +0.003]	−0.007 [−0.016, +0.002]	**+0.009** [+0.006, +0.012]
JOINT-RETAINED	**+0.005** [+0.003, +0.008]	−0.001 [−0.011, +0.009]	−0.010 [−0.016, −0.005]	−0.013 [−0.025, −0.003]	+0.000 [−0.002, +0.002]
G-ONLY	**+0.010** [+0.007, +0.013]	+0.004 [−0.005, +0.013]	+0.002 [−0.003, +0.008]	−0.006 [−0.012, +0.000]	+0.001 [−0.001, +0.003]
SEQ-FREEZE	+0.001 [−0.000, +0.002]	−0.006 [−0.017, +0.004]	+0.002 [−0.005, +0.009]	−0.004 [−0.013, +0.004]	+0.000 [−0.001, +0.002]
SEQ-ANCHOR	**+0.003** [+0.002, +0.004]	−0.013 [−0.021, −0.006]	+0.000 [−0.007, +0.008]	−0.006 [−0.015, +0.002]	+0.000 [−0.002, +0.002]
Feature cosine, after Hungarian matching					
SEQ-REG	**+0.136** [+0.132, +0.140]	**+0.288** [+0.280, +0.294]	**+0.231** [+0.221, +0.242]	**+0.340** [+0.316, +0.362]	**+0.015** [+0.012, +0.018]
JOINT-RETAINED	**+0.117** [+0.091, +0.133]	**+0.229** [+0.218, +0.239]	**+0.184** [+0.164, +0.201]	**+0.271** [+0.249, +0.293]	**+0.005** [+0.003, +0.007]
G-ONLY	**+0.136** [+0.132, +0.140]	**+0.209** [+0.186, +0.231]	**+0.205** [+0.180, +0.226]	**+0.239** [+0.205, +0.267]	**+0.004** [+0.002, +0.006]
SEQ-FREEZE	**+0.018** [+0.013, +0.023]	**+0.102** [+0.090, +0.112]	**+0.116** [+0.101, +0.130]	**+0.067** [+0.057, +0.078]	+0.002 [−0.001, +0.004]
SEQ-ANCHOR	**+0.066** [+0.055, +0.077]	**+0.105** [+0.090, +0.118]	**+0.124** [+0.109, +0.139]	**+0.096** [+0.082, +0.110]	+0.002 [−0.000, +0.004]

conditions retain a smaller gap that is resolved on every domain except QM9. Edge structure does not: in slot order the edge-AUROC gap for SEQ-REG is positive on all five domains, but after Hungarian matching it is resolved only on Synthetic, CiteSeer, and QM9, by at most 0.017, and after matching the edge AUROC of the matched decoding lies between 0.49 and 0.51 for every condition on the three citation domains. We therefore claim per-graph correspondence for node attributes and read the structural round-trip as a distribution-level statement.

E LIMITATIONS

This study isolates mechanisms rather than maximizing scale or leaderboard performance. Its claims hold for compact GIN autoencoders trained on a single GPU across five domains, with a GAT ablation and an ogbn-arxiv study as the only backbone and scale checks. External generators enter as single-GPU reference rows, not as tuned or budget-matched competitors, and a dedicated molecular generator trained on one GPU comes close to the real-versus-real floor on QM9 under our evaluator, which the compact dual-use model does not. The C2ST verdict is relative to a density-matched Erdős-Rényi null and the chosen discriminator. Comprehension is graph-level classification throughout, and QM9 is scored as graph structure rather than for chemical validity. The null matches

Table A.26: Non-generative unlabeled-pool control. SEQ-SSL keeps the Phase-1 pretraining, the Phase-2 schedule, and the pool mini-batch sequence of SEQ-REG but replaces $\lambda\mathcal{L}_{\text{gen}}$ on each pool mini-batch by a graph-level contrastive loss with weight λ_{ssl} and a frozen decoder (Appendix D.9). Its pool exposure equals that of SEQ-REG in Table A.14, and its deployed generator is the Phase-1 generator of LINEAR-PROBE. SEQ-REG and MAE-DISCARD repeat Tables 3 and 4. The lower panels give paired accuracy differences, and bold marks a CI that excludes zero.

	Synthetic SBM	CiteSeer	Cora	PubMed	QM9
Accuracy, mean ± standard deviation					
SEQ-SSL, λ_{ssl}=1	0.847 ± 0.024	0.576 ± 0.029	0.698 ± 0.023	0.714 ± 0.024	0.794 ± 0.030
SEQ-SSL, λ_{ssl}=0.1	0.837 ± 0.028	0.584 ± 0.033	0.677 ± 0.029	0.716 ± 0.025	0.804 ± 0.029
SEQ-REG	0.924 ± 0.028	0.616 ± 0.024	0.683 ± 0.033	0.671 ± 0.047	0.840 ± 0.034
MAE-DISCARD	0.843 ± 0.034	0.571 ± 0.033	0.682 ± 0.027	0.690 ± 0.030	0.835 ± 0.036
CKA, mean ± standard deviation					
SEQ-SSL, λ_{ssl}=1	0.742 ± 0.045	0.424 ± 0.065	0.498 ± 0.051	0.477 ± 0.061	0.722 ± 0.074
SEQ-SSL, λ_{ssl}=0.1	0.763 ± 0.065	0.417 ± 0.055	0.441 ± 0.071	0.420 ± 0.031	0.564 ± 0.152
SEQ-REG	0.863 ± 0.031	0.518 ± 0.054	0.492 ± 0.075	0.650 ± 0.081	0.708 ± 0.087
MAE-DISCARD	0.515 ± 0.053	0.313 ± 0.077	0.394 ± 0.088	0.261 ± 0.061	0.308 ± 0.097
Accuracy at λ_{ssl}=1, paired Δ [95% CI]					
SEQ-REG minus SEQ-SSL	$+\mathbf{0.077}$ [+0.058, +0.099]	$+\mathbf{0.040}$ [+0.018, +0.063]	-0.016 [−0.036, +0.004]	$-\mathbf{0.043}$ [−0.083, −0.012]	$+\mathbf{0.046}$ [+0.032, +0.061]
SEQ-SSL minus MAE-DISCARD	$+0.003$ [−0.015, +0.024]	$+0.005$ [−0.018, +0.029]	$+0.016$ [−0.003, +0.032]	$+\mathbf{0.025}$ [+0.007, +0.048]	$-\mathbf{0.041}$ [−0.059, −0.021]
Accuracy at λ_{ssl}=0.1, paired Δ [95% CI]					
SEQ-REG minus SEQ-SSL	$+\mathbf{0.087}$ [+0.068, +0.109]	$+\mathbf{0.032}$ [+0.014, +0.049]	$+0.006$ [−0.011, +0.026]	$-\mathbf{0.045}$ [−0.077, −0.015]	$+\mathbf{0.036}$ [+0.025, +0.046]
SEQ-SSL minus MAE-DISCARD	-0.007 [−0.020, +0.008]	$+0.014$ [−0.006, +0.033]	-0.006 [−0.022, +0.010]	$+\mathbf{0.026}$ [+0.009, +0.043]	$-\mathbf{0.031}$ [−0.049, −0.012]

sizes and one pooled edge density, not the per-graph density distribution. The L2-SP control shows whether weight anchoring can reproduce a comprehension gain, not by which mechanism the generative loss produces it, and it covers only the tested anchor weights and, for equal unlabeled exposure, one contrastive objective (Appendix D.9). Round-trip compatibility is a relative, distribution-level result that under C2ST is no closer than generation and does not clear the null on QM9, and per-graph round-trip recovery is established for node attributes, whereas edge structure round-trips only at the distribution level, and the reconstruction loss is slot aligned rather than permutation invariant. The grid does not compare modalities and does not isolate representation granularity, so it cannot say whether the interfaces it separates differ from those that govern unified vision and language models, and whether they govern large pretrained graph generators remains open.