JUST ALIGN $\boldsymbol{x}$: ALIGNING PREDICTIONS, NOT REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation alignment has become an effective way to accelerate diffusion training, but its benefits do not transfer reliably to pixel-space clean-image prediction. In JiT, we find that auxiliary feature alignment can improve access to semantic features while reducing access to image variation needed for clean-image prediction, creating a mismatch between the auxiliary objective and the denoising task. This suggests a different principle: *auxiliary supervision should improve the prediction target itself rather than impose a separate representation target*. We introduce **JA$\boldsymbol{x}$** (Just Align $\boldsymbol{x}$), a prediction-supervision method that aligns clean-image predictions across noise levels. **JA$\boldsymbol{x}$** couples a noisier student observation with a cleaner observation through a Markov degradation that preserves the original JiT input distribution. Under this coupling, the oracle prediction from the cleaner state has the same conditional mean as the optimal JiT target, while its conditional target covariance is no greater. Thus, oracle prediction alignment preserves the population JiT objective up to a constant while providing a lower-variance training target. To make this construction practical with an imperfect EMA teacher, **JA$\boldsymbol{x}$** combines ground-truth supervision with a reliability-gated coupling band that selects nearby teacher states based on prediction risk. On ImageNet 256×256, **JA$\boldsymbol{x}$** consistently improves FID and accelerates convergence across JiT-B/16, L/16, and H/16, without an external encoder or changes to the architecture or sampling procedure. Gradient diagnostics further show reduced minibatch gradient variance, while ablations demonstrate that the gains cannot be explained by time reweighting alone. These results show that prediction-space supervision provides a simple and principled alternative to representation alignment for pixel-space generative models.

1 INTRODUCTION

Diffusion and flow-based models have emerged as dominant paradigms for visual generation, transforming noise into data through iterative denoising (Ho et al., 2020) or continuous transport (Song et al., 2021; Lipman et al., 2023; Liu et al., 2023). Diffusion Transformers (DiTs) provide a scalable architecture for modeling these generative dynamics (Peebles & Xie, 2023). Beyond architectural scaling, recent studies (Yu et al., 2025; Singh et al., 2026a; Jiang et al., 2026; Wang et al., 2026; Chefer et al., 2026; Wang & He, 2025) have shown that aligning or regularizing internal representations can substantially accelerate DiT training. However, most of this progress has been developed in latent diffusion (Rombach et al., 2022), where an autoencoder compresses images before generative modeling. This introduces a reconstruction–generation trade-off: latent representations must preserve image details while remaining suitable for generation (Leng et al., 2025; Yao et al., 2025; Zheng et al., 2026; Singh et al., 2026b).

Pixel-space generation removes this intermediate representation but requires the model to directly predict high-dimensional images. Motivated by the low-dimensional manifold structure of natural images, JiT (Li & He, 2026) addresses this challenge through clean-image prediction with large pixel patches, providing a simple and effective framework without pretrained tokenizers or auxiliary representation objectives. This difference in prediction space raises a fundamental question: *what auxiliary supervision is appropriate for pixel-space $\boldsymbol{x}$-prediction?*

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.

A direct transfer of representation alignment is not always effective. PixelREPA (Shin et al., 2026) shows that applying REPA can accelerate early learning but may degrade generation quality later in training, motivating adaptations to the mismatch between semantic features and pixel-level reconstruction. In the JiT settings we evaluate, transferred hidden self-alignment and regularization methods provide limited gains, while the advantage of REPA in low-resolution image generation can also reverse as model size decreases. One important difference is the input dimensionality: DiT-B/2 maps a 16-dimensional latent patch to a 768-dimensional hidden token, whereas JiT-B/16 maps 768 pixel values to the same hidden width. Although these dimensions alone do not determine model capacity, they highlight a potential conflict between auxiliary feature learning and clean-image prediction. Our probes show increased linear accessibility of aligned semantic features together with reduced accessibility of clean-image variation not linearly explained by those features (Section 3). These observations motivate a different principle: rather than imposing an auxiliary representation target, auxiliary supervision should directly support the prediction target of the generative objective.

We therefore ask whether an auxiliary prediction target can preserve the conditional mean required by JiT while reducing the conditional target variance. Based on this principle, we introduce **JA**$\boldsymbol{x}$ (Just Align $\boldsymbol{x}$), which aligns clean-image predictions across noise levels. Specifically, we couple a cleaner observation with the student's noisier input through a Markov degradation that preserves the original student-input marginal. Under this coupling, the optimal clean-image predictions satisfy a martingale consistency property: averaging the cleaner prediction conditioned on the noisier observation recovers the optimal prediction at that noise level. With the original JiT time distribution and loss weighting, this oracle supervision induces the same population regression objective up to a parameter-independent constant, while its conditional target covariance is no greater than that of direct clean-image supervision.

In practice, **JA**$\boldsymbol{x}$ uses an exponential-moving-average (EMA) teacher to predict $\boldsymbol{x}$ from the cleaner observation. Because teacher errors can bias the auxiliary target, we retain supervision from the ground-truth clean image and introduce a reliability-gated coupling band that selects nearby teacher states according to an empirical prediction-risk criterion. This selection balances the potential variance reduction from cleaner targets against the error introduced by the evolving teacher. **JA**$\boldsymbol{x}$ requires no external encoder and leaves the JiT architecture and sampling procedure unchanged. Across model scales, it improves generation quality and accelerates convergence in training epochs.

Our contributions are summarized as follows:

- We empirically examine the limitations of hidden representation alignment in JiT and identify conditional target consistency as a guiding principle for auxiliary prediction supervision.
- We introduce Markov-coupled prediction supervision and establish conditional-mean consistency, population loss equivalence, and conditional covariance reduction for oracle predictions.
- We develop reliability-gated pair selection for an evolving EMA teacher and demonstrate improved generation quality and faster convergence across JiT model scales.

2 Related Works

Pixel-space diffusion Early diffusion models directly learn the generative process in RGB space (Ho et al., 2020; Song et al., 2021), whereas modern large-scale systems commonly adopt latent diffusion to reduce the spatial and computational burden through a pretrained autoencoder (Rombach et al., 2022). This efficiency, however, makes generation dependent on a separately learned latent space and its reconstruction properties. Recent works (Baade et al., 2026; Chen et al., 2025b; 2026; Yu et al., 2026; Ma et al., 2026; Li & He, 2026) revisit end-to-end pixel-space generation from several directions: PixelFlow progressively increases spatial resolution along the generative trajectory (Chen et al., 2025b); PixelDiT and DeCo separate semantic or low-frequency modeling from pixel-level details (Yu et al., 2026; Ma et al., 2026); JiT instead studies a particularly minimal formulation based on large pixel patches and direct clean-image prediction, showing that high-dimensional RGB observations can be modeled without a tokenizer or an auxiliary decoder (Li & He, 2026). Our work retains this minimal JiT architecture and focuses on accelerating its optimization without introducing a separate representation space or modifying its sampling process.

Representation Learning. Beyond serving as a compressed generation space (Rombach et al., 2022), learned representations can provide useful structure and supervision for generative modeling. Semantic tokenizers and representation autoencoders explore generation spaces that combine semantic information with image reconstruction (Chen et al., 2025a; Zheng et al., 2026; Singh et al., 2026b), while representation alignment methods use learned features to guide the optimization of generative models. REPA (Yu et al., 2025) aligns intermediate diffusion features with representations from pretrained visual encoders, improving the hidden representation leading to better generation quality and training efficiency in latent DiTs. Subsequent work explores the choice of alignment targets (Singh et al., 2026a), the placement and duration of supervision (Liu et al., 2025), and the joint adaptation of the tokenizer and denoising network (Leng et al., 2025). Self-alignment methods (Jiang et al., 2026; Wang et al., 2026; Chefer et al., 2026) further investigate whether useful representation guidance can be obtained from the generative model itself, reducing reliance on external encoders. However, changing the source of supervision does not necessarily remove the additional representational demands imposed by hidden-feature alignment. PixelREPA adapts feature alignment to pixel-space clean-image prediction (Shin et al., 2026). Nevertheless, it retains hidden representations as the alignment target, potentially imposing demands that compete with pixel reconstruction within shared model capacity. This motivates aligning clean-image predictions themselves: our method provides additional supervision for the denoising task without requiring hidden features to match a separate representation target, with conditional consistency and variance reduction established under a Markov coupling for optimal predictions.

3 Motivation and Analysis

Preliminaries: x-prediction. Let $X \sim p_{\text{data}}(x \mid c)$ and $Z_t = tX + (1-t)\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ is independent of X, and $t = 0$ denotes noise while $t = 1$ denotes clean data. JiT directly predicts the clean image by minimizing

$$\mathcal{L}_{\text{JiT}} = \mathbb{E}\big[w(t)\, \|f_\theta(Z_t, t, c) - X\|_2^2\big], \qquad w(t) = \frac{1}{(1-t)^2}, \tag{1}$$

where the expectation is over the data, conditioning, corruption noise, and the JiT time distribution $t \sim p(t)$ on $(0, 1)$. Writing $U_t = (Z_t, t, c)$ for the student input, the population regression target is $m_t(Z_t, c) = \mathbb{E}[X \mid U_t]$.

3.1 Auxiliary Supervision in JiT

Conventional feature alignment supervises an intermediate representation $h_\theta^\ell(Z_t, t, c)$ toward a feature target $\phi(X)$. Under squared regression, the corresponding conditional statistic is

$$r_t(Z_t, c) = \mathbb{E}[\phi(X) \mid U_t], \qquad r_t \not\equiv \phi \circ m_t \quad \text{in general.} \tag{2}$$

A sufficiently expressive representation may support both statistics. Their interaction nevertheless merits examination in JiT: DiT-B/2 maps a 16-dimensional latent patch into a 768-dimensional hidden token, whereas JiT-B/16 maps 768 pixel values into the same width. These dimensions do not determine effective model capacity, but motivate studying how auxiliary feature learning interacts with the representational demands of clean-image prediction.

Training behavior. PixelREPA (Shin et al., 2026) relates the difficulty of applying REPA in pixel space to the mismatch between high-resolution pixels and compressed semantic features, reporting improvements at ImageNet 32×32 but later degradation at 256×256. We revisit the low-resolution regime with a smaller JiT-S/4 (Fig. 1). REPA improves FID from 16.37 to 16.08 at 50 epochs and from 12.89 to 12.03 at 100 epochs. At 200 and 400 epochs, however, it obtains FIDs of 10.79 and 9.90, compared with 9.15 and 7.65 for vanilla JiT. Thus, reducing resolution does not by itself ensure sustained gains from feature alignment. Our JiT-B/16 comparisons also show limited gains from directly transferred hidden self-alignment and regularization methods (Table 1).

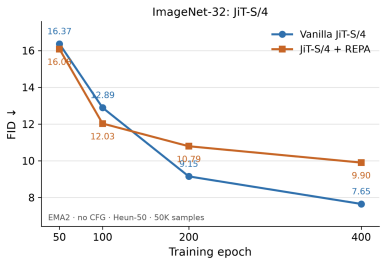

Figure 1: REPA initially improves FID on JiT-S/4 at ImageNet 32×32, but the advantage reverses later in training.

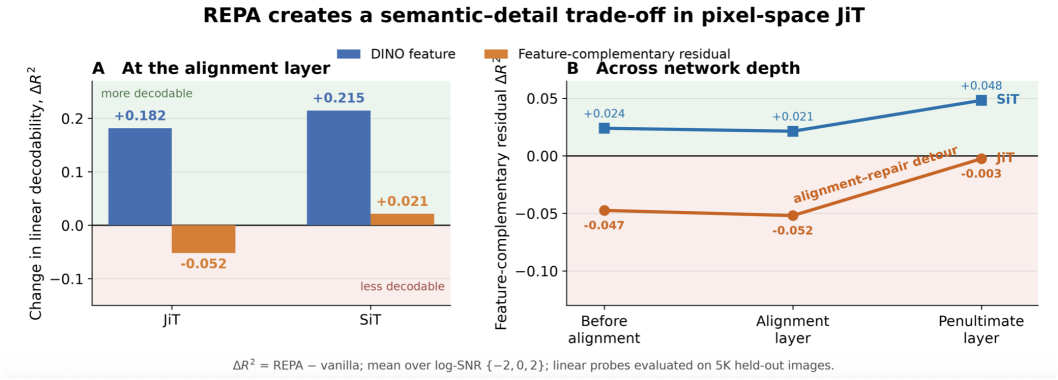

Figure 2: Linear-probe observations at epoch 200. (A) REPA increases DINO-feature decodability in both JiT and SiT, while feature-complementary residual decodability decreases in JiT and increases in SiT. (B) JiT exhibits an *alignment–repair detour* in linear decodability: its residual deficit narrows toward the output. SiT shows gains at all measured layers.

These observations motivate examining the interaction between auxiliary feature learning and clean-image prediction in the tested JiT settings.

Linear-probe observations. We probe intermediate representations at epoch 200 using two clean-image targets: the DINO feature $\phi(X)$ (Oquab et al., 2024), and the feature-complementary residual $X_\perp = X - \Pi_\phi X$, where $\Pi_\phi X$ is the linear prediction of X from $\phi(X)$. The latter measures access to clean-image variation not linearly explained by the aligned features. Figure 2 reports changes in probe R^2 relative to vanilla models, averaged over log-SNR levels $\{-2, 0, 2\}$. At the alignment layer, REPA increases DINO-feature decodability by 0.182 in JiT and 0.215 in SiT (Ma et al., 2024). Residual decodability, however, decreases by 0.052 in JiT and increases by 0.021 in SiT. Across depth, JiT's residual decodability changes are -0.047 before the alignment layer, -0.052 at that layer, and -0.003 at the penultimate layer; SiT shows positive changes at all three layers.

We refer to this pattern as an *alignment–repair detour*: the deficit in linear decodability of feature-complementary variation at intermediate layers largely diminishes toward the output. This pattern concerns linear accessibility and does not by itself establish information loss or an additional optimization cost. These observations are consistent with a capacity-sensitive trade-off in the evaluated JiT setting: greater linear accessibility of the aligned features accompanies reduced accessibility of complementary clean-image variation. They provide an empirical motivation to study auxiliary supervision whose relation to the denoising target can be analyzed directly.

3.2 What Should an Auxiliary Prediction Target Preserve?

These observations motivate auxiliary supervision directly on the clean-image prediction. To keep this supervision aligned with JiT's denoising task, a target Y with finite second moments should satisfy *conditional target consistency*:

$$\mathbb{E}[Y \mid U_t] = \mathbb{E}[X \mid U_t] = m_t(Z_t, c). \tag{3}$$

This condition preserves the population regression target required by JiT.

Consistency alone, however, does not explain how an auxiliary target can improve training: the clean-image target X already satisfies this condition. A potential benefit lies in reducing the variability of the supervision. Conditional on a noisy input U_t, clean-image targets vary around the posterior mean; under squared regression, this target variation contributes to stochastic-gradient variability. We therefore seek a target that retains the same conditional mean while reducing conditional target variance, expressed by the covariance ordering

$$\mathrm{Cov}(Y \mid U_t) \preceq \mathrm{Cov}(X \mid U_t), \tag{4}$$

where $\preceq$ denotes the positive-semidefinite order. This provides a possible optimization benefit while preserving the underlying regression task.

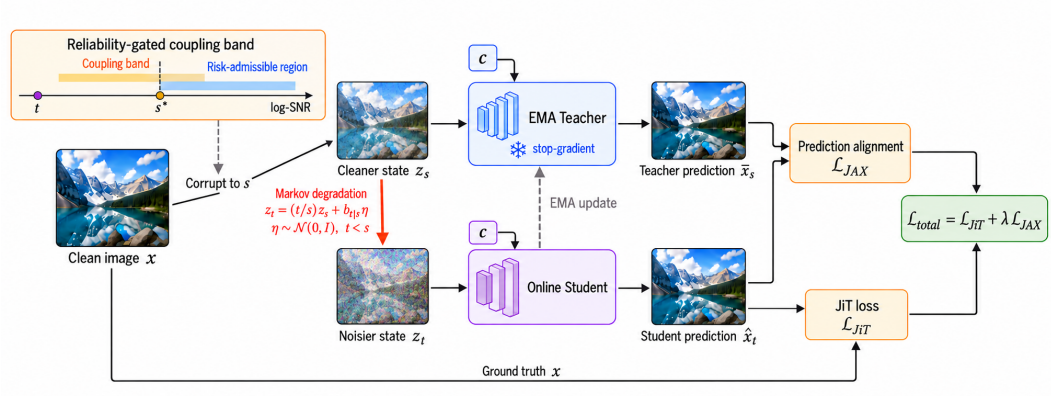

Figure 3: Overview of **JA$\boldsymbol{x}$** . An EMA teacher predicts the clean image from a cleaner observation, while the online student receives a Markov-coupled noisier observation. Training combines direct JiT supervision with prediction alignment, without modifying the architecture or sampling procedure.

In Section 4, we construct a target from a cleaner, Markov-coupled observation and establish both properties for oracle clean-image predictions. We then address the approximation error of an actual EMA teacher through ground-truth supervision and reliability-gated pair selection.

4 Method

We present **JA$\boldsymbol{x}$** , which supervises clean-image predictions across Markov-coupled noise levels. We first construct an oracle target satisfying the conditional consistency and variance criteria of Section 3.2, then instantiate the method with an EMA teacher and reliability-gated pair selection.

4.1 Markov-Coupled Observations

We construct the cleaner and noisier observations through a Markov degradation that preserves the student's original corruption marginal. Figure 3 illustrates the resulting training pipeline.

Let $0 < t < s < 1$, so that Z_s is the cleaner state. We first corrupt the clean image and then apply an incremental Gaussian degradation:

$$Z_s = sX + (1-s)\epsilon_s, \quad Z_t = a_{t|s}Z_s + b_{t|s}\eta, \quad a_{t|s} = \tfrac{t}{s}, \quad b_{t|s}^2 = (1-t)^2 - \tfrac{t^2}{s^2}(1-s)^2. \tag{5}$$

Here $\epsilon_s, \eta \sim \mathcal{N}(0, I)$ are independent of each other and of (X, c), and both predictions use the same conditioning c. This transition is well defined for $t < s$ and preserves the JiT marginal, $Z_t \mid (X = x, c) \sim \mathcal{N}(tx, (1-t)^2 I)$, while inducing the Markov chain

$$X \longrightarrow Z_s \longrightarrow Z_t \qquad \text{conditional on } (c, t, s). \tag{6}$$

Thus, the student receives the same marginally corrupted input as in standard JiT, now coupled to a cleaner observation. This Markov structure yields the conditional consistency established below; independent corruptions of the same image do not generally provide this property. Appendix A.1 verifies the construction.

4.2 Oracle Properties

Recall that $U_t = (Z_t, t, c)$ and $m_t(Z_t, c) = \mathbb{E}[X \mid U_t]$. Under the Markov coupling, the cleaner posterior prediction provides a conditionally consistent target for the student.

Proposition 1 (Martingale consistency)**.** *For any $0 < t < s < 1$, the oracle clean-image predictions satisfy*

$$\mathbb{E}\big[m_s(Z_s, c) \mid U_t, s\big] = m_t(Z_t, c). \tag{7}$$

The cleaner prediction can exploit information unavailable in Z_t, so the two predictions need not agree for each paired sample. Averaging over cleaner observations compatible with the student input

nevertheless recovers exactly its posterior mean. This conditional consistency gives the posterior predictions their martingale structure, formalized in Appendix A.1.

Let $\pi(t, s) = p(t)q(s \mid t)$ be a pair distribution supported on $t < s$, with the JiT student-time marginal $p(t)$. The pair is sampled independently of (X, c) and the corruption noises. Using the same weight $w(t)$ as JiT, define the oracle objective

$$\mathcal{L}^{\star}_{\mathrm{JA}x}(\theta) = \mathbb{E}_{\pi}\Big[w(t)\big\|f_\theta(U_t) - \mathrm{sg}[m_s(Z_s, c)]\big\|_2^2\Big], \tag{8}$$

where sg denotes stop-gradient and the expectation also includes the data and coupled corruptions. Proposition 1 implies

$$\mathcal{L}^{\star}_{\mathrm{JA}x}(\theta) = \mathcal{L}_{\mathrm{JiT}}(\theta) + K, \qquad \nabla_\theta \mathcal{L}^{\star}_{\mathrm{JA}x} = \nabla_\theta \mathcal{L}_{\mathrm{JiT}}, \tag{9}$$

where K is independent of θ. The oracle objective therefore preserves the population loss up to a constant, including its minimizers and expected gradient.

Although the population gradients agree, the regression targets have different conditional variability. The Markov coupling yields the following covariance ordering.

Proposition 2 (Variance reduction)**.** *For any fixed* $0 < t < s < 1$*,*

$$\mathrm{Cov}\big(m_s(Z_s, c) \mid U_t\big) = \mathrm{Cov}(X \mid U_t) - \mathbb{E}\big[\mathrm{Cov}(X \mid Z_s, c) \mid U_t\big] \preceq \mathrm{Cov}(X \mid U_t). \tag{10}$$

For fixed t, the covariance reduction is nonincreasing as s increases. The oracle target approaches $m_t(Z_t, c)$ as $s \downarrow t$ and X as $s \uparrow 1$, with both limits in mean square.

The oracle target removes the posterior uncertainty that remains unresolved at the cleaner observation while retaining the conditional mean required by JiT. Thus, it satisfies both criteria in Section 3.2, and $K \leq 0$. For fixed student parameters, replacing X with this oracle target also preserves the expected gradient and does not increase its covariance. Proofs, endpoint limits, and the corresponding gradient analysis are given in Appendices A.2 and A.4.

4.3 Practical Prediction Supervision

The oracle predictor m_s is unavailable in practice. We approximate it with an exponential-moving-average (EMA) teacher, whose parameters $\bar{\theta}$ track the online student. Given the cleaner observation, the teacher produces the target $\bar{x}_s = \mathrm{sg}[f_{\bar{\theta}}(Z_s, s, c)]$.

Let $a_k(t) \in \{0, 1\}$ indicate whether alignment is active at update k, and let $s^{\star}_k(t)$ be the selected teacher time, as specified in Section 4.4. With the teacher and selection rule held fixed during the student update, we combine prediction alignment with direct clean-image supervision:

$$\mathcal{L}_{\mathrm{JA}x,k}(\theta) = \mathbb{E}\Big[a_k(t)w(t)\big\|f_\theta(U_t) - \bar{x}_{s^{\star}_k(t)}\big\|_2^2\Big], \quad \mathcal{L}_{\mathrm{total},k} = \mathcal{L}_{\mathrm{JiT}} + \lambda\mathcal{L}_{\mathrm{JA}x,k}. \tag{11}$$

The alignment integrand is zero for inactive times, without requiring a teacher prediction.

Approximation error can shift the teacher target's conditional mean and alter its covariance, so the EMA objective does not automatically inherit the oracle guarantees. Direct clean-image supervision attenuates the teacher-induced shift in the conditional regression target. We further restrict alignment to empirically reliable teacher states through the selection rule below. Appendix A.5 analyzes the teacher bias and the effective target and weighting of the combined objective.

4.4 Reliability-Gated Pair Selection

The oracle analysis favors nearby teacher states, which provide greater variance reduction. As s approaches t, however, the conditional bias of an EMA teacher need not vanish: alignment approaches same-state EMA self-distillation, whose usefulness depends on teacher accuracy. At the opposite endpoint, the oracle target approaches X as $s \to 1$, leaving little variance reduction. We therefore seek an empirically reliable teacher within a bounded separation from the student.

Coupling band. We parameterize time by the log-SNR $\ell_t = 2 \log \frac{t}{1-t}$. Under Eq. 5, the correlation between the standardized corruption noises is

$$r(t,s) = \frac{t(1-s)}{s(1-t)} = \exp\left(-\frac{\ell_s - \ell_t}{2}\right). \tag{12}$$

The gap $\Delta\ell = \ell_s - \ell_t$ therefore controls the shared corruption noise. We constrain active pairs to $\Delta_{\min} \le \Delta\ell \le \Delta_{\max}$, with $\Delta_{\min} > 0$, to avoid nearly identical or excessively separated states.

Empirical reliability frontier. We periodically estimate the teacher's mean squared clean-image prediction risk $R_k(\ell)$ by log-SNR. Using estimates available before the current batch, we define

$$\ell_{\text{safe}}(k) = \inf\{\ell : R_k(\ell) \le R^\star\} + \delta_{\text{safe}}, \tag{13}$$

where $R^\star$ is the admissible risk and δ_{safe} is a margin toward cleaner states. This frontier approximates an admissible region on the cleaner side of the risk crossing. Prediction risk motivates this empirical reliability criterion through its relation to teacher bias (Appendix A.5).

For a student time t, we choose

$$\ell_s^\star = \max\{\ell_t + \Delta_{\min}, \ell_{\text{safe}}(k)\}, \qquad s_k^\star(t) = \text{sigmoid}(\ell_s^\star/2). \tag{14}$$

We set $a_k(t) = 1$ only if the frontier is available, $\ell_s^\star - \ell_t \le \Delta_{\max}$, and $\ell_s^\star \le \ell_{\max}$; otherwise, the sample receives only JiT supervision. Within this frontier-defined feasible set, the selected state is the nearest admissible teacher and maximizes the oracle covariance reduction. This motivates the rule as an empirical response to the bias–variance trade-off; its effectiveness with an EMA teacher is evaluated experimentally.

Batch objective. Let $\mathcal{A}_k = \{i : a_k(t_i) = 1\}$ denote the active samples in a batch of size B. The alignment loss is

$$\widehat{\mathcal{L}}_{\text{JA}x,k} = \frac{1}{B} \sum_{i \in \mathcal{A}_k} \frac{\left\| f_\theta(Z_{t_i}, t_i, c_i) - \bar{x}_{s_k^\star(t_i)} \right\|_2^2}{(1-t_i)^2}. \tag{15}$$

Normalization uses the full batch size, so inactive samples contribute zero alignment loss and each active sample retains its weight as the activation fraction changes. Training pseudocode (Algorithm 1), calibration settings, and other implementation details are provided in Appendix C.

5 Experiments

Experimental setup. We evaluate on ImageNet 256×256 using the JiT-B/16, L/16, and H/16 architectures and base training recipe (Li & He, 2026), with a shared alignment weight $\lambda = 0.5$. We report FID (Heusel et al., 2017) and Inception Score (IS) (Salimans et al., 2016) using 50K generated samples. Sampling follows JiT with a 50-step Heun ODE solver and CFG (Ho & Salimans, 2022) over the interval $[0.1, 1]$ (Kynkäänniemi et al., 2024). Full training, EMA, and evaluation settings are provided in Appendix C.

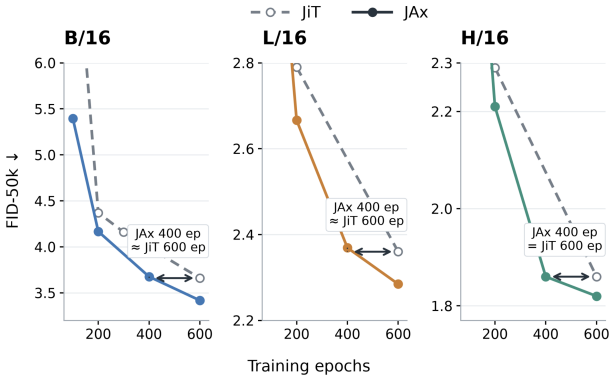

Figure 4: Training convergence across JiT scales. Dashed and solid lines denote JiT and **JA$\boldsymbol{x}$** , respectively. JiT-L/16 and H/16 baselines use the reported 200- and 600-epoch results.

Generation quality and convergence. Figure 4 shows improved FID and faster convergence in training epochs across model scales. At 600 epochs, **JA$\boldsymbol{x}$** reduces FID from 3.66 to 3.41 on B/16, from 2.36 to 2.26 on L/16, and from 1.86 to 1.82 on H/16. It also reaches approximately the corresponding JiT model's 600-epoch FID at 400 epochs. These results support the effectiveness of prediction supervision across the evaluated scales. Full metrics and training-cost analysis are provided in Appendix E; broader system-level comparisons appear in Appendix F.1.

Comparison with feature alignment methods. Table 1 compares auxiliary supervision methods on JiT-B/16 after 200 epochs. Directly transferring existing methods yields limited gains: SRA 2, SRA, and REPA underperform vanilla JiT, while Self-Flow improves FID from 4.37 to 4.31. By supervising clean-image predictions, **JA$\boldsymbol{x}$** achieves 4.17 without an external encoder, supporting our motivation to align auxiliary supervision with the denoising target. PixelREPA achieves a lower FID of 4.00 using external semantic supervision. Combining it with **JA$\boldsymbol{x}$** further improves FID to 3.91, supporting the compatibility of prediction supervision and pixel-specific feature alignment in this setting.

Table 1: Auxiliary supervision on JiT-B/16 after 200 epochs. External representations are supplied by a separate pretrained encoder. Bold and underlining mark the best FID overall and without external representations, respectively.

Method	External representations	FID ↓
JiT	None	4.37
JiT + SRA 2	VAE	4.48
JiT + SRA	None	4.38
JiT + Self-Flow	None	4.31
JiT + REPA	DINOv2	5.12
JA$\boldsymbol{x}$ (ours)	None	<u>4.17</u>
PixelREPA	DINOv2	4.00
PixelREPA + **JA$\boldsymbol{x}$**	DINOv2	**3.91**

Optimization behavior. Figure 5 compares paired minibatch gradients under JiT supervision and the combined **JA$\boldsymbol{x}$** objective, normalized by $1 + \lambda$ to remove the overall scale difference. We evaluate our epoch-200 vanilla and **JA$\boldsymbol{x}$** checkpoints, together with an official vanilla JiT checkpoint and our 600-epoch **JA$\boldsymbol{x}$** checkpoint, using their corresponding epoch EMA teacher. All paired samples used in this diagnostic have active prediction alignment, i.e., $a_k(t) = 1$. Normalized **JA$\boldsymbol{x}$** gradients exhibit lower variance at every evaluated noise level, with an average reduction of 9.6% across noise levels and student states (9.4% at epoch 200 and 9.9% in the 600-epoch comparison). Accounting also for the mean-gradient shift, the estimated mean-squared deviation from the corresponding JiT population gradient decreases by 6.5% (6.4% at epoch 200 and 6.6% for 600). These measurements support the variance-reduction motivation of our oracle analysis in the tested EMA setting; the decomposition and theoretical scope are given in Appendix A.4.

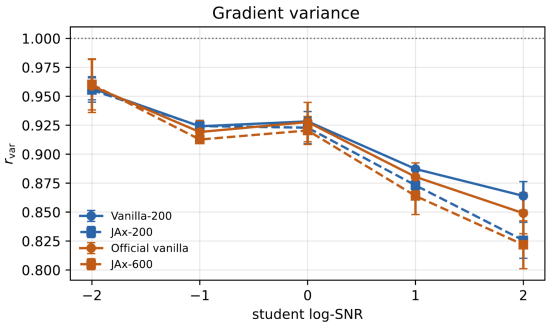

Figure 5: Minibatch gradient variance relative to JiT at epoch 200 and epoch 600, after normalizing the combined **JA$\boldsymbol{x}$** gradient by $1 + \lambda$. **JA$\boldsymbol{x}$** has lower gradient variance across training stages.

Qualitative results. Figure 6 presents samples generated by **JA$\boldsymbol{x}$** across diverse ImageNet categories. These examples illustrate the visual quality of the generated images and complement the quantitative improvements in FID. The reported FID gains are achieved by modifying only the prediction supervision during training, while retaining the JiT-H architecture and the original sampling procedure, with no additional inference cost.

6 Ablation Study

We examine teacher-state selection, observation coupling, and the distinction between target modification and time reweighting on JiT-B/16. Table 2 summarizes 200-epoch results on ImageNet 256×256.

Reliability-gated pair selection. Fixed-gap alignment at $\Delta\ell = 2$, enabled after warmup, and conditional sampling from the JiT time distribution restricted to $s > t$ yield FIDs of 4.48 and 4.43, respectively, compared with 4.37 for vanilla JiT. The fixed-gap control retains the same EMA, alignment weight, time-feasibility constraints, and full-batch normalization as **JA$\boldsymbol{x}$** . Every evaluated reliability-gated band improves upon the vanilla baseline. With the minimum gap fixed at 2, tightening the upper bound from 4 to 2.5 improves FID from 4.29 to 4.17. These results support selecting nearby teacher states through an empirical reliability criterion: neither Markov coupling alone nor fixed-gap alignment reproduces the gains of reliability-gated selection in this setting.

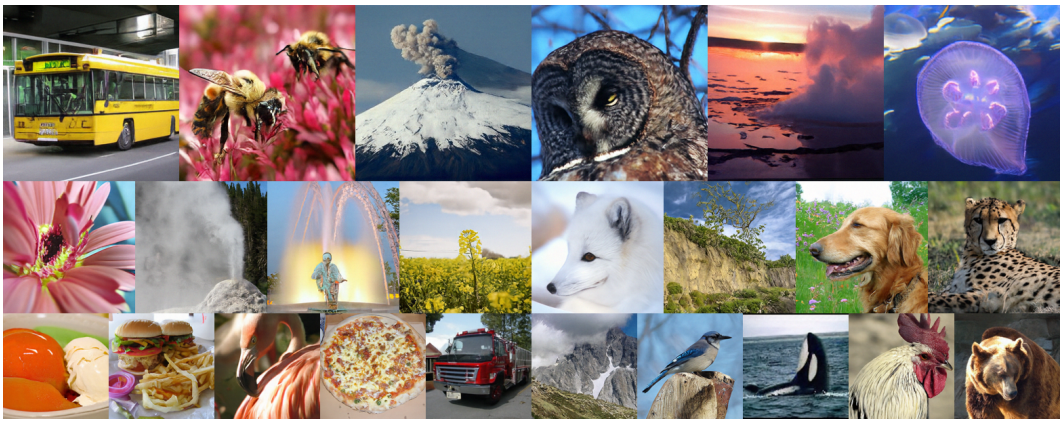

Figure 6: Curated samples generated by **JA$\boldsymbol{x}$** on ImageNet 256 $\times$ 256 using the standard JiT-H architecture and sampling procedure.

Markov coupling. Within the same reliability-gated band $[2, 2.5]$, Markov coupling achieves 4.17 FID, compared with 4.49 for the non-Markov variant, using the same pair-selection rule, teacher configuration, and alignment weight. This comparison supports our choice of Markov-coupled prediction supervision and shows that reliability-gated selection alone is insufficient in the tested setting.

Teacher targets and time reweighting. The combined objective changes both the regression target and the effective time weighting (Appendix A.5). To isolate the teacher target's contribution, the reweighting-only control replays the saved **JA$\boldsymbol{x}$** activation schedule $a_k(t)$ and optimizes $w(t)(1 + \lambda a_k(t))\|f_\theta(U_t) - X\|_2^2$, with initialization, data order, and the base training protocol matched. It yields 4.39 FID, compared with 4.37 for vanilla JiT and 4.17 for **JA$\boldsymbol{x}$** . Thus, time reweighting alone does not reproduce the improvement, supporting the contribution of teacher targets beyond the induced weighting in this setting.

Table 2: Ablations on JiT-B/16 at 200 epochs. RCB denotes the reliability-gated coupling band. Prediction-supervision variants use Markov coupling unless noted. FID uses 50K samples.

Method	$\Delta\ell$	FID ↓
Baseline and controls		
Vanilla JiT	–	4.37
Reweighting only	–	4.39
Fixed gap	2	4.48
Conditional sampling	> 0	4.43
Reliability-gated bands		
RCB	$[2, 4]$	4.29
RCB	$[1.5, 4]$	4.32
RCB	$[2, 3.5]$	4.26
RCB	$[2, 3]$	4.23
RCB (**JA$\boldsymbol{x}$**)	$[2, 2.5]$	**4.17**
Observation coupling		
RCB (Non-Markov)	$[2, 2.5]$	4.49

Summary of the appendix. Appendix A provides proofs and further analysis of Markov coupling, variance reduction, and teacher approximation error. Appendices B–D discuss related methods, implementation details, and limitations. Appendices E–F provide additional quantitative comparisons, cost analysis, and qualitative results.

7 Conclusion

We presented **JA$\boldsymbol{x}$** , a prediction-supervision method for JiT motivated by the observed limitations of hidden representation alignment. Under Markov coupling, oracle clean-image predictions preserve the required conditional mean and reduce conditional target covariance, with a population loss equivalent to JiT's up to a constant. The practical method combines an EMA teacher with direct clean-image supervision and reliability-gated pair selection to address teacher approximation error. Experiments demonstrate improved final FID and faster convergence in training epochs across JiT scales, without external encoders or changes to the architecture and sampling procedure. Ablations show that time reweighting alone and fixed-gap alignment after warmup do not reproduce these gains, supporting the contributions of both teacher targets and reliability-gated selection. Gradient diagnostics further show lower normalized gradient variance and mean-squared deviation from the JiT population gradient. Combining **JA$\boldsymbol{x}$** with PixelREPA also improves FID in the evaluated JiT-B/16 setting.

AI USE STATEMENT

Generative AI tools were used to assist with writing, editing, and proofreading during the preparation of this manuscript. AI was also used to double check the code the authors wrote. All AI-assisted content was reviewed and verified by the authors. The authors are responsible for the final content, methodology, experimental results, and conclusions presented in this paper.

ETHICS STATEMENT

This work does not involve human-subject experiments, user studies, or the collection of personal or sensitive information. The experiments use ImageNet and standard image-generation benchmarks for evaluating representation and generative modeling methods. We use the datasets and pretrained models in accordance with their applicable licenses and usage terms. No additional ethical concerns beyond those associated with the underlying datasets and generative modeling technology are introduced by the proposed training method.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of JAx, including the Markov-coupled prediction objective, reliability-gated coupling band, EMA teacher procedure, training settings, and theoretical assumptions in the paper and the appendix. We report model architectures, optimization settings, evaluation protocols, and ablation configurations for experiments on ImageNet 256×256. We will release the source code, training configurations, and evaluation scripts upon publication to facilitate reproduction of the reported results.

REFERENCES

Alan Baade, Eric Ryan Chan, Kyle Sargent, Changan Chen, Justin Johnson, Ehsan Adeli, and Li Fei-Fei. Latent forcing: Reordering the diffusion trajectory for pixel-space image generation. *arXiv preprint arXiv:2602.11401*, 2026.

Hila Chefer, Patrick Esser, Dominik Lorenz, Dustin Podell, Vikash Raja, Vinh Tong, Antonio Torralba, and Robin Rombach. Self-supervised flow matching for scalable multi-modal synthesis. *arXiv preprint arXiv:2603.06507*, 2026.

Hansheng Chen, Jan Ackermann, Minseo Kim, Gordon Wetzstein, and Leonidas Guibas. Asymmetric flow models. *arXiv preprint arXiv:2605.12964*, 2026.

Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. In *International Conference on Machine Learning (ICML)*, 2025a.

Shoufa Chen, Chongjian Ge, Shilong Zhang, Peize Sun, and Ping Luo. PixelFlow: Pixel-space generative models with flow. *arXiv preprint arXiv:2504.07963*, 2025b.

Giannis Daras, Yuval Dagan, Alexandros G. Dimakis, and Constantinos Daskalakis. Consistent diffusion models: Mitigating sampling drift by learning to be consistent. In *Advances in neural information processing systems (NeurIPS)*, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in neural information processing systems (NeurIPS)*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in neural information processing systems (NeurIPS)*, pp. 6840–6851, 2020.

Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans. Simpler diffusion: 1.5 FID on ImageNet512 with pixel-space diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18062–18071, 2025.

Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. No other representation component is needed: Diffusion transformers can provide representation guidance by themselves. In *International Conference on Learning Representations (ICLR)*, 2026.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in neural information processing systems (NeurIPS)*, pp. 122458–122483, 2024.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for end-to-end tuning with latent diffusion transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18262–18272. IEEE, 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Deyuan Liu, Peng Sun, Xufeng Li, and Tao Lin. Efficient generative model training via embedded representation warmup. *arXiv preprint arXiv:2504.10188*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision (ECCV)*, pp. 23–40. Springer, 2024.

Zehong Ma, Longhui Wei, Shuai Wang, Shiliang Zhang, and Qi Tian. DeCo: Frequency-decoupled pixel diffusion for end-to-end image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43600–43610, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in neural information processing systems (NeurIPS)*, pp. 2234–2242, 2016.

Jaeyo Shin, Jiwook Kim, and Hyunjung Shim. Representation alignment for Just Image Transformers is not easier than you think. In *European Conference on Computer Vision (ECCV)*. Springer, 2026.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? In *International Conference on Learning Representations (ICLR)*, 2026a.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026b.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning (ICML)*, 2023.

Mengmeng Wang, Dengyang Jiang, Liuzhuozheng Li, Yucheng Lin, Guojiang Shen, Xiangjie Kong, Yong Liu, Guang Dai, and Jingdong Wang. SRA 2: Variational autoencoder self-representation alignment for efficient diffusion training. *arXiv preprint arXiv:2601.17830*, 2026.

Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization. *arXiv preprint arXiv:2506.09027*, 2025.

Donglin Yang, Yongxing Zhang, Xin Yu, Liang Hou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Renjie Liao. Stable velocity: A variance perspective on flow matching. *arXiv preprint arXiv:2602.05435*, 2026.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712. IEEE, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations (ICLR)*, 2025.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. PixelDiT: Pixel diffusion transformers for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14273–14282, 2026.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations (ICLR)*, 2026.

A Theoretical Analysis and Proofs

The analysis connects the construction in Section 4 to its training role. Markov coupling preserves the JiT input distribution and the oracle regression target; conditional averaging then reduces target and gradient variability. We also quantify how teacher error enters the practical objective, explaining the roles of direct image supervision and reliability-gated selection. Appendix A.3 gives an analytical comparison of observation couplings, and Appendix B discusses the relationship to prior work.

A.1 Markov Coupling and Martingale Consistency

Standing assumptions and notation. Let $X \in \mathbb{R}^d$ satisfy $\mathbb{E}\|X\|_2^2 < \infty$, and let c denote the conditioning information. For deterministic $u \in (0,1)$, $m_u(z,c)$ is the posterior mean under the corruption $Z_u = uX + (1-u)\epsilon$, with $\epsilon \sim \mathcal{N}(0, I_d)$ independent of (X, c). All primitive Gaussian noise draws below are independent of (X, c) and of each other. Unless otherwise stated, times are fixed; when $(t, s) \sim \pi(t, s) = p(t)q(s \mid t)$, the pair is independent of (X, c) and the noises, and $p(t)$ is the JiT time distribution. Write Var for a covariance matrix, equivalently Cov in the main text, and $\preceq$ for the positive-semidefinite order. All conditional identities hold almost surely. We assume the weighted losses in use are finite. For statements about population gradients and their covariances, we additionally assume the relevant gradients are square-integrable and differentiation can be interchanged with expectation. These requirements matter because $w(t) = (1-t)^{-2}$ is unbounded near $t = 1$.

Well-posedness and preservation of the corruption marginal. For $0 < t, s < 1$, set

$$a_{t|s} = \frac{t}{s}, \qquad b_{t|s}^2 = (1-t)^2 - \frac{t^2}{s^2}(1-s)^2. \tag{16}$$

Then $b_{t|s}^2 \geq 0$ if and only if $t \leq s$. For such pairs, choose $b_{t|s} \geq 0$ and construct $Z_s = sX + (1-s)\epsilon_s$ and $Z_t = a_{t|s}Z_s + b_{t|s}\eta$. This construction has the JiT marginal and satisfies $X \perp Z_t \mid (Z_s, c)$.

Proof. Since all four quantities $t, s, 1-t, 1-s$ are positive,

$$b_{t|s}^2 \geq 0 \quad \Longleftrightarrow \quad s(1-t) \geq t(1-s) \quad \Longleftrightarrow \quad t \leq s.$$

Furthermore,

$$Z_t = tX + \frac{t(1-s)}{s}\epsilon_s + b_{t|s}\eta.$$

The two noise terms are independent, centered Gaussian vectors with total covariance $(1-t)^2 I_d$. Thus $Z_t \mid (X = x, c) \sim \mathcal{N}(tx, (1-t)^2 I_d)$, as required. Given (Z_s, c), the remaining randomness in Z_t comes only from η, which is independent of (X, Z_s, c). This proves the conditional Markov property in Eq. 6. □

Proof of Proposition 1. The Markov property implies $\mathbb{E}[X \mid Z_s, Z_t, c] = m_s(Z_s, c)$. Therefore,

$$\mathbb{E}[m_s(Z_s, c) \mid Z_t, c] = \mathbb{E}[\mathbb{E}[X \mid Z_s, Z_t, c] \mid Z_t, c] = m_t(Z_t, c). \tag{17}$$

Here the nested sigma-fields are $\sigma(Z_t, c)$ and $\sigma(Z_s, Z_t, c)$; neither $\sigma(Z_t, c)$ nor $\sigma(Z_s, c)$ alone is generally contained in the other. For random times, the same argument is first applied conditional on (t, s), giving Eq. 7, and then averaged over $s \mid t$. Preservation of the marginal ensures that conditioning on s does not change the posterior of X given (Z_t, t, c). □

Composition and the meaning of martingale consistency. For $0 < t < s_1 < s_2 < 1$, generate the chain $X \to Z_{s_2} \to Z_{s_1} \to Z_t$ using independent transition noises. The coefficients satisfy

$$a_{t|s_1}a_{s_1|s_2} = a_{t|s_2}, \qquad a_{t|s_1}^2 b_{s_1|s_2}^2 + b_{t|s_1}^2 = b_{t|s_2}^2. \tag{18}$$

Indeed, the first equality follows by cancellation, and the second follows by inserting Eq. 16 and cancelling the two terms containing $(1-s_1)^2$. Thus, after marginalizing out Z_{s_1}, the law of (X, Z_{s_2}, Z_t, c) agrees with the direct two-time construction. The same argument gives compatible

couplings for any finite set $0 < \tau_1 < \cdots < \tau_n < 1$. For this chain, define $\mathcal{F}_i = \sigma(c, Z_{\tau_1}, \ldots, Z_{\tau_i})$. Conditional independence gives $\mathbb{E}[X \mid \mathcal{F}_i] = m_{\tau_i}(Z_{\tau_i}, c)$. Consequently these posterior means form a martingale with respect to $(\mathcal{F}_i)$ as the time index increases toward cleaner observations. This is the precise finite-dimensional sense of the consistency property; it does not assert pointwise equality of the two predictions.

Log-SNR representation. Define $\ell_u = 2\log(u/(1-u))$ and $\xi_u = (Z_u - uX)/(1-u)$. The coupling yields

$$\xi_t = r(t,s)\xi_s + \sqrt{1 - r(t,s)^2}\,\eta, \qquad r(t,s) = \frac{t(1-s)}{s(1-t)} = \exp\left(-\frac{\ell_s - \ell_t}{2}\right). \tag{19}$$

In particular, $\mathrm{Cov}(\xi_t, \xi_s) = r(t,s) I_d$. This proves Eq. 12; the stated correlation is between the standardized corruption noises, rather than between the complete observations.

A.2 Objective Equivalence and Variance Reduction

For fixed $t < s$, abbreviate

$$M_s = m_s(Z_s, c), \qquad C_t = \mathrm{Var}(X \mid Z_t, c),$$
$$V_t(s) = \mathrm{Var}(M_s \mid Z_t, c), \quad D_t(s) = \mathbb{E}[\mathrm{Var}(X \mid Z_s, c) \mid Z_t, c].$$

Here $D_t(s)$ is the covariance removed by the oracle target, whereas $V_t(s)$ is the covariance that remains.

Conditional squared-loss decomposition. Let $U = (Z_t, t, c)$ and let Y be a square-integrable target satisfying $\mathbb{E}[Y \mid U] = m_t(Z_t, c)$. Since $f_\theta(U)$ is U-measurable,

$$\mathbb{E}[w(t)\|f_\theta(U) - Y\|_2^2 \mid U] = w(t)\|f_\theta(U) - m_t(Z_t, c)\|_2^2 + w(t)\,\mathrm{tr}\,\mathrm{Var}(Y \mid U). \tag{20}$$

To verify this identity, expand the square around $m_t(Z_t, c)$. The cross term is zero because $\mathbb{E}[Y - m_t(Z_t, c) \mid U] = 0$.

Proof of the objective equivalence in Eq. 9. Proposition 1 gives the required conditional mean both for $Y = X$ and for $Y = M_s$, including when the times are sampled from π. Applying Eq. 20 first at fixed (t, s) and then averaging over π proves

$$\mathcal{L}^\star_{\mathrm{JAx}}(\theta) = \mathcal{L}_{\mathrm{JiT}}(\theta) + K, \qquad K = \mathbb{E}_\pi[w(t)\,\mathrm{tr}\{V_t(s) - C_t\}]. \tag{21}$$

Expectations with subscript π also include the data and coupled noises. The common marginal $p(t)$ and common weight $w(t)$ are essential here. The constant K is independent of θ, so the two population objectives have the same parameter-space minimizers and, wherever differentiable, the same gradients. This statement concerns the oracle objective in Eq. 8, before the sample-dependent activation in Eq. 15.

Proof of Proposition 2: covariance identity and monotonicity. The conditional law of total covariance, using the nested information (Z_t, c) and (Z_s, Z_t, c), gives

$$\begin{aligned} C_t &= \mathbb{E}[\mathrm{Var}(X \mid Z_s, Z_t, c) \mid Z_t, c] + \mathrm{Var}(\mathbb{E}[X \mid Z_s, Z_t, c] \mid Z_t, c) \\ &= D_t(s) + V_t(s). \end{aligned} \tag{22}$$

The second equality uses the Markov property in both terms. Since $D_t(s) \succeq 0$, this proves Eq. 10. It also gives the explicit constant

$$K = -\mathbb{E}_\pi[w(t)\,\mathrm{tr}\,D_t(s)] = -\mathbb{E}_\pi[w(t)\|X - M_s\|_2^2] \le 0. \tag{23}$$

For monotonicity, use the compatible chain $X \to Z_{s_2} \to Z_{s_1} \to Z_t$ from Eq. 18, where $t < s_1 < s_2$. The Markov property and Proposition 1 imply

$$\mathbb{E}[M_{s_2} \mid Z_{s_1}, Z_t, c] = M_{s_1}.$$

Applying conditional total covariance to M_{s_2} now gives

$$V_t(s_2) - V_t(s_1) = \mathbb{E}[\mathrm{Var}(M_{s_2} \mid Z_{s_1}, c) \mid Z_t, c] \succeq 0. \tag{24}$$

All pair marginals in this calculation agree with Eq. 5. Therefore $V_t(s)$ is nondecreasing and $D_t(s) = C_t - V_t(s)$ is nonincreasing in s, in the positive-semidefinite order. At fixed t, the same ordering holds with respect to the log-SNR gap. □

Endpoint limits in Proposition 2. Under the standing second-moment assumption, the precise limits are

$$\begin{aligned} s \downarrow t: &\quad V_t(s) \longrightarrow 0, \quad D_t(s) \longrightarrow C_t, \\ s \uparrow 1: &\quad V_t(s) \longrightarrow C_t, \quad D_t(s) \longrightarrow 0. \end{aligned} \tag{25}$$

These are L^1 limits in matrix trace norm under the common student marginal of (Z_t, c). The corresponding oracle targets converge in mean square to $m_t(Z_t, c)$ and X, respectively, under the pair coupling.

Proof. For $s \uparrow 1$, the posterior mean minimizes squared error among functions of (Z_s, c). Comparing it with Z_s/s gives

$$\mathbb{E}\,\mathrm{tr}\, D_t(s) = \mathbb{E}\|X - M_s\|_2^2 \le \mathbb{E}\|X - Z_s/s\|_2^2 = d\left(\frac{1-s}{s}\right)^2 \longrightarrow 0. \tag{26}$$

Because $D_t(s)$ is positive semidefinite, its trace is its trace norm; Eq. 22 gives both limits at this endpoint.

For $s \downarrow t$, realize all two-time pairs using the same (X, c, ϵ, η), by setting $Z_s = sX + (1-s)\epsilon$ and $Z_t^{(s)} = (t/s)Z_s + b_{t|s}\eta$. Both observations converge almost surely to $Z_t^0 = tX + (1-t)\epsilon$. For almost every c, the Gaussian likelihood gives the continuous version

$$m_u(z, c) = \frac{\int x \exp\{-\|z - ux\|_2^2/[2(1-u)^2]\}\, P_c(dx)}{\int \exp\{-\|z - ux\|_2^2/[2(1-u)^2]\}\, P_c(dx)}, \qquad 0 < u < 1, \tag{27}$$

where P_c is the conditional law of X given c. The denominator is strictly positive, and dominated convergence applies to the numerator and denominator because $\int \|x\|_2\, P_c(dx) < \infty$. Thus $M_s - m_t(Z_t^{(s)}, c) \to 0$ almost surely.

For completeness, this convergence is also in mean square. Put $Q = \|X\|_2^2$. Conditional Jensen gives $\|M_s\|_2^2 \le \mathbb{E}[Q \mid Z_s, c]$, and similarly for $\|m_t(Z_t^{(s)}, c)\|_2^2$. The family of conditional expectations of a fixed integrable, nonnegative Q is uniformly integrable: for any sigma-field $\mathcal{G}$, $H = \mathbb{E}[Q \mid \mathcal{G}]$, and $A = \{H > L\}$,

$$\mathbb{E}[H\mathbf{1}_A] = \mathbb{E}[Q\mathbf{1}_A] \le \mathbb{E}[Q\mathbf{1}_{\{Q>M\}}] + M\,\mathbb{E}[Q]/L.$$

Taking first $L \to \infty$ and then $M \to \infty$ proves the assertion uniformly in $\mathcal{G}$. Hence the squared differences above are uniformly integrable and their expectations tend to zero. Finally,

$$\mathbb{E}\,\mathrm{tr}\, V_t(s) = \mathbb{E}\|M_s - m_t(Z_t^{(s)}, c)\|_2^2 \longrightarrow 0.$$

Each $Z_t^{(s)}$ has the same student marginal, so this is the claimed L^1 covariance limit. Eq. 22 gives the remaining limit for $D_t(s)$. □

A.3 An Analytical Comparison of Observation Couplings

A scalar Gaussian example isolates the role of coupling without teacher approximation or optimization effects. Let $X \sim \mathcal{N}(0, \tau^2)$, with $\tau > 0$, and fix $0 < t < s < 1$. Construct $Z_u = uX + (1-u)\epsilon_u$ for $u \in \{t, s\}$, where (ϵ_t, ϵ_s) is jointly Gaussian, independent of X, with standard normal marginals and correlation ρ. All choices of ρ preserve the individual corruption marginals. Define

$$Q_u = u^2\tau^2 + (1-u)^2, \qquad C_{ts} = ts\tau^2 + (1-t)(1-s)\rho, \qquad m_u(z) = \frac{u\tau^2}{Q_u}z. \tag{28}$$

Gaussian conditioning gives the oracle target's conditional mean and variance:

$$\mathbb{E}[m_s(Z_s) \mid Z_t = z] = \frac{s\tau^2 C_{ts}}{Q_s Q_t}z, \qquad \mathrm{Var}(m_s(Z_s) \mid Z_t) = \left(\frac{s\tau^2}{Q_s}\right)^2\left(Q_s - \frac{C_{ts}^2}{Q_t}\right). \tag{29}$$

Subtracting $m_t(z)$ from the conditional mean yields

$$\mathbb{E}[m_s(Z_s) \mid Z_t = z] - m_t(z) = \frac{\tau^2(1-s)\{s(1-t)\rho - t(1-s)\}}{Q_s Q_t}z. \tag{30}$$

Table 3: Analytical target statistics for $X \sim \mathcal{N}(0,1)$, $t = 1/3$, and $s = 1/2$. The required conditional mean is $m_t(z) = 0.6z$. These values are exact Gaussian calculations.

Target / coupling	ρ	$\mathbb{E}[Y \mid Z_t = z]$	$\mathrm{Var}(Y \mid Z_t)$
Clean image X	–	$0.6z$	0.80
Oracle, independent noise	0	$0.3z$	0.45
Oracle, shared noise	1	$0.9z$	0.05
Oracle, Markov coupling	0.5	$0.6z$	0.30

Thus, within this family, conditional consistency holds precisely at $\rho = t(1-s)/(s(1-t))$, the Markov correlation in Eq. 12. Independent noise ($\rho = 0$) and shared noise ($\rho = 1$) both shift the conditional mean for nonzero z.

Table 3 illustrates why the two target criteria in Section 3.2 are needed together. Shared-noise supervision has the smallest conditional variance in this example, but changes the conditional regression target. Markov coupling preserves that target while reducing its variance relative to clean-image supervision. This provides a concrete explanation for the coupling design; the JiT ablation in Table 2 evaluates its practical effect with an EMA teacher.

A.4 Implications for Stochastic Gradients

Fix the current student parameters θ, and hold every target fixed when differentiating. Let $U = (Z_t, t, c)$ and $J_\theta(U) = \partial f_\theta(U)/\partial\theta \in \mathbb{R}^{d\times p}$, where p is the number of parameters. For a stop-gradient target Y,

$$g(Y) = 2w(t)J_\theta(U)^\top(f_\theta(U) - Y). \tag{31}$$

We use this expression to translate the target covariance reduction in Proposition 2 into a corresponding reduction in stochastic-gradient covariance, and then relate it to the normalized gradient statistic evaluated in Figure 5.

At fixed (t, s), the matrix $J_\theta(U)$ is fixed conditional on U. Thus, writing $J = J_\theta(U)$ and $w = w(t)$,

$$\begin{aligned}\mathbb{E}[g(M_s) \mid U] &= \mathbb{E}[g(X) \mid U] = 2wJ^\top(f_\theta(U) - m_t(Z_t, c)),\\ \mathrm{Var}(g(Y) \mid U) &= 4w^2 J^\top \mathrm{Var}(Y \mid U)J.\end{aligned} \tag{32}$$

Consequently, Proposition 2 gives

$$\mathrm{Var}(g(X) \mid U) - \mathrm{Var}(g(M_s) \mid U) = 4w^2 J^\top D_t(s) J \succeq 0. \tag{33}$$

Thus the oracle prediction target preserves the expected gradient while reducing its conditional covariance. For random pairs, the same conclusion follows by conditioning first on (U, s) and then on U, since the conditional mean above does not depend on s. The corresponding unconditional ordering then follows from the law of total covariance.

The normalized combined oracle gradient. Figure 5 compares the combined objective with JiT after dividing its gradient by $1 + \lambda$. For an active pair, the corresponding oracle quantity is

$$G^\star_\lambda = \frac{g(X) + \lambda g(M_s)}{1+\lambda}, \qquad \lambda \geq 0. \tag{34}$$

The two gradients share (X, Z_s, Z_t) and are generally correlated. Set $R = X - M_s$. Since $\mathbb{E}[R \mid Z_s, Z_t, c] = 0$, R is conditionally orthogonal to the oracle prediction fluctuation $M_s - m_t(Z_t, c)$, and $\mathbb{E}[RR^\top \mid U] = D_t(s)$. It follows that

$$\mathrm{Var}(G^\star_\lambda \mid U) = 4w^2 J^\top \left[V_t(s) + \frac{D_t(s)}{(1+\lambda)^2}\right] J, \tag{35}$$

$$\mathrm{Var}(g(X) \mid U) - \mathrm{Var}(G^\star_\lambda \mid U) = 4w^2 \left(1 - \frac{1}{(1+\lambda)^2}\right) J^\top D_t(s) J \succeq 0. \tag{36}$$

The identity follows from $g(X) = g(M_s) - 2wJ^\top(X - M_s)$ and the conditional orthogonality above. The normalized combined oracle gradient therefore has the same conditional mean as the JiT

gradient and no larger conditional covariance. The same conclusion extends to random pairs and to unconditional covariances as above. For a mean of B independent, identically distributed sample gradients at fixed parameters θ, each covariance is divided by B. No such variance ordering against JiT is implied for the *unnormalized* combined gradient.

Empirical evidence with an EMA teacher. Figure 5 tests the same variance-reduction mechanism using a frozen EMA teacher at the evaluated student checkpoints. Normalizing the combined gradient by $1 + \lambda$ removes the overall scale difference. The resulting 9.6% average variance reduction is accompanied by a 6.5% reduction in mean-squared deviation from the corresponding JiT population gradient (averaged over epoch 200 and 600). The latter accounts for both gradient variance and a possible mean-gradient shift, so the observed reduction is not explained by variance alone.

The oracle analysis above identifies the gradient-level mechanism induced by the Markov-coupled prediction target. The EMA measurements exhibit the same variance-reduction behavior in the evaluated practical training setting. Appendix A.5 further characterizes how teacher approximation error enters this objective and how reliability gating and direct JiT supervision control its effect.

A.5 EMA Approximation and Reliability Gating

We now characterize the practical construction used by **JA$\boldsymbol{x}$** , including the EMA teacher, reliability-gated pair selection, and direct clean-image supervision.

Condition throughout this subsection on the training history $\mathcal{H}_k$ before the current batch. The student parameters, EMA parameters, and risk estimates based on previous updates are then fixed. Assume the current data and corruption noises are freshly drawn from the specified training distribution, and that pair selection uses only the student time and $\mathcal{H}_k$, as in Section 4.4. All expectations below are conditional on $\mathcal{H}_k$, which is suppressed in the notation. The teacher and selection rule are held fixed when differentiating the current loss; no derivative is taken through their training history.

Feasibility and the nearest admissible teacher. Assume $0 < \Delta_{\min} \le \Delta_{\max} < \infty$ and $\ell_{\max} < \infty$. Given a finite frontier $\ell_{\mathrm{safe}}(k)$, define the frontier-based admissible set

$$\mathcal{I}_k(t) = [\ell_t + \Delta_{\min}, \ell_t + \Delta_{\max}] \cap [\ell_{\mathrm{safe}}(k), \ell_{\max}]. \tag{37}$$

Put

$$L_k(t) = \max\{\ell_t + \Delta_{\min}, \ell_{\mathrm{safe}}(k)\}, \qquad U_k(t) = \min\{\ell_t + \Delta_{\max}, \ell_{\max}\}.$$

Then $\mathcal{I}_k(t)$ is nonempty if and only if $L_k(t) \le U_k(t)$, which is exactly the activation test in Section 4.4. Whenever active, $\ell_s^\star = L_k(t)$ is its smallest element and $s^\star = \mathrm{sigmoid}(\ell_s^\star/2)$ satisfies $t < s^\star < 1$. By Eq. 24, this choice maximizes the oracle covariance reduction $D_t(s)$ within $\mathcal{I}_k(t)$. If the frontier is unavailable or the set is empty, define $a_k(t) = 0$; otherwise set $a_k(t) = 1$.

The frontier is defined from the empirical prediction-risk curve and serves as the practical reliability criterion. For a nonmonotone empirical curve $R_k(\ell)$, the definition

$$\ell_{\mathrm{safe}}(k) = \inf\{\ell : R_k(\ell) \le R^\star\} + \delta_{\mathrm{safe}}$$

does not by itself imply $R_k(\ell) \le R^\star$ for every $\ell \ge \ell_{\mathrm{safe}}(k)$. Such a statement would require additional regularity of the risk curve together with an estimation guarantee. The margin δ_{safe} is therefore a design margin rather than a finite-sample confidence bound.

Conditional bias of an imperfect teacher. For an active pair, write

$$\bar{x}_s = M_s + e_s, \qquad e_s = f_{\bar{\theta}}(Z_s, s, c) - m_s(Z_s, c), \qquad \beta_{t,s}(U) = \mathbb{E}[e_s \mid U].$$

Here $s = s^\star(t)$ is fixed given the student time and history. Proposition 1 gives

$$\mathbb{E}[\bar{x}_s \mid U] = m_t(Z_t, c) + \beta_{t,s}(U).$$

Let $h = f_\theta(U) - m_t(Z_t, c)$. Direct expansion, retaining the conditional expectation in every random term, yields

$$\begin{aligned}\mathbb{E}[w\|f_\theta(U) - \bar{x}_s\|_2^2 \mid U] = w\|h\|_2^2 - 2w\langle h, \beta_{t,s}(U)\rangle \\ + w\mathbb{E}[\|\bar{x}_s - m_t(Z_t, c)\|_2^2 \mid U].\end{aligned} \tag{38}$$

The last term is independent of the current student parameters under the frozen-teacher convention. Hence

$$\mathbb{E}[g(\bar{x}_s) - g(X) \mid U] = -2wJ^\top \beta_{t,s}(U). \tag{39}$$

Thus the conditional mean of the teacher error directly determines the shift in the expected gradient.

Teacher approximation also enters the target covariance:

$$\begin{aligned}\mathrm{Var}(\bar{x}_s \mid U) = V_t(s) + \mathrm{Var}(e_s \mid U) + \mathrm{Cov}(M_s, e_s \mid U) \\ + \mathrm{Cov}(e_s, M_s \mid U).\end{aligned} \tag{40}$$

This decomposition separates the covariance of the oracle target from the additional teacher-error terms. It makes explicit the two aspects controlled by the practical construction: the conditional mean shift is governed by $\beta_{t,s}(U)$, while the remaining variability depends on the magnitude and correlation structure of the teacher error.

Prediction risk as a measure of teacher reliability. Clean-image prediction error provides an empirically estimable proxy for teacher approximation error. The Markov property gives $\mathbb{E}[X - M_s \mid Z_s, Z_t, c] = 0$, while e_s is measurable with respect to (Z_s, c). Consequently,

$$\mathbb{E}[\|\bar{x}_s - X\|_2^2 \mid U] = \mathbb{E}[\|e_s\|_2^2 \mid U] + \operatorname{tr} D_t(s), \qquad \|\beta_{t,s}(U)\|_2^2 \le \mathbb{E}[\|e_s\|_2^2 \mid U]. \tag{41}$$

Define the population prediction risk at $s = \mathrm{sigmoid}(\ell/2)$ by

$$\mathcal{R}_k(\ell) = \mathbb{E}\|f_{\bar{\theta}}(Z_s, s, c) - X\|_2^2.$$

For each fixed selected pair,

$$\mathbb{E}\|\beta_{t,s}(U)\|_2^2 \le \mathbb{E}\|e_s\|_2^2 = \mathcal{R}_k(\ell_s) - \mathbb{E}\|X - M_s\|_2^2 \le \mathcal{R}_k(\ell_s). \tag{42}$$

Thus small clean-image prediction risk controls the average squared conditional bias of the teacher target, providing a direct motivation for the reliability criterion used in Section 4.4.

The bound concerns the population risk and controls the bias on average, rather than pointwise for every input. The online quantity R_k estimates $\mathcal{R}_k$ and should not be substituted for it in an exact inequality without an estimation guarantee. If the implemented risk is averaged over the d coordinates, the corresponding population quantity above is scaled by d. A numerical gradient-bias bound additionally depends on $w(t)$ and the network Jacobian, as indicated by Eq. 39.

The mixed target and time weighting. For a fixed teacher and selection rule, let $q_k(t) = \lambda a_k(t)$. On an active pair, completing the square gives

$$Y_{\mathrm{mix},k} = \frac{X + q_k(t)\bar{x}_{s^\star(t)}}{1 + q_k(t)}, \tag{43}$$

$$\begin{aligned}\ell_{\mathrm{total},k} &= w(t)\|f_\theta(U) - X\|_2^2 + w(t)q_k(t)\|f_\theta(U) - \bar{x}_{s^\star(t)}\|_2^2 \\ &= w(t)(1 + q_k(t))\|f_\theta(U) - Y_{\mathrm{mix},k}\|_2^2 + \frac{w(t)q_k(t)}{1 + q_k(t)}\|X - \bar{x}_{s^\star(t)}\|_2^2.\end{aligned} \tag{44}$$

The final term is independent of θ under the frozen-teacher convention. For inactive times, set $Y_{\mathrm{mix},k} = X$ and the final term to zero, without evaluating a teacher.

The alignment therefore changes both the regression target and the effective time weight, from $w(t)$ to $w(t)(1 + \lambda a_k(t))$. This identity motivates the reweighting-only control in Table 2: replaying the same activation schedule while regressing only to X isolates the contribution of the teacher target from the induced change in weighting.

Bias–variance trade-off with an imperfect teacher. For an active pair, let $\alpha = \lambda/(1 + \lambda)$, so that $Y_{\mathrm{mix},k} = (1 - \alpha)X + \alpha(M_s + e_s)$. Write $m_t = m_t(Z_t, c)$ and define

$$v = \operatorname{tr} V_t(s), \qquad \delta = \operatorname{tr} D_t(s), \qquad e_2 = \mathbb{E}[\|e_s\|_2^2 \mid U], \qquad b = \mathbb{E}[\langle M_s - m_t, e_s\rangle \mid U].$$

Since $X - M_s$ has zero conditional mean given (Z_s, Z_t, c), it is conditionally orthogonal to both $M_s - m_t$ and e_s. Expanding the mixed target's error gives the exact identity

$$\mathbb{E}[\|Y_{\mathrm{mix},k} - m_t\|_2^2 - \|X - m_t\|_2^2 \mid U] = -\alpha(2 - \alpha)\delta + 2\alpha b + \alpha^2 e_2. \tag{45}$$

The first term is the oracle variance benefit, while the remaining terms quantify teacher approximation error and its correlation with the oracle prediction. This decomposition makes explicit the trade-off underlying the practical rule: nearby states provide a larger oracle variance reduction, whereas prediction accuracy controls the additional approximation terms.

By $|b| \le \sqrt{ve_2}$, a sufficient condition for the mixed target to have no greater conditional squared error around m_t than X is

$$2\sqrt{ve_2} + \alpha e_2 \le (2-\alpha)\delta. \tag{46}$$

This condition is sufficient rather than necessary and is not used directly by the pair-selection rule. It illustrates one regime in which the oracle variance benefit dominates teacher approximation error. The same decomposition applies to conditional gradient deviation by using $\|J^\top z\|_2^2$ in place of $\|z\|_2^2$ and multiplying by $4w^2$.

The actual gated objective and its minimizer. Since Eq. 15 divides by the full batch size, its conditional population counterpart is

$$\mathcal{L}_{\mathrm{align},k}(\theta) = \mathbb{E}[a_k(t)w(t)\|f_\theta(U) - \bar{x}_{s^\star(t)}\|_2^2]. \tag{47}$$

For inactive times the integrand is defined to be zero without requiring a teacher state. Set $\beta_k(U) = \beta_{t,s^\star(t)}(U)$ on active times and zero otherwise. Combining Eq. 38 with the JiT term gives, up to a term independent of θ,

$$\begin{aligned}
&\mathcal{L}_{\mathrm{JiT}} + \lambda\mathcal{L}_{\mathrm{align},k} \\
&\quad = \mathbb{E}\big[w(t)(1+\lambda a_k(t))\|f_\theta(U) - m_t(Z_t,c)\|_2^2\big] \\
&\qquad - 2\lambda\mathbb{E}[a_k(t)w(t)\langle f_\theta(U) - m_t(Z_t,c), \beta_k(U)\rangle] + \mathrm{const}.
\end{aligned} \tag{48}$$

Completing the square shows that the pointwise optimal predictor over unrestricted square-integrable functions is

$$f_k^{\mathrm{opt}}(U) = m_t(Z_t,c) + \frac{\lambda a_k(t)}{1+\lambda a_k(t)}\,\beta_k(U). \tag{49}$$

This identity makes the role of direct JiT supervision explicit. On an active pair, the conditional mean shift is attenuated from $\beta_k(U)$ under teacher-only regression to

$$\frac{\lambda}{1+\lambda}\beta_k(U)$$

under the combined objective. For the experimental setting $\lambda = 0.5$, the attenuation factor is $1/3$, so the shift has one third of the teacher-only magnitude and one ninth of its squared magnitude.

The expression above describes the frozen-teacher conditional regression optimum; the trained network additionally reflects its function class and optimization. Even with an oracle teacher ($\beta_k = 0$), gating changes the quadratic weight from $w(t)$ to $w(t)(1+\lambda a_k(t))$. The unrestricted Bayes predictor remains m_t, while in a restricted network class the parameter-space optimum can depend on this weighting. Without gating and with an oracle teacher, Eq. 21 recovers the exact identity

$$\mathcal{L}_{\mathrm{JiT}} + \lambda\mathcal{L}^\star_{\mathrm{JAx}} = (1+\lambda)\mathcal{L}_{\mathrm{JiT}} + \lambda K.$$

Together, these results clarify the complementary roles of the three design choices. Markov coupling supplies the conditionally consistent low-variance target, reliability-gated pair selection balances the oracle preference for nearby states with empirical teacher accuracy, and direct clean-image supervision attenuates residual teacher bias. The practical EMA construction therefore follows the same prediction-supervision principle identified by the oracle analysis.

B Relation to Existing Supervision Methods

Comparison with representation-based supervision. Figure 7 compares **JAx** with four representative strategies for augmenting generative training. Auxiliary representation learning introduces self-supervised tasks, such as masked reconstruction, while external representation alignment uses a pretrained encoder to supervise hidden features. SRA replaces the external encoder with an EMA

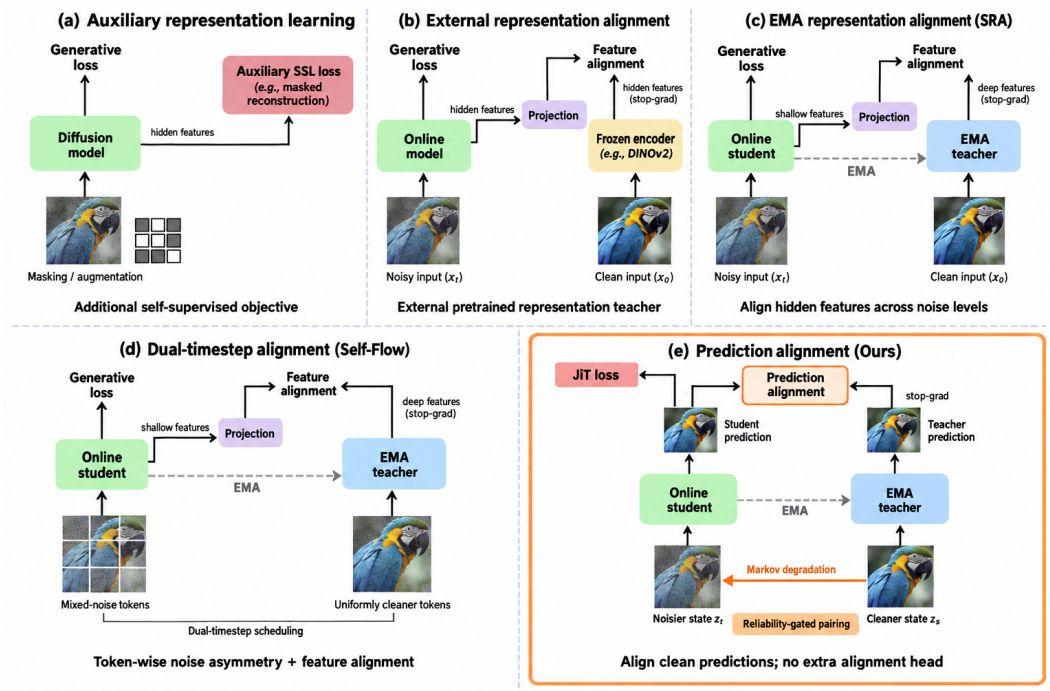

Figure 7: **Comparison of auxiliary supervision strategies for generative modeling.** (a) Auxiliary self-supervised learning, such as masked reconstruction. (b) Hidden-feature alignment with a frozen pretrained encoder. (c) SRA: alignment between shallow student and deeper EMA-teacher features. (d) Self-Flow: feature alignment with dual-timestep noise scheduling. (e) **JA$\boldsymbol{x}$** : direct clean-prediction alignment under reliability-gated Markov coupling, without an additional trainable alignment head. All methods retain the base generative objective. Dashed arrows indicate EMA updates, and teacher targets are stop-gradient. Inputs are schematic; conditioning, latent encoding, and ground-truth target connections are omitted for clarity.

teacher and aligns shallow student features with deeper teacher features. Self-Flow further introduces dual-timestep scheduling, using mixed-noise student inputs and uniformly cleaner teacher inputs for feature alignment. **JA$\boldsymbol{x}$** instead applies alignment directly to clean predictions, using reliability-gated Markov coupling to construct teacher–student pairs. This formulation requires neither an external representation encoder nor an additional trainable alignment head, and avoids imposing an auxiliary matching constraint on hidden representations. Since prediction alignment and representation alignment act at different levels, they are complementary and can be combined.

Stochastic and trajectory consistency. Consistent Diffusion Models (Daras et al., 2023) enforce a martingale property under the model's learned reverse process to mitigate sampling drift. **JA$\boldsymbol{x}$** constructs teacher–student observations directly through an explicit forward Markov coupling $X \to Z_s \to Z_t$, preserving the JiT corruption marginals without simulating model-generated trajectories. This coupling makes the cleaner oracle prediction a valid regression target for the noisier student: its conditional expectation equals the student's desired posterior mean. The resulting oracle alignment objective therefore has the same population gradient as JiT, with reduced conditional target covariance. Consistency Models (Song et al., 2023) instead learn a common endpoint prediction along each probability-flow ODE trajectory for one- or few-step generation. Our stochastic coupling permits oracle predictions to differ across individual pairs while preserving their conditional expectation. This distinction enables auxiliary prediction supervision within the existing JiT training and sampling framework.

Low-variance prediction targets. Stable Velocity (Yang et al., 2026) introduces StableVM, which reduces target variance through a weighted aggregation of conditional velocities over mul-

Algorithm 1 One training update of **JA$\boldsymbol{x}$**

Require: Student θ, EMA teacher $\bar{\theta}$, EMA decay ρ, previous risk estimates R_k
1: Draw a batch $\{(X_i, c_i)\}_{i=1}^{B}$ and student times $t_i \sim p(t)$.
2: Determine activation using Eqs. 13–14 and the feasibility conditions; set $s_i^\star = s_k^\star(t_i)$ for active samples.
3: **for** $i = 1, \ldots, B$ **do**
4: **if** $a_k(t_i) = 1$ **then**
5: Draw $(Z_{s_i^\star}, Z_{t_i})$ using Eq. 5; set $\bar{x}_{s_i^\star} = \mathrm{sg}[f_{\bar{\theta}}(Z_{s_i^\star}, s_i^\star, c_i)]$.
6: **else**
7: Draw $Z_{t_i} = t_i X_i + (1 - t_i)\epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, I)$.
8: **end if**
9: **end for**
10: Update θ using the batch JiT loss plus $\lambda \widehat{\mathcal{L}}_{\mathrm{JA}x,k}$.
11: Update $\bar{\theta} \leftarrow \rho\bar{\theta} + (1 - \rho)\theta$.
12: At calibration intervals, refresh the risk estimates for subsequent updates.

tiple reference data points. **JA$\boldsymbol{x}$** obtains its oracle target by posterior conditioning on a coupled, cleaner observation of the same image. The Markov construction jointly establishes conditional-mean correctness and identifies the exact covariance removed: the expected posterior uncertainty remaining at the teacher observation. Both properties matter, since a low-variance target can still supervise the wrong conditional mean, as illustrated by the Gaussian example in Appendix A.3. The analysis also informs teacher-time selection: closer oracle targets provide greater variance reduction, whereas practical EMA targets must be sufficiently reliable. Our reliability-gated coupling band operationalizes this trade-off using an empirical reliability frontier and bounded log-SNR separation, retaining JiT-only supervision when no feasible pair exists. The contribution thus connects an explicit coupling, a posterior-based target construction, and reliability-aware pair selection into a practical prediction-alignment method.

C Implementation Details

Algorithm 1 summarizes a training update. Table 4 summarizes the backbone architectures, training configurations, alignment settings, and evaluation protocol.

JiT-S/4. For the ImageNet 32×32 experiments in Figure 1, we construct JiT-S/4 from JiT-B by halving the hidden dimension from 768 to 384 and using 4×4 patches.

D Limitation and Future Work

Our evaluation focuses on class-conditional image generation, leaving large-scale text-to-image and video generation unexplored. We expect **JA$\boldsymbol{x}$** to be effective in these settings as well, since its prediction alignment mechanism does not rely on an external pretrained representation encoder. This property may be particularly valuable for video generation, where higher-dimensional spatiotemporal representations can intensify competition for model capacity between representation alignment and denoising. By aligning clean predictions directly, **JA$\boldsymbol{x}$** avoids introducing an additional hidden-representation matching objective. Moreover, **JA$\boldsymbol{x}$** is complementary to hidden-representation alignment methods and can be used together with them, offering a promising direction for future work. Our reliability frontier and several hyperparameter choices are empirically determined. In the future, we plan to systematically study how to better translate the benefits of oracle targets to practical EMA teachers and develop more principled strategies for hyperparameter selection.

E Generation Quality, Convergence, and Training Cost

Sustained FID improvements across scales. Table 5 reports results on ImageNet 256×256 after 600 training epochs. With unchanged student architectures and sampling procedures, **JA$\boldsymbol{x}$** reduces FID from 3.66 to 3.41 on B/16, from 2.36 to 2.26 on L/16, and from 1.86 to 1.82 on H/16. Notably,

Table 4: Implementation details of **JA$\boldsymbol{x}$** on ImageNet 256×256. The backbone architectures and base training configurations follow JiT. The alignment weight is shared across model scales. Training-teacher EMA settings are listed separately from the EMA candidates used for generation evaluation.

Configuration	**JA$\boldsymbol{x}$** -B/16	**JA$\boldsymbol{x}$** -L/16	**JA$\boldsymbol{x}$** -H/16
Architecture			
Depth	12	24	32
Hidden dimension	768	1024	1280
Attention heads	12	16	16
Image size	256×256		
Patch size	16×16		
Bottleneck dimension	128	128	256
Dropout	0	0	0.2
In-context class tokens	32		
In-context start block	4	8	10
Prediction parameterization	$\boldsymbol{x}$-prediction		
Base training			
Training epochs	600		
Training hardware	8 NVIDIA RTX Pro 6000 or 8 NVIDIA A100 GPUs		
Optimizer	AdamW, $(\beta_1, \beta_2) = (0.9, 0.95)$		
Batch size	1024		
Learning rate	2×10^{-4}		
Learning-rate schedule	Constant after warmup		
Warmup epochs	5		
Weight decay	0		
Student time distribution	$\mathrm{logit}(t) \sim \mathcal{N}(-0.8, 0.8^2)$		
Noise scale	1.0		
Minimum division denominator	0.05		
Class-conditioning dropout	0.1		
Prediction alignment and pair selection			
Alignment weight λ	0.5		
Teacher EMA decay	0.9996		
Teacher initialization	Copy of the initial student parameters		
Observation coupling	Markov coupling, Eq. 5		
Minimum log-SNR gap $\Delta_{\min}$	2.0		
Maximum log-SNR gap $\Delta_{\max}$	2.5		
Maximum teacher log-SNR $\ell_{\max}$	$2\log(0.95/0.05) \approx 5.889$		
Admissible prediction risk $R^\star$	0.0155		
Safety margin δ_{safe}	0.25		
Risk-estimation grid spacing	0.25		
Calibration update interval	Every epoch		
Alignment normalization	Full batch size B		
Infeasible pairs	JiT supervision only		
Sampling and evaluation			
ODE solver	Heun		
ODE steps	50		
Sampling timesteps	Uniformly spaced in $[0, 1]$		
CFG interval	$[0.1, 1.0]$		
Final CFG candidates (600 epochs)	$1.0 - 4.0$		
Evaluation EMA candidates	$\{0.9996, 0.9998, 0.9999\}$		
Samples for CFG/EMA selection	10,000 for sweeping; 50,000 for final evaluation		
Generation batch size per GPU	128		
Samples for final evaluation	50,000		
Evaluation metrics	FID $\downarrow$, IS $\uparrow$		

JA$\boldsymbol{x}$ -H/16 matches the FID of the larger JiT-G/16 model (1.82), using approximately half as many parameters (953M versus 2B).

Table 5: Results on ImageNet 256×256 after 600 epochs, with unchanged JiT architectures and sampling procedures. Bold marks the lower FID at each scale.

Scale	FID ↓		IS ↑	
	JiT	**JA$\boldsymbol{x}$**	JiT	**JA$\boldsymbol{x}$**
B/16	3.66	**3.41**	275.1	271.82
L/16	2.36	**2.26**	298.5	284.93
H/16	1.86	**1.82**	303.4	302.8

Accounting for the practical overhead. The additional training workload consists of stop-gradient teacher evaluation on active samples, elementwise EMA updates, and periodic risk calibration. The teacher workload scales with the active fraction, while full-batch normalization keeps the loss weight of each active target fixed. The student architecture and sampling procedure remain unchanged, so **JA$\boldsymbol{x}$** introduces no additional inference cost.

For an approximate compute accounting, let F denote the cost of a full-batch student forward pass and count a backward pass as two forward passes. A baseline JiT step then costs approximately $3F$. Evaluating the teacher on an active fraction α adds αF, giving a relative cost of

$$\frac{(3+\alpha)F}{3F} = 1 + \frac{\alpha}{3},$$

excluding the lightweight EMA update and periodic calibration. For the log-SNR coupling band $[2,\ 2.5]$, the saturated active fraction is $\alpha = 0.625$, yielding a multiplier of approximately $1.208\times$ before calibration. Calibration probes at most two log-SNR points using 16 batches per point each epoch, contributing an amortized cost of approximately $0.009\times$ the baseline training cost. The resulting estimated multiplier is approximately $1.22\times$ in the saturated regime and $1.21\times$ averaged over the full training run, during which the active fraction is initially lower.

Wall-clock profiling uses 8 GPUs, a global batch size of 1024, and SDPA with bfloat16 precision. For JiT-L/16, the step time increases from 0.544 seconds with teacher evaluation disabled to 0.649 seconds when the active fraction saturates, corresponding to a $1.19\times$ multiplier. Teacher overhead is lower before saturation, when fewer samples receive alignment supervision. For JiT-B/16, the saturated step time is 0.205 seconds. A separate paired measurement under matched load gives 0.71 seconds without teacher evaluation and 0.83 seconds with it, corresponding to a $1.17\times$ multiplier. At 1,251 steps per epoch, the reported saturated step times correspond to approximately 4.3 minutes for B/16 and 13.5 minutes for L/16, excluding any calibration time not already included in the step measurements. The training speed is broadly comparable to that of methods such as Self-Flow and SRA, while **JA$\boldsymbol{x}$** requires no additional trainable alignment layers.

F Additional Results

F.1 System-level Comparison

Table 6 provides a broader comparison with latent- and pixel-space diffusion models. We produce better results than our vanilla JiT baselines across scales and even achieve the same performance (FID=1.82) as JiT-G with our **JA$\boldsymbol{x}$** -H which is only 1/2 of the model size.

F.2 Additional qualitative results

Table 6: System-level comparison on ImageNet 256×256. Parameters include the generator and tokenizer decoder used at inference, excluding training-only components. GFLOPs measure one generator forward pass excluding tokenizer computation. **JAx** improves FID over JiT at each evaluated scale without changing inference computation. The specific REPA encoder used by DeCo is unspecified.

Method	External components	Params (M)	Epochs	GFLOPs	FID ↓	IS ↑
Latent-space diffusion						
DiT-XL/2 (Peebles & Xie, 2023)	VAE	$675 + 49$	1400	238	2.27	278.2
SiT-XL/2 (Ma et al., 2024)	VAE	$675 + 49$	1400	238	2.06	277.5
SiT-XL/2 + REPA (Yu et al., 2025)	VAE, DINOv2	$675 + 49$	200	238	1.96	264.0
SiT-XL/2 + REPA (Yu et al., 2025)	VAE, DINOv2	$675 + 49$	800	238	1.42	305.7
Pixel-space diffusion: existing systems						
SiD2, UViT/1 (Hoogeboom et al., 2025)	None	–	–	1306	1.38	–
PixelFlow-XL/4 (Chen et al., 2025b)	None	677	320	5818	1.98	282.1
PixelDiT-XL/16 (Yu et al., 2026)	DINOv2	797	320	311	1.61	292.7
PixelDiT-XL/16 (Yu et al., 2026)	DINOv2	797	800	311	1.54	297.0
DeCo-XL/16 (Heun) (Ma et al., 2026)	REPA encoder	682	600	–	1.69	304.0
DeCo-XL/16 (Euler) (Ma et al., 2026)	REPA encoder	682	800	–	1.62	301.0
JiT backbone comparisons						
JiT-B/16 (Li & He, 2026)	None	131	600	50	3.66	275.1
JAx -B/16	None	131	600	50	3.41	271.82
JiT-L/16 (Li & He, 2026)	None	459	600	176	2.36	298.5
JAx -L/16	None	459	600	176	2.26	284.93
JiT-H/16 (Li & He, 2026)	None	953	600	364	1.86	303.4
PixelREPA-H/16 (Shin et al., 2026)	DINOv2	953	600	364	1.81	317.2
JAx -H/16	None	953	600	364	1.82	302.8
Larger-backbone reference						
JiT-G/16 (Li & He, 2026)	None	2B	600	766	1.82	292.6

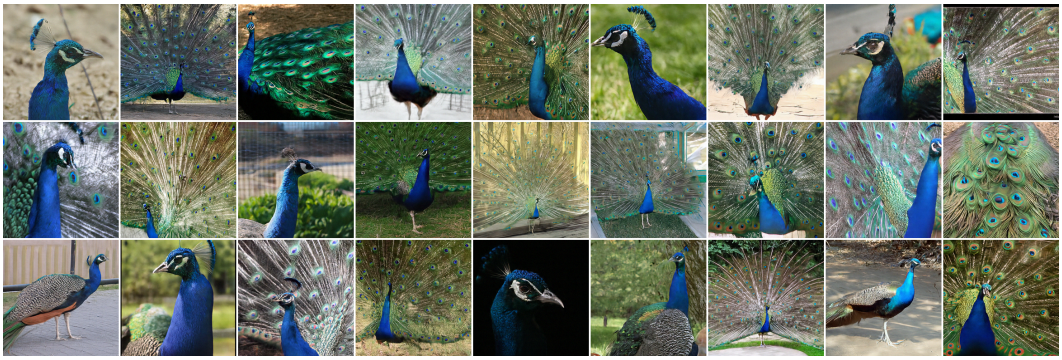

Figure 8: Uncurated samples from Class 084 (peacock)



Figure 9: Uncurated samples from Class 141 (redshank)



Figure 10: Uncurated samples from Class 213 (Irish setter)



Figure 11: Uncurated samples from Class 352 (impala)



Figure 12: Uncurated samples from Class 388 (giant panda)



Figure 13: Uncurated samples from Class 406 (altar)



Figure 14: Uncurated samples from Class 829 (streetcar)



Figure 15: Uncurated samples from Class 927 (trifle)



Figure 16: Uncurated samples from Class 932 (pretzel)



Figure 17: Uncurated samples from Class 937 (broccoli)



Figure 18: Uncurated samples from Class 941 (acorn squash)