Toward Early Quality Assessment of Text-to-Image Diffusion Models

Supplementary Material

7. Implementation Details

7.1. Network Structure and Data Processing Pipeline

Signals and Inputs. During sampling, we tap an intermediate denoiser activation $h_t \in \mathbb{R}^{C\times H\times W}$ at an early timestep t and optionally a text-side embedding e_{text} depending on the target evaluator. A sinusoidal timestep embedding is computed and projected to $e_t \in \mathbb{R}^{d_t}$, which conditions subsequent residual blocks. In addition, for SD3 we first unpatch h_t to recover the dense feature map (this step is not needed for SD2). We then resize the feature map by bilinear interpolation to $C \times 48 \times 48$. Next, we apply PCA along the channel dimension and keep the top 48 components, which yields a tensor of size $48 \times 48 \times 48$. This compression reduces GPU memory by a large margin and, in our experiments, does not cause a noticeable drop in evaluation performance.

Time-Conditioned Feature Encoder. We encode h_t using a stack of attention–residual down blocks with time modulation. Each block contains a time-conditioned ResNet sub-block followed by self-attention, and 3×3 convolution downsampling.

We use GroupNorm with 16 groups in the residual blocks, `SiLU` as the residual activation, and `conv` downsampling. The attention head dimension is set to 5. The sinusoidal timestep embedding has 64 channels and is projected to a time MLP of width 512; the residual blocks consume this time embedding through scale–shift conditioning.

The resulting feature map is aggregated by average pooling and flattened to a compact vector, which is then passed through a two-layer image-side MLP with one hidden layer of width 512 (`ReLU`, dropout 0.5, BatchNorm enabled).

Optional Text Alignment. For evaluators that depend on the prompt semantics (e.g., CLIP- or BLIP-based metrics, ImageReward, HPS), we project the text embedding to the same 512-dimensional space. We then form a concatenated vector by stacking u, the text vector v, their element-wise product $u \odot v$, and the absolute difference $|u - v|$. This enriched representation is fed to the final prediction head. We also compute a cosine-similarity matrix between u and v for analysis.

Prediction Head. A three-layer MLP (hidden width $512, 256$) with a Sigmoid output maps the representation to a scalar in $[0, 1]$ that serves as the early quality score.

Head Multiplicity and Dimensions. We instantiate one probe network per target evaluator key (e.g., `CLIPScore`, `BLIP_ITM`, `ImgReward`, `HPSv21`). Keys that use text rely on a text projector, while text-free keys skip this branch. Implementation uses d_t=64 for the sinusoidal timestep input, a time-MLP width of 512, attention groups of 32, and a final two-layer MLP with hidden sizes $[512, 256]$ and Sigmoid output. Pooling uses a 3×3 average operator by default.

8. Additional Results

8.1. Full Visualization

This section provides comprehensive visualizations of the structural signals evolving within the denoiser network across different timesteps, supplementing the core finding in Sec. 5.2. We visualize the feature maps by applying PCA along the channel dimension for a better view of their underlying structure, similar to Fig. 3 in main text.

Trajectory Final Images. Figure 7 shows the final generated images for the trajectories visualized in this section.

SD2 Full PCA Visualization. Fig. 8 and Fig. 9 extend the analysis of Stable Diffusion 2 (SD2) features. They show the PCA visualization for all Down, Mid, and Up blocks across timesteps $t = 0.2$ to $t = 0.9$ for two different prompts. This full set confirms the observation: while high-resolution details are absent early on, coarse structural cues (like layout and large object boundaries) emerge and stabilize in the mid-to-late layers (especially Up **3** as selected for probing) as early as $t = 0.2$ to $t = 0.3$. The features in Up **3** remain relatively stable compared to the noisier, high-frequency layers (Down 1, Up 4).

SD3-M & SD3-L Full PCA Visualization. Fig. 10 and Fig. 11 present the PCA visualizations for the intermediate blocks of Stable Diffusion 3.5 Medium (SD3-M) and Stable Diffusion 3.5 Large (SD3-L), respectively. Since SD3 series model contains more than 20 blocks, we visualize the feature every 4 blocks (strat from 0). These visualizations confirm that the structural stability phenomenon generalizes to transformer-based denoisers as well. Layers corresponding to mid-to-high resolutions (e.g., Block 8 and Block 12 in SD3-M, Block 16 and Block 20 in SD3-L) rapidly converge to a stable representation of the main objects and their layout (e.g., the bicycle shape and its components), even when $t \leq 0.3$. This cross-backbone consistency further validates the structural signal discovery for early quality assessment.

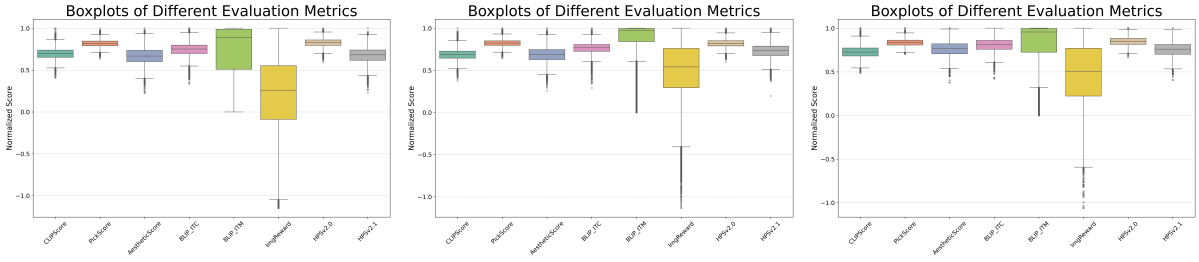
Finally, we use the output of block 20 of SD3-M and block 28 of SD3-L for representative feature for ProbeSelect.

8.2. Supplement Results for Main Paper

This section provides additional quantitative and visual results referenced in the main paper, including the full-time

Time	ClipScore	PickScore	AeS	BLIP-ITC	BLIP-ITM	ImageReward	HPSv2.0	HPSV2.1
SD2								
0.4	0.71	0.79	0.65	0.65	0.98	0.99	0.71	0.64
0.5	0.71	0.79	0.65	0.65	0.98	0.99	0.71	0.64
0.6	0.71	0.79	0.65	0.65	0.98	0.99	0.71	0.64
SD3-M								
0.4	0.78	0.84	0.67	0.74	0.98	0.99	0.82	0.80
0.5	0.78	0.84	0.67	0.74	0.98	0.99	0.82	0.80
0.6	0.78	0.84	0.67	0.74	0.98	0.99	0.83	0.80
SD3-L								
0.4	0.80	0.85	0.71	0.74	0.98	0.99	0.82	0.78
0.5	0.80	0.85	0.71	0.74	0.98	0.99	0.82	0.78
0.6	0.80	0.85	0.71	0.74	0.98	0.99	0.82	0.78
FLUX.1-dev								
0.4	0.75	0.86	0.69	0.72	0.98	0.99	0.79	0.78
0.5	0.75	0.86	0.69	0.72	0.98	0.99	0.79	0.78
0.6	0.75	0.86	0.69	0.72	0.98	0.99	0.79	0.78

Table 3. Spearman correlation between predicted score and each evaluator from latent feature in different time stamps.



(a) SD2 (b) SD3M (c) Flux

Figure 6. Boxplots of evaluation metrics of samples generated by each model. All metric scores are normalized by dividing by their respective maximum values for visualization consistency.

range of Spearman correlations (extending Tab. 1) and the boxplots for all generative backbones (extending Fig. 4).

Table 3 presents the additional set of Spearman correlations between the early probe predictions ($\hat{y}_{t,m}$) and the final ground-truth metric scores ($R_m(x_1)$) across four backbones and eight evaluators. This table supplements the subset of results presented in the main paper's Tab. 1 by including correlations at timesteps $t = 0.3$ and $t = 0.5$.

The complete results confirm the primary finding of high and stable correlations : the correlation scores across all metrics are already strong at $t = 0.2$ and remain highly stable, changing only marginally up to $t = 0.6$. Specifically, metrics like BLIP-ITM and ImageReward achieve correlations near $0.98 - 0.99$ across all checkpoints, validating that reliable ranking signals emerge very early in the diffusion process, making $t = 0.2$ a robust choice for early pruning.

Figure 6 provides the boxplots for the normalized scores of all eight evaluation metrics (CLIPScore, PickScore, AeS, etc.) on samples generated by SD2, SD3-M, and FLUX.1-dev. These plots are provided for completeness (as referenced in 4) and illustrate the inherent distribution and dynamic range of each reference metric across the samples.

As noted in the main text Sec. 5.3, metrics with a broader, less-saturated score distribution, such as ImageReward (ImgReward), naturally produce a more stable relative ordering, which translates into the higher Spearman correlations. Conversely, metrics with narrower distributions, such as the HPSv2.x scores, often result in tighter clusters and a greater likelihood of ties in ranking.

Figure 7. Final image of trajectory visualization.

8.3. Earlier Checkpoints and Scheduler Robustness

To justify the default choice of $t = 0.2$, we further evaluate earlier checkpoints on SD3-L. Table 4 shows that the correlation is still weak at very early stages such as $t = 0.05$, improves substantially at $t = 0.1$ and $t = 0.15$, and becomes consistently strong by $t = 0.2$. This supports our claim that $t = 0.2$ offers a strong balance between prediction quality and compute savings.

We also repeat the analysis under different samplers, including Euler and Heun schedulers, and observe similar Spearman correlations at the same normalized checkpoint. This indicates that the early structural signal is robust to the specific scheduler choice and is not tied to one sampling implementation.

t	CS	PS	AS	BIC	BIM	IR
0.05	0.24	0.16	0.18	0.16	0.28	0.29
0.10	0.64	0.68	0.57	0.62	0.76	0.78
0.15	0.75	0.82	0.67	0.69	0.92	0.92
0.20	0.79	0.84	0.70	0.74	0.98	0.99

Table 4. Spearman correlation at very early checkpoints on SD3-L. Correlations become consistently strong at $t = 0.2$, supporting our default choice.

8.4. Cross-Backbone Transfer of Probe Networks

Although our main experiments train one probe per backbone, we find that the learned probe transfers well across different diffusion models after the shared PCA-based feature processing. Table 5 reports the transfer results across SD2, SD3-M, SD3-L, and FLUX.1-dev. A probe trained on one backbone remains close to peak performance on the others, suggesting that the captured early structural signal is not highly model-specific. This substantially reduces the practical cost of deployment, since one trained probe can serve as a plug-in evaluator for multiple backbones.

Source \ Target	SD2	SD3-M	SD3-L	FLUX.1-dev
SD2	0.98	0.96	0.97	0.96
SD3-M	0.86	0.98	0.95	0.98
SD3-L	0.86	0.98	0.98	0.97
FLUX.1-dev	0.89	0.97	0.97	0.98

Table 5. Cross-backbone transfer of probe networks. Each entry reports the Spearman correlation when a probe trained on the source backbone is applied to the target backbone.

8.5. Additional Baseline: Decoding and Scoring Early

A simple alternative to our method is to decode the predicted clean image $\hat{x}_0$ at an early checkpoint and directly apply a standard evaluator. We test this baseline at $t = 0.2$ using BLIP-ITM (BIM). The resulting Spearman correlation is only around 0.52, which is much lower than the correlation achieved by Probe-Select (typically above 0.9 under the same setting).

This gap is expected: standard evaluators are trained on fully formed images and are not designed to assess the potential of blurry or partially denoised outputs. In contrast, Probe-Select is trained directly on intermediate denoiser features and learns to map early structural signals to the final quality. This result shows that early quality assessment requires dedicated probing rather than directly reusing final-image evaluators on incomplete samples.

8.6. Ablation and Sensitivity Analysis

We conduct an ablation study to understand the influence of the key hyperparameters governing our joint training objective: the listwise ranking temperature $\tau_{\text{list,max,}}$ the listwise ranking margin α_{max} , and the weight of the contrastive alignment loss λ_{Align} . All ablations are performed on the FLUX.1-dev backbone using the ImageReward evaluator as the target metric, with probe predictions taken at the early checkpoint $t = 0.2$. The baseline configuration uses the values $\tau_{\text{list, max}} = \tau_{\text{Align, max}} = 1.0, \alpha_{\max} = 0.4\sigma$, and $\lambda_{\text{Align}} = 10$.

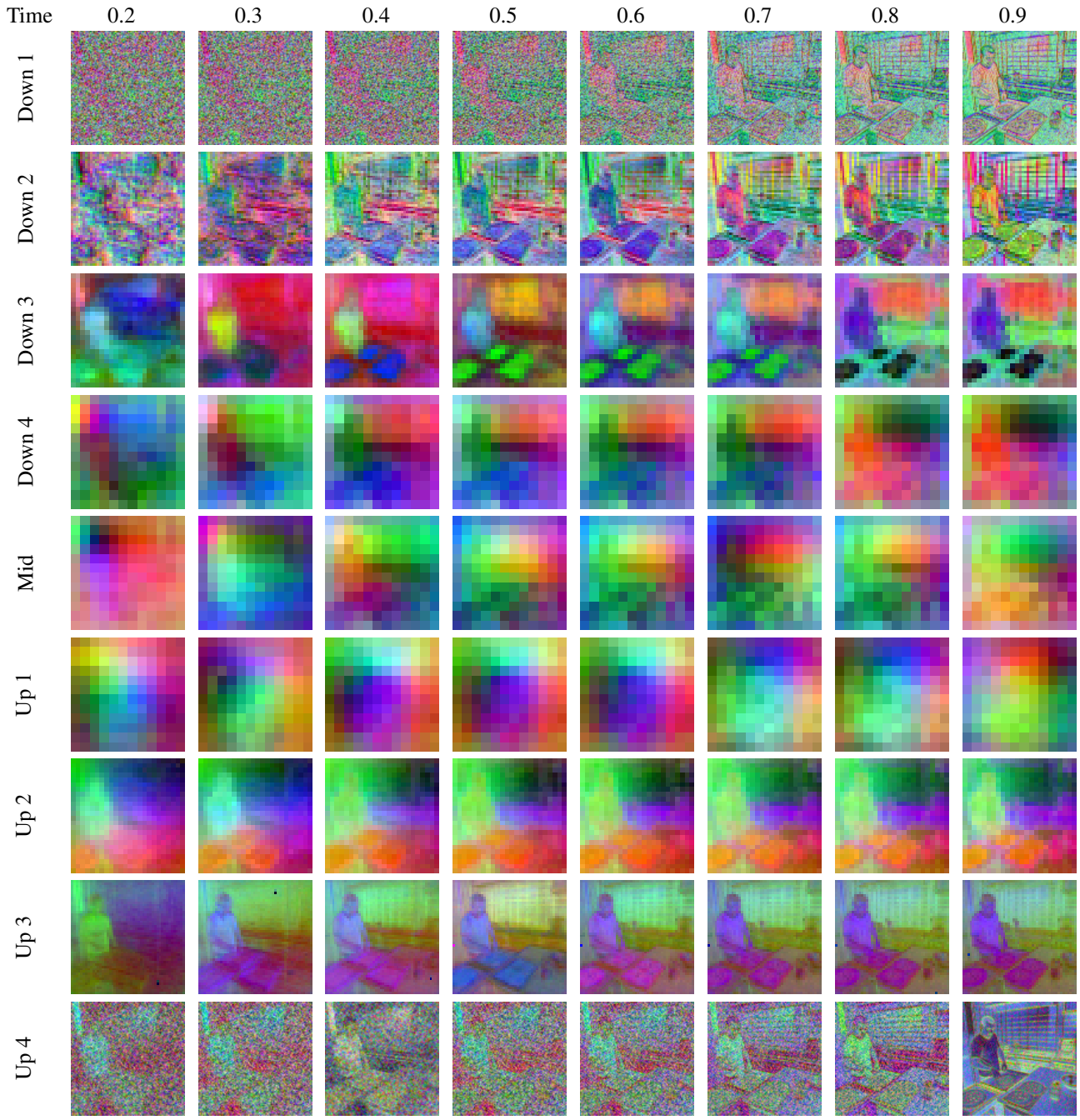


Figure 8. PCA visualization for denoising network of Stable Diffusion 2 across time.

Effect of Alignment Loss Weight (λ_{Align}). The weight λ_{Align} balances the primary ranking objective $\mathcal{L}_{\text{list}}$ and the auxiliary semantic alignment objective $\mathcal{L}_{\text{Align}}$ (Equation 7). $\mathcal{L}_{\text{Align}}$ helps maintain prompt awareness in the probe's latent space. We investigate its values with $\tau_{\text{list,max}} = 1.0$ and $\alpha_{\text{max}} = 0.4\sigma$ fixed. The $\lambda_{\text{Align}} = 0$ case corresponds to only using the listwise ranking loss.

Results presented in Table 6 show that the contrastive alignment loss is crucial for high-fidelity quality prediction. Without it ($\lambda_{\text{Align}} = 0$), the Spearman correlation drops significantly to 0.66 . The correlation improves steadily as λ_{Align} increases, reaching its peak fidelity of 0.99 at the baseline value of $\lambda_{\text{Align}} = 10$. This demonstrates that simply learning to rank based on visual cues is insufficient; the probe must also be guided to align its representation with the text prompt embedding to successfully forecast text-

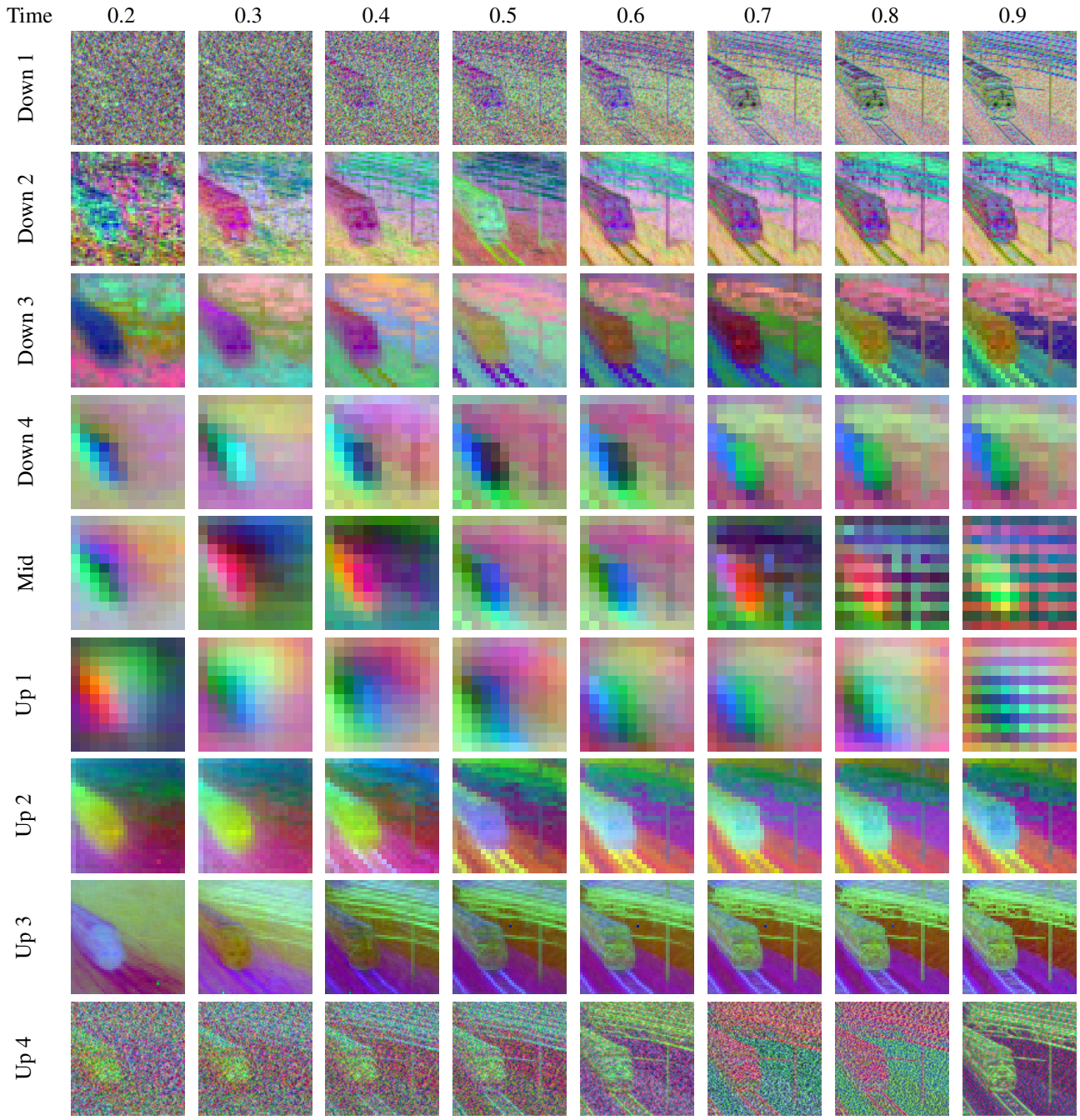


Figure 9. PCA visualization for denoising network of Stable Diffusion 2 across time.

aware quality metrics like ImageReward. The slight drop at $\lambda_{\text{Align}} = 50$ suggests an overly strong focus on text alignment can hurt ranking performance.

Effect of Temperature ($\tau_{\text{list, max}}$ and $\tau_{\text{Align, max}}$). The temperature τ controls the smoothness of the softmax functions in both the listwise ranking loss ($\mathcal{L}_{\text{list}}$) and the contrastive loss ($\mathcal{L}_{\text{Align}}$), where we set $\tau_{\text{list,max}} = \tau_{\text{Align,max}}$. A lower temperature leads to a stricter loss. We investigate its effect while keeping $\alpha_{\max} = 0.4\sigma$ and $\lambda_{\text{Align}} = 10$ fixed.

As shown in Tab. 8, the Spearman correlation is highly robust to the choice of τ in the range of $[0.1, 10.0]$, remaining exceptionally high across all tested values (between 0.96 and 0.99). The best performance is achieved with $\tau \in \{0.5, 1.0\}$, indicating that a moderate degree of strictness in the ranking is beneficial. The general stability suggests that once the prompt-aware features are extracted (due

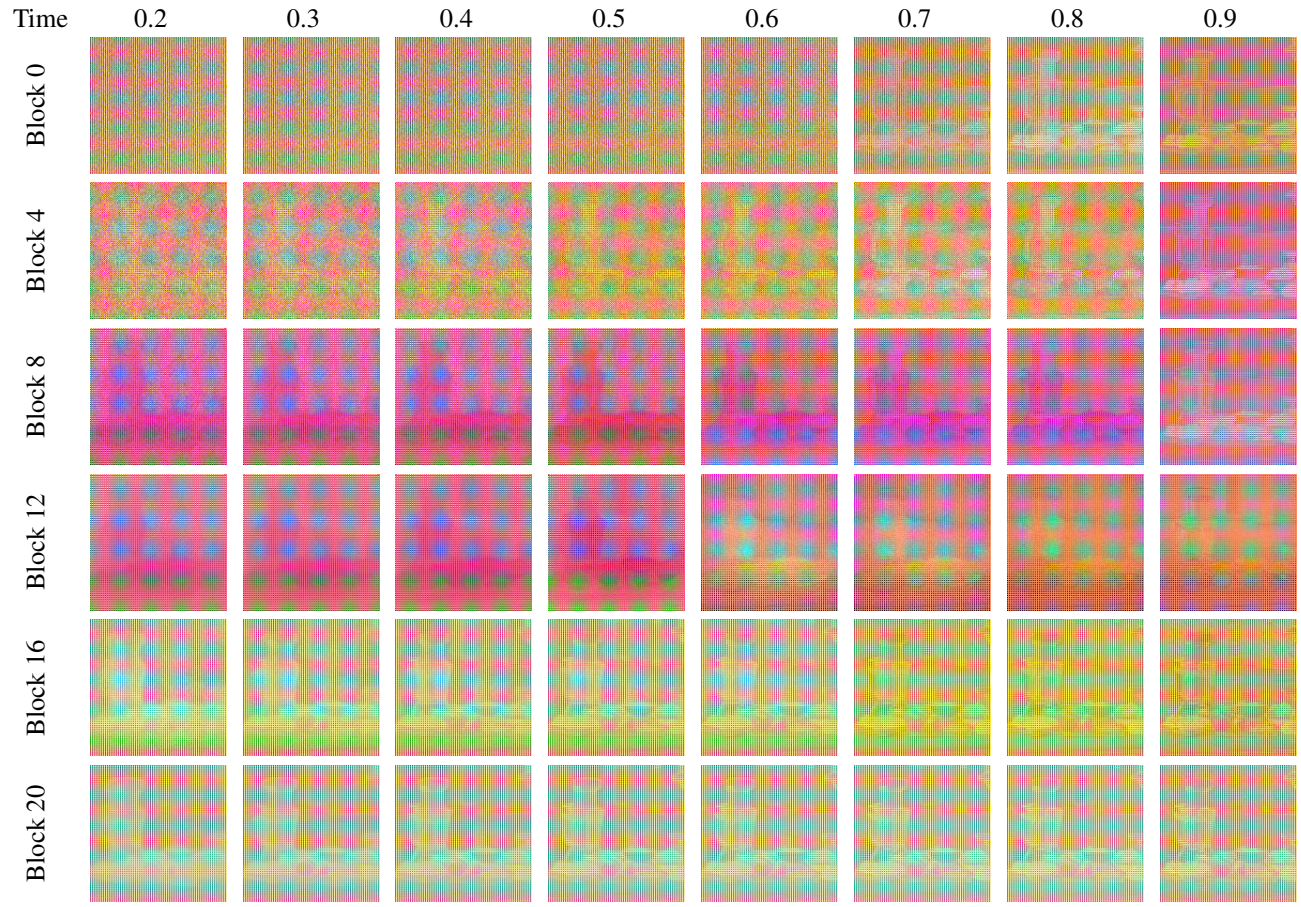


Figure 10. PCA visualization for denoising network of Stable Diffusion 3.5 medium across time.

Dataset	Method	CS	PS	BIC	BIM	IR
DrawBench	Baseline	0.32	0.23	0.48	0.90	1.08
	Probe-Select	0.35	0.23	0.52	0.93	1.55
GenEval	Baseline	0.34	0.23	0.49	0.90	1.07
	Probe-Select	0.35	0.23	0.52	0.96	1.53
HPD	Baseline	0.34	0.22	0.48	0.95	1.12
	Probe-Select	0.36	0.23	0.51	0.98	1.38
T2I-CompBench	Baseline	0.32	0.22	0.49	0.95	1.13
	Probe-Select	0.33	0.23	0.52	0.98	1.49

Table 6. Generalization to additional text-to-image benchmarks. Probe-Select consistently improves final quality metrics over the no-selection baseline.

τ	Validation Spearman Correlation (IR)
0	0.66
1	0.79
2	0.84
5	0.94
10	0.99
20	0.98
50	0.93

Table 7. Ablation on the contrastive alignment loss weight λ_{Align}

to $\lambda_{\text{Align}} > 0$), the precise shaping of the ranking loss via the temperature is less critical.

Effect of Listwise Ranking Margin ($\alpha_{\max}$). The margin $\alpha_{\max}$ in $\mathcal{L}_{\text{list}}$ ensures a minimum score separation is enforced between preferred and less-preferred samples. We study its impact with $\tau_{\text{list,max}} = 1.0$ and $\lambda_{\text{Align}} = 10$ fixed. The margin is expressed as a fraction of σ, the standard deviation of the ImageReward scores.

The results in Table 9 indicate that the margin is a critical hyperparameter. Very small margins, such as 0.01σ, result in a significantly degraded correlation of 0.90 . The ranking performance steadily improves as the margin increases, with the optimal performance of 0.99 achieved at the baseline value of $\alpha_{\max} = 0.4\sigma$. This suggests that the probe needs a large separation in the predicted scores to effectively align with the finegrained relative preferences

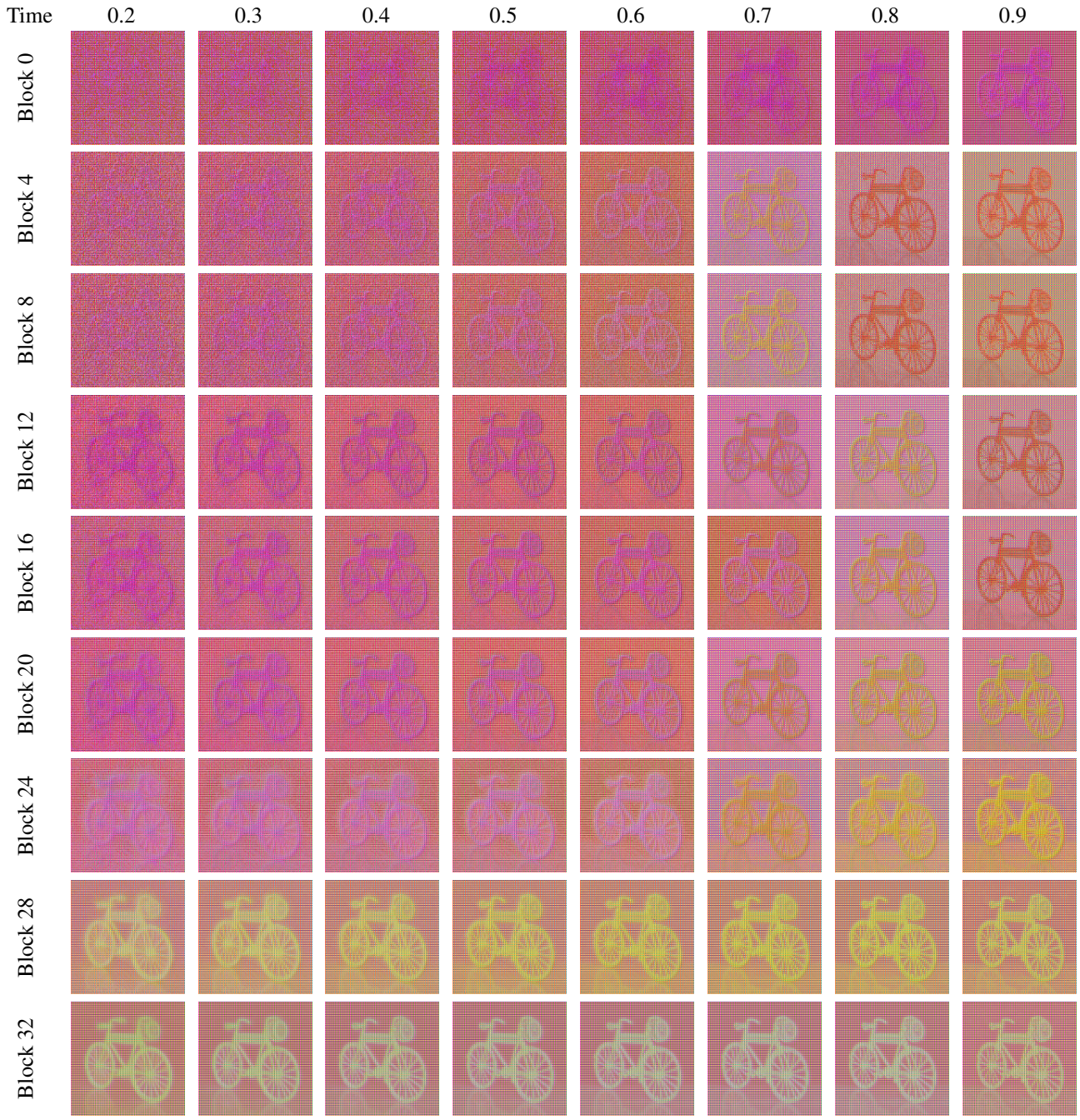


Figure 11. PCA visualization for denoising network of Stable Diffusion 3.5 large across time.

captured by the ImageReward metric. Setting the margin too high (0.5σ) slightly degrades performance, likely because it over-penalizes the model when only very small differences exist in the ground-truth scores.

Effect of Latent Channel Dimension C_{PCA} In our Probe-Select framework, we compress the intermediate denoiser activation h_t via Principal Component Analysis (PCA) along the channel dimension to a reduced size of $C_{\text{PCA}} = 48$. This compression is crucial for reducing GPU memory footprint and computational overhead. To validate this design choice, we perform an ablation study on the reduced channel dimension $C_{\text{PCA}} \in \{16, 32, 48, 96, 128, 192, 256\}$.

We train the ImageReward probe on the FLUX.1-dev backbone at $t = 0.2$ using the baseline hyperparameters ($\tau = 1.0, \alpha_{\max} = 0.4\sigma, \lambda_{\text{Align}} = 10$). For each dimen-

τ	Validation Spearman Correlation (IR)
0.1	0.98
0.5	0.99
1.0	0.99
2.0	0.97
3.0	0.96
5.0	0.96
10.0	0.96

Table 8. Table 4. Ablation on the temperature $\tau_{\text{list, max}}$ and $\tau_{\text{Align, max}}$.

τ	Validation Spearman Correlation (IR)
0.01σ	0.90
0.02σ	0.94
0.05σ	0.95
0.1σ	0.94
0.2σ	0.95
0.3σ	0.98
0.4σ	0.99
0.5σ	0.97

Table 9. Ablation on the maximum listwise ranking margin $\alpha_{\max}$

sion, we report the validation Spearman correlation and the probe's total number of parameters (to quantify the overhead). The results are summarized in Tab. 10.

C_{PCA}	Validation Spearman Correlation (IR)
16	0.84
32	0.96
48	0.99
96	0.99
128	0.99
192	0.99
256	0.99

Table 10. Ablation on on the reduced channel dimension C_{PCA} for the denoiser feature map.

The results demonstrate that prediction fidelity is sensitive to severe compression: using a very low dimension like $C_{\text{PCA}} = 16$ leads to a significant drop in correlation to 0.84 , indicating insufficient structural information is retained. However, the correlation quickly saturates to the peak value of $\mathbf{0.99}$ once the dimension reaches $C_{\text{PCA}} = 48$. Since the probe's computational cost and parameter count scale with C_{PCA}, and no further performance gain is observed beyond 48 , we select $C_{\text{PCA}} = 48$ as the optimal dimension to balance high prediction fidelity with the required storage and computational overhead, aligning with our goal of a lightweight, efficient early assessment mechanism.