# PRIOR SUBSPACE RECOMPOSITION: A UNIFIED PRIOR INITIALIZATION FRAMEWORK FOR CONDITIONAL GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Initial noise plays a crucial yet underexplored role in continuous conditional generative models. While existing golden-noise methods demonstrate that carefully selected or optimized noises can improve generation quality, they are often tailored to specific generation regimes or predefined quality objectives, limiting their adaptability across tasks and evaluation criteria. In this paper, we propose **Prior Subspace Recomposition** (PSR), a unified framework that reformulates noise initialization as metric-induced prior discovery followed by condition-aware anchor recomposition. Given a downstream evaluation suite, PSR first empirically identifies a metric-induced superior region in the Gaussian prior space and distills it into a reusable set of semantic-agnostic prior anchors. To handle complex conditions, PSR further estimates concept-conditioned anchor fitness and converts multi-dimensional metric responses into relative anchor preferences, which supervise a lightweight local ranker based on first-step denoising velocity fields. At inference time, local ranking predictions are resolved into a global anchor ordering through graph-based optimization, and the top-ranked anchors define a condition-specific local recomposition region on the Gaussian prior shell. The final initialization is obtained via geometry-preserving spherical averaging, integrating complementary anchor priors while preserving the prior distribution geometry. Extensive experiments across text-to-image, text-to-video, and image-to-video generation demonstrate that PSR consistently improves generation quality without modifying the underlying generative backbones.

## 1 INTRODUCTION

Continuous generative models (Wan Team et al., 2025; HaCohen et al., 2026; Labs, 2024) have become the dominant paradigm for high-fidelity visual generation. By transporting samples from a simple prior distribution to a complex data distribution, these models have demonstrated remarkable capability in synthesizing photorealistic and semantically rich content.

A largely underutilized degree of freedom in this process is the initial noise. Existing generative pipelines typically draw the initial noise from a standard Gaussian prior without considering its compatibility with the conditioning prompt. However, for deterministic samplers such as Probability Flow ODEs or flow-based trajectories, each initial noise defines a distinct path toward the data manifold. Therefore, even when random seeds lead to modest variance in aggregate benchmark scores, their semantic suitability can differ across prompts and concepts, especially in compositional scenarios involving multiple entities, spatial relations, or fine-grained attributes. This suggests an opportunity to improve generation not by modifying the generative backbone, but by systematically identifying, ranking, and combining more favorable prior initializations.

Recent studies on initial noise optimization have shown that carefully selected or optimized initial noises can substantially improve diffusion-based generation. In text-to-image (T2I) generation, existing methods typically optimize or predict initial noises to enhance prompt-image semantic faithfulness and compositional consistency. For instance, several works exploit attention-derived signals to identify semantically valid noise regions and guide the initial latent toward them (Mao et al., 2024; Guo et al., 2024; Zhou et al., 2025). In text-to-video (T2V) generation, recent methods further demonstrate the importance of noise initialization by refining the initial video latent to improve



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.

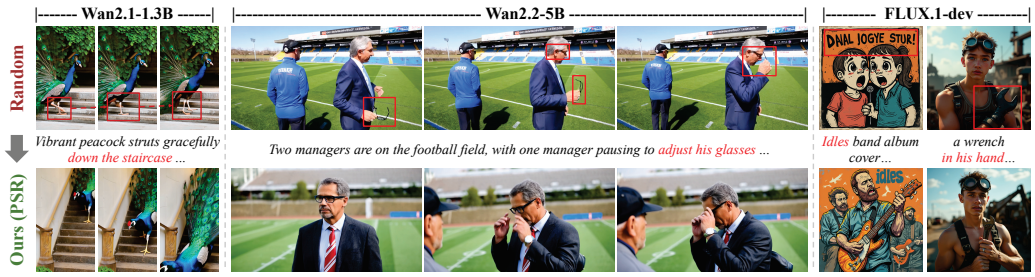


Figure 1: Qualitative comparisons of random initialization (top) and PSR (bottom) across T2V and T2I generation. Red boxes highlight representative errors in the random-initialization results.

temporal consistency, stabilize subject appearance, or compensate for spatio-temporal frequency gaps (Wu et al., 2024; Bai et al., 2025). Other video-oriented noise calibration methods are tailored to content-preserving enhancement, where the optimized noise is expected to improve visual quality while maintaining the input content (Yang et al., 2024).

Despite these promising results, existing noise initialization methods remain limited in two aspects. First, existing methods are largely tailored to a specific generation task, such as T2I or T2V, and thus lack a unified prior-initialization formulation that can generalize across different conditional generation regimes. Second, the optimization objective itself is often biased toward a particular quality dimension. For example, T2I methods mainly emphasize semantic faithfulness, while video initialization methods often prioritize temporal consistency and subject appearance stability. Such objective-specific optimization may improve the targeted metric, but it fails to provide a mechanism to adaptively discover which regions of the prior space are favorable under different downstream evaluation criteria. This motivates the central question of this work: *Can we construct a unified prior-initialization framework, where the evaluation objective defines the prior subspace and the conditioning signal determines how prior anchors are selected and combined?*

To answer this question, we propose **Prior Subspace Recomposition** (PSR), a unified framework for continuous conditional generation. Our framework decouples prior initialization into three stages: (1) *Metric-induced Prior Subspace Discovery.* Given a downstream evaluation suite, we empirically discover a metric-induced prior subspace by randomly sampling a large pool of noises, and distilling the top-performing candidates into a reusable anchor set. Therefore, routing denotes a condition-dependent mechanism that ranks, selects, and combines suitable anchors within this metric-induced subspace. Since the evaluation suite defines what constitutes high-quality generation, changing the evaluation suite naturally adapts the discovered anchors across different tasks and models. (2) *Semantic-Conditioned Prior Anchor Ranking.* Motivated by "Lottery Ticket Hypothesis", we construct a subject-level semantic pool and estimate concept-conditioned anchor fitness under the targeted metric suite. The resulting normalized fitness scores define relative anchor preferences, which act as supervisory signals to train a lightweight local ranker that predicts anchor compatibility from first-step velocity fields. (3) *Globally-aware Golden Anchor Ensemble.* At inference time, local ranking predictions are resolved into a global anchor ordering via graph optimization, and the top anchors define a condition-specific local recomposition region on the Gaussian prior shell, from which the final golden initialization is obtained. In contrast to existing methods that optimize noise for a specific generation regime or a predefined quality dimension, our framework allows the evaluation metric to induce the prior subspace, while the conditioning signal determines how prior anchors are ranked and recomposed within this subspace. This design yields a unified formulation that can generalize across diverse generation tasks and different downstream evaluation criteria.

We summarize our contributions as four-folds: (1) We reformulate noise initialization optimization as a prior-subspace routing problem. Instead of searching for an isolated golden noise, we identify a metric-induced high-yield region in the Gaussian space and distill it into a set of semantic-agnostic anchors; (2) We propose a semantic-conditioned prior anchor ranking mechanism to align anchors with fine-grained semantic concepts. Conditioned on multi-dimensional metric responses, we derive relative anchor preferences as supervision and train a local ranker based on initial denoising velocity fields to capture concept-specific prior suitability; (3) We introduce a partial-to-global anchor ensemble strategy for prior subspace recomposition. By resolving local anchor preferences into a global ordering through graph optimization, the top-ranked anchors are composed into a condition-specific local region on the Gaussian prior shell, from which a robust golden initialization is obtained via geometry-preserving spherical averaging; (4) The proposed framework is a unified and metric-

adaptive paradigm applicable across generation tasks. Experiments across models and benchmarks demonstrate that PSR improves generation quality over random initialization with manageable additional inference overhead. Representative visual comparisons are shown in Figure 1.

## 2 Related Work

**Initial Priors in Diffusion and Flow-Based Generative Models.** Flow Matching and Rectified Flow learn continuous transport fields between prior and data samples (Lipman et al., 2023; Liu et al., 2023). Similar formulations have been extended to video generation, from Video Diffusion Models (Ho et al., 2022) to recent transformer-based models such as HunyuanVideo (Kong et al., 2024), LTX-Video (HaCohen et al., 2026), and Wan (Wan Team et al., 2025). While these models directly adopt random noise for generation, recent work shows that initialization strongly affects generation. Mao et al. (2024) identified semantic winning-ticket noises; InitNO optimizes noise using attention-based validity signals (Guo et al., 2024). Based on that, Zhou et al. (2025) proposed learning prompt-conditioned favorable noise transformations. FIND adapts the initial noise distribution via policy optimization (Chen et al., 2024). Moreover, FreeInit (Wu et al., 2024), FastInit (Bai et al., 2025), and Noise Calibration (Yang et al., 2024) refine video noise for temporal consistency or content-preserving enhancement. Unlike these prompt-, video-, or enhancement-specific refinements, our work discovers reusable metric-induced prior anchors and routes generation through them according to semantic concepts.

**Structured Latent and Prior Subspaces.** Generative models often contain structured semantic spaces. For example, InterFaceGAN (Shen et al., 2020) identifies linear attribute directions, GANSpace discovers interpretable PCA directions (Härkönen et al., 2020), and closed-form factorization extracts semantic controls from generator weights (Shen & Zhou, 2021). As for diffusion models, Null-text Inversion proposed optimizing unconditional embeddings for accurate inversion (Mokady et al., 2023); Asyrp identifies semantic latent directions inside pretrained diffusion models (Kwon et al., 2023). Closest to initial priors, Wang et al. (2025) showed that diffusion noise seeds contain compositional cues revealed by patch PCA. These works mainly manipulate learned latents, features, attention maps, or inversion variables after or during generation. We instead study the initial Gaussian prior space before sampling begins, using "prior anchor" to denote an empirical high-yield region induced by evaluation metrics.

**Local to Global Ranking Aggregation.** Learning to rank studies how to infer an ordering among candidates from supervision signals such as pairwise preferences, listwise labels, or ranking-aware objectives. Representative methods include pairwise models such as RankNet (Burges et al., 2005), listwise formulations such as ListNet and ListMLE (Cao et al., 2007; Xia et al., 2008), and differentiable metric-based objectives such as NeuralNDCG (Pobrotyn & Białobrzeski, 2021). More importantly, our formulation is closely related to HodgeRank, which provides a graph-theoretic framework for recovering a global ranking from incomplete and noisy local pairwise comparisons (Jiang et al., 2011). Inspired by HodgeRank, we construct semantic-conditioned local preference relations among prior anchors and aggregate them into a global routing decision over the anchor graph.

## 3 Methodology

### 3.1 Metric-induced prior subspace discovery

Continuous generative models construct a mapping from a simple prior distribution, typically $\mathcal{N}(\mathbf{0}, \mathbf{I})$ in $\mathbb{R}^d$, to a complex data space $\mathcal{X}$. While the prior distribution is isotropic over $\mathbb{R}^d$, the deterministic nature of sampling trajectories, such as Probability Flow ODEs in diffusion and flow matching frameworks, suggests a critical geometric property. Since natural visual data are concentrated near a structured low-dimensional manifold in $\mathcal{X}$, their corresponding initial noise representations, induced by the inverse deterministic trajectory, are not equally favorable across the entire prior space under a given evaluation suite $\mathcal{M}$. This observation motivates the existence of a metric-induced subset of initial noise vectors that are more likely to generate high-fidelity and semantically meaningful samples. We call this structured subset the Prior Subspace, denoted by $\mathcal{U}_{\mathcal{M}} \subset \mathbb{R}^d$.

Directly characterizing $\mathcal{U}_{\mathcal{M}}$ is intractable due to the high dimensionality of the prior space and the cost of full generative evaluation. To obtain the support set of $\mathcal{U}_{\mathcal{M}}$ while mitigating subject-specific bias, we propose an empirical prior distillation strategy guided by a Semantic-Agnostic Common Prompt $p_c$. Deviating from arbitrary textual descriptions, $p_c$ is designed to be semantically neutral

yet sufficiently comprehensive, covering the intrinsic evaluation axes of downstream benchmarks; detailed design protocols are provided in Appendix A.1. Specifically, we first initialize an extensive candidate set $N_{init} = \{n_i\}_{i=1}^{N}$ with each $n_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \in \mathbb{R}^d$. For each candidate $n_i$, we evaluate its universal prior fitness by unrolling the generative mapping $\mathcal{F}$ conditioned on $p_c$, formulated as

$$\text{score}_i = \Psi(\mathcal{M}(\mathcal{F}(n_i|p_c))), \tag{1}$$

where $\mathcal{M} = \{m^{(1)}, \ldots, m^{(L)}\}$ denotes a holistic evaluation suite that quantifies generative performance over $L$ distinct dimensions (e.g., motion smoothness and temporal consistency), and $\Psi(\cdot)$ denotes the aggregation operator that fuses these metrics into a scalar composite score. By ranking these candidates based on their composite scores and retaining the Top-$K$ elements ($K \ll N$), we distill a compact set of subspace anchors, explicitly denoted as $\mathcal{A} = \{a_1, a_2, \ldots, a_K\}$. These anchors provide an empirical, semantic-agnostic approximation to the high-yield region of the prior space associated with $\mathcal{U}_{\mathcal{M}}$, serving as foundation for the subsequent fine-grained semantic routing.

### 3.2 Semantic-conditioned prior anchor ranking

While the universal subspace anchors distilled in Section 3.1 provide a strong generic prior for initialization, they are not sufficient for complex prompts that involve multiple entities, spatial relations, or dense semantic attributes. Recent studies on the Lottery Ticket Hypothesis in denoising (Mao et al., 2024) reveal that noise initializations exhibit pronounced concept specificity: an initialization that is favorable for one semantic concept may be ineffective, or even detrimental, for another. Consequently, relying on a single subspace anchor to satisfy a complex prompt can impose an undesirable generative compromise. This concept-specific view implies that a global initialization may act as an "ambiguous ticket", attempting to simultaneously accommodate disjoint semantic layouts. During the generative trajectory, this initialization ambiguity can intensify cross-attention interference among distinct textual tokens, frequently culminating in various compositional failures.

To overcome this limitation, we introduce a fine-grained distribution alignment strategy to identify entity-specific prior anchors for individual semantic concepts. Specifically, we first construct a subject-level semantic pool $\mathcal{S} = \{s_1, \ldots, s_m\}$ by applying syntactic parsing to a collection of prompts. The extracted concepts primarily correspond to subject entities and salient noun phrases, which serve as the semantic units for subsequent prior-anchor alignment. To rigorously eliminate contextual interference, we isolate and quantify the specific generative fitness of each subspace anchor $a_i \in \mathcal{A}$ via the $L$-dimensional evaluation suite $\mathcal{M}$, strictly conditioned on the individual semantic concept $s_j \in \mathcal{S}$. However, the intrinsic generation difficulty can vary substantially across concepts, and the metrics in $\mathcal{M}$ naturally operate on heterogeneous numerical scales. Directly aggregating raw metric values would therefore lead to an unstable and biased supervisory signal. To obtain comparable supervision across both concepts and metrics, we apply a hierarchical normalization procedure. For notation brevity, we define the standard Z-score operator as $\mathcal{Z}(x; \mu, \sigma) = (x - \mu)/\sigma$. For a given semantic concept $s_j \in \mathcal{S}$ and a subspace anchor $a_k \in \mathcal{A}$, we denote the raw response of the $\ell$-th evaluation metric as $y_{k,j}^{(l)} = m^{(l)}(\mathcal{F}(a_k|s_j))$. We compute the empirical mean $\mu_j^{(l)}$ and standard deviation $\sigma_j^{(l)}$ of this specific metric over the anchor set $\mathcal{A}$. To eliminate scale disparities among different metrics, we apply the Z-score normalization independently:

$$z_{k,j}^{(l)} = \mathcal{Z}\left(y_{k,j}^{(l)}; \mu_j^{(l)}, \sigma_j^{(l)}\right). \tag{2}$$

This normalization makes metric responses comparable within each concept by reducing scale-induced disparities across evaluation dimensions. Subsequently, we aggregate the standardized responses into a composite score via summation: $\hat{y}_{k,j} = \sum_{l=1}^{L} z_{k,j}^{(l)}$. The resulting score serves as the relative fitness of anchor $a_k$ for generating semantic concept $s_j$. Importantly, $\hat{y}_{k,j}$ is not intended to provide an absolutely calibrated quality value across all concepts; rather, it defines a concept-specific preference relation among anchors. This converts metric-based anchor evaluation into a learning-to-rank problem for semantic-conditioned prior routing. Directly learning a global ordering over all $K$ anchors is unnecessary and inflexible, especially when the anchor pool size may vary with the evaluation suite or computational budget. Instead, we learn local preference relations over small sampled anchor subsets and later aggregate them into a global ranking. Accordingly, we train a lightweight local ranker $r_\theta$ of length-$\tau$ to predict anchor preferences within sampled subsets. To make ranking efficient, the ranker should assess anchor suitability without unrolling the full generative trajectory. We therefore use the first-step velocity field as a lightweight predictive proxy:

$$v_{k,j} = \mathcal{F}_1(a_k \mid s_j), \tag{3}$$

where $\mathcal{F}_1$ denotes the first-step velocity-field predictor induced by the generative dynamics. This initial velocity captures the early transport direction from the prior anchor toward the data manifold, providing a computationally efficient yet informative feature for estimating the compatibility between anchor $a_k$ and semantic concept $s_j$.

During training, for a given semantic concept $s_j$, we randomly sample a sub-sequence of $\tau$ anchors ($\tau \ll K$) from the candidate pool, denoted as $\mathcal{T} \subset \mathcal{A}$. For all anchors in $\mathcal{T}$, we obtain their first-step velocity fields through a single-step forward pass and collect them as $\mathcal{V}_{\mathcal{T},j} = \{v_{u,j}\}_{a_u \in \mathcal{T}}$. We instantiate $r_\theta$ as a lightweight dual-stream multimodal Transformer. Given the velocity-field sequence $\mathcal{V}_{\mathcal{T},j}$ and the textual embedding of $s_j$, the length-$\tau$ ranker outputs a ranking score vector

$$S_{\mathcal{T},j} = r_\theta(\mathcal{V}_{\mathcal{T},j}, s_j) = \{S_{u,j} | a_u \in \mathcal{T}\} \in \mathbb{R}^\tau. \tag{4}$$

Here, each element $S_{u,j}$ denotes the predicted suitability score of anchor $a_u$ for semantic concept $s_j$. Comprehensive details regarding the network architecture and optimization are deferred to the Appendix A.3. Given the normalized fitness scores $\hat{y}_{\mathcal{T},j} = \{\hat{y}_{u,j} | a_u \in \mathcal{T}\}$, we train the ranker with a hybrid pairwise and list-wise learning-to-rank objective:

$$\mathcal{L} = \sum_{a_u, a_w \in \mathcal{T}} \max\left(0, \mathbf{1}(\hat{y}_{u,j} > \hat{y}_{w,j})(\gamma - (S_{u,j} - S_{w,j}))\right) + \lambda \mathcal{L}_{NeuralNDCG}(S_{\mathcal{T},j}, \hat{y}_{\mathcal{T},j}). \tag{5}$$

The first term is a pairwise margin loss that encourages $r_\theta$ to predict higher scores for superior anchors in each pair with predefined margin $\gamma$. The second term denotes the list-wise NeuralNDCG loss (Pobrotyn & Białobrzeski, 2021), which further aligns the predicted ordering over the entire anchor subset $\mathcal{T}$ with the ordering induced by $\hat{y}_{\mathcal{T},j}$. The coefficient $\lambda$ balances the pairwise and list-wise ranking objectives. By training on local anchor subsets, the rank-$\tau$ model $r_\theta$ learns to compare prior anchors using only their first-step velocity fields. The subsequent aggregation of these local predictions into a global routing decision during inference will be detailed in Section 3.3.

### 3.3 Globally-aware golden anchor ensemble

During the inference phase, when conditioned on a novel complex prompt $p$, our objectives are twofold: deriving the global ranking of all anchors by resolving the local predictions yielded by the ranker $r_\theta$, and ensembling the top-performing anchors to establish a robust, high-yield initialization.

Inspired by HodgeRank theory (Jiang et al., 2011), which posits that a globally consistent ranking can be robustly reconstructed by resolving local, potentially incomplete preference flows over a directed graph, we formulate the global ranking derivation as a graph-based optimization problem. Conditioned on the test prompt $p$, we first compute the initial velocity field set $\mathcal{V}_{\mathcal{A},p} = \{v_{k,p} \triangleq \mathcal{F}_1(a_k|p)\}_{k=1}^K$ for all $K$ anchors via a single-step forward pass. This process begins by first constructing an observation pool $\{\mathcal{V}^i_{\mathcal{A},p}\}_{i=1}^{N_p}$ comprising $N_p$ randomly shuffled variants of the velocity field set $\mathcal{V}_{\mathcal{A},p}$. For each such permuted variant $\mathcal{V}^i_{\mathcal{A},p}$, the $K$ velocity proxies are divided into $\lceil K/\tau \rceil$ non-overlapping sub-sequences of length $\tau$. Each sub-sequence is then ranked by $r_\theta$ to yield local relative scores, which are subsequently translated into explicit pairwise preferences. These comparative relationships populate a preference vector $\mathcal{P}^i$, where the entry $\mathcal{P}^i_{uw}$ for any proxy pair $(u, w)$ is assigned 1 if $S_{u,p} > S_{w,p}$, $-1$ for the converse, and 0 if they remain uncompared within the sub-sequence. Each preference vector $\mathcal{P}^i$ is naturally mapped to a directed edge set $E^i = \{(u, w) \mid \mathcal{P}^i_{uw} \neq 0\}$. By aggregating these directed edge sets across all $N_p$ permutations, we construct a comprehensive directed preference graph that covers the entire candidate pool. The derivation of the global ranking score vector $\mathbf{r}$ is then mathematically cast as resolving the global potential on this graph via least squares. Specifically, we utilize the graph incidence matrix $D^i \in \mathbb{R}^{|E^i| \times K}$ to encode the directed topological structure of $E^i$, leading to

$$\mathbf{r}^* = \arg\min_{\mathbf{r}} \sum_{i=1}^{N_p} \frac{1}{2N_p} \|D^i \mathbf{r} - \mathcal{P}^i\|_2^2, \tag{6}$$

where each row of $D^i$ corresponds to an active directed edge $(u, w)$, containing 1 and $-1$ at the $u$-th and $w$-th columns respectively and 0 elsewhere.

Based on the derived global ranking $\mathbf{r}^*$, a naive inference strategy would simply select the single top-ranked anchor. However, we depart from this isolated selection strategy and instead perform a multi-anchor spherical ensemble. The motivation is twofold. First, aggregating multiple

top-performing anchors allows the final initialization to integrate complementary generative priors, thereby better accommodating diverse semantic and structural requirements. Second, the distilled anchor set $\mathcal{A}$ is only a discrete approximation to the metric-induced Prior Subspace. Selecting a single anchor restricts inference to this finite candidate set, whereas composing multiple top-ranked anchors induces a local continuous region on the Gaussian prior shell. In this sense, our method returns from discrete anchor retrieval to ***Prior Subspace Recomposition***: the top-ranked anchors define a condition-specific local recomposition region within the discovered prior subspace, and the final initialization is selected from this region.

To respect the geometry of the prior space, the recomposition should remain on the typical set of the Gaussian prior. According to the *Gaussian Annulus Theorem* in high-dimensional probability, samples from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ concentrate on a thin hyperspherical shell with radius approximately $\sqrt{d}$. A naive linear interpolation among anchors may shrink the norm and move the recomposed initialization away from this high-probability region. Therefore, we employ a Normalized Spherical Averaging strategy. For a Top-$M$ ensemble, let $\{a_{(1)}, a_{(2)}, \ldots, a_{(M)}\}$ denote the $M$ highest-ranked anchors according to $\mathbf{r}^*$. We define their condition-specific spherical recomposition region as

$$\mathcal{R}_p^M = \left\{ \sqrt{d} \cdot \frac{\sum_{m=1}^{M} \beta_m a_{(m)}}{\left\| \sum_{m=1}^{M} \beta_m a_{(m)} \right\|_2} \;\middle|\; \boldsymbol{\beta} \in \Delta^{M-1} \right\}, \tag{7}$$

where $\Delta^{M-1}$ denotes the probability simplex. The final recomposed initialization is obtained as

$$z^*_{\text{init}} = \sqrt{d} \cdot \frac{\sum_{m=1}^{M} \beta_m a_{(m)}}{\left\| \sum_{m=1}^{M} \beta_m a_{(m)} \right\|_2}. \tag{8}$$

Here, $a_{(m)}$ represents the $m$-th highest-ranked anchor, $d$ is the dimensionality of the prior space, and $\beta_m$ controls the contribution of each anchor. This normalization projects the recomposed initialization back onto the Gaussian prior shell, allowing it to integrate complementary anchor directions while preserving the prior geometry. Thus, the output is not merely a retrieved golden anchor, but a recomposed golden initialization from a local region of the metric-induced Prior Subspace. In our implementation, we instantiate the above recomposition with the top-2 golden anchors for efficiency and robustness. We obtain the final initialization through a variance-preserving spherical blending: $z^*_{\text{init}} = \sqrt{1-\alpha^2}\, a_{(1)} + \alpha\, a_{(2)}$.

## 4 Experiments

**Evaluation Benchmarks.** To comprehensively evaluate the proposed PSR framework, we conduct experiments on three representative benchmarks spanning text-to-image, text-to-video, and image-to-video generation. (1) For the text-to-image task, we adopt **HPSv3** (Ma et al., 2025) to assess the human preference alignment of generated images. The HPSv3 benchmark contains 12,000 prompts sampled from 12 categories, and we report all HPSv3 scores following the official protocol. (2) For the text-to-video task, we employ **VMBench** (Ling et al., 2025) to evaluate the dynamic quality of generated videos. VMBench contains 1,050 diverse motion prompts, covering 6 movement modes. We report all five perception-driven metrics: Perceptible Amplitude Score (PAS), Object Integrity Score (OIS), Temporal Coherence Score (TCS), Motion Smoothness Score (MSS), and Commonsense Adherence Score (CAS). (3) For the image-to-video task, we adopt **VBench++** (Huang et al., 2025; Zheng et al., 2025) to evaluate video synthesis quality. It consists of 1,118 standardized sample pairs, where individual samples are evaluated across 10 dimensions.

**Implementation Details.** We build our framework on representative image and video generation backbones, including FLUX.1-dev (Labs, 2024) for text-to-image generation, as well as the T2V/I2V variants of the Wan series (Wan Team et al., 2025) and LTX-2.3 (HaCohen et al., 2026) for video generation. Specifically, we randomly sample a candidate pool of $N = 1000$ initial noises during offline prior discovery and retain the top $K = 50$ candidates as the anchor pool. We construct a semantic pool containing approximately 300 subject concepts for semantic-conditioned anchor supervision. The local ranker operates on subsets of length $\tau = 5$, and we use $N_p = 5$ permutations to construct the global preference graph. For both training and inference, images are generated at $1024 \times 1024$ resolution with FLUX.1-dev, while videos are generated at $480 \times 832$ resolution with 81 frames for Wan2.1 and at $1280 \times 704$ resolution with 121 frames for Wan2.2. We optimize the

Table 1: Quantitative comparison on VMBench for text-to-video generation. Random reports the mean and standard deviation of benchmark-level scores over 10 full-benchmark runs, with initial noise independently resampled for each video in every run. Better results are highlighted in bold.

| Model | Method | VMBench | | | | | |
|---|---|---|---|---|---|---|---|
| | | PAS | OIS | TCS | MSS | CAS | Avg. |
| Wan2.1-1.3B | Random | 11.46±0.06 | 67.27±0.19 | 93.77±0.03 | 73.42±0.23 | 58.18±0.09 | 60.82 |
| | +Ours | **11.60** | **69.81** | **94.07** | **75.07** | **62.18** | **62.55** |
| Wan2.2-5B | Random | 14.86±0.57 | 68.65±0.65 | 95.21±0.18 | 79.46±0.55 | 58.58±0.24 | 63.35 |
| | +Ours | **24.84** | **70.43** | **96.31** | **84.00** | **63.32** | **67.78** |
| LTX-2.3 | Random | 11.46±0.01 | 57.16±0.16 | 94.09±0.08 | 80.40±0.09 | 38.11±0.20 | 56.24 |
| | +Ours | **18.52** | **58.61** | **94.66** | **80.48** | **41.89** | **58.83** |

Table 2: Quantitative comparison on HPSv3 for text-to-image generation. Random reports the mean and standard deviation of benchmark-level scores over 10 full-benchmark runs, with initial noise independently resampled for each video in every run. Better results are highlighted in bold.

| Model | Method | HPSv3 | | | | | |
|---|---|---|---|---|---|---|---|
| | | Characters | Arts | Design | Architecture | Animals | Natural Scenery |
| FLUX.1-dev | Random | 11.60±0.02 | 10.32±0.02 | 9.49±0.01 | 11.01±0.02 | 10.67±0.02 | 10.12±0.03 |
| | +Ours | **11.81** | **10.62** | **9.69** | **11.17** | **10.86** | **10.30** |
| **Model** | **Method** | **HPSv3** | | | | | |
| | | **Transportation** | **Products** | **Others** | **Plants** | **Food** | **Science** |
| FLUX.1-dev | Random | 10.91±0.02 | 10.84±0.03 | 10.46±0.04 | 10.83±0.03 | 11.58±0.02 | 9.32±0.02 |
| | +Ours | **11.08** | **11.06** | **10.72** | **11.09** | **11.81** | **9.57** |

ranker using the AdamW optimizer with a base learning rate of $10^{-5}$. The ranker is trained for 40 epochs with a global batch size of 32. All experiments are conducted on 8 NVIDIA H200 GPUs. To verify the effectiveness of PSR without introducing specific tuning, we consistently set $\alpha = 0.3$.

### 4.1 Quantitative analysis

**Text-to-Video Generation.** Table 1 reports the quantitative comparison on VMBench. Across all evaluated T2V backbones, the proposed method consistently improves the average VMBench score over random initialization. For Wan2.1-1.3B, our method improves the average score from $60.82$ to $62.55$, with gains across all five metrics. Notably, the improvement is especially clear on OIS, MSS, and CAS, suggesting that the recomposed initialization helps stabilize both object-level generation and motion-level plausibility. Moreover, the improvement becomes more pronounced on the stronger Wan2.2-5B backbone. Our method increases the average VMBench score from $63.35$ to $67.78$, with substantial gains on PAS, MSS, and CAS. In particular, PAS improves from $14.86$ to $24.84$, indicating that the discovered prior anchors are effective in activating more perceptible motion patterns rather than merely preserving static visual quality. Similar trends are observed on LTX-2.3, where PSR improves the average score from $56.24$ to $58.83$. The consistent improvements across different T2V architectures demonstrate the structural applicability of the proposed PSR.

**Text-to-Image Generation.** Table 2 evaluates PSR on HPSv3 with FLUX.1-dev. Compared with random initialization, our method achieves better performance in all categories. The gains are consistent across diverse semantic categories, indicating that the proposed framework is not restricted to motion-centric video metrics, but can also adapt to image-level human preference objectives.

**Image-to-Video Generation.** Table 3 further evaluates our method on VBench++ for I2V generation. Here are several observations. First of all, the most notable gain appears on Dynamic Degree, which increases from $52.89$ to $61.38$, indicating that our method can encourage richer motion while maintaining temporal stability. Meanwhile, Subject Consistency and Background Consistency also improve, showing that the generated motion does not come at the cost of severe identity or scene drift. Secondly, a more nuanced observation is that our method does not improve over random initialization on I2V Background and Camera Motion as well as showing a slight decrease on Imaging Quality. This suggests that the current metric-induced prior anchors may favor dynamic expressiveness and temporal stability more strongly than camera-motion control or low-level imaging fidelity. Such trade-offs are reasonable under a multi-dimensional evaluation setting and further support the central motivation of this work: different evaluation suites can induce different favorable prior regions. With a metric suite that assigns larger weights to camera motion or imaging quality, the same framework can be adapted to discover anchors better aligned with these objectives.

Table 3: Quantitative comparison on VBench++ for image-to-video generation. Random reports the mean and standard deviation of benchmark-level scores over 10 full-benchmark runs, with initial noise independently resampled for each video in every run. Better results are highlighted in bold.

| Model | Method | VBench++ I2V Subject | I2V Background | Camera Motion | Subject Consistency | Background Consistency |
|---|---|---|---|---|---|---|
| Wan2.2-5B | Random | 98.52±0.06 | 99.44±0.06 | 40.56±1.31 | 95.40±0.08 | 97.33±0.12 |
| | +Ours | **98.63** | 99.39 | 37.48 | **96.24** | **97.83** |
| **Model** | **Method** | **VBench++ Motion Smoothness** | **Dynamic Degree** | **Aesthetic Quality** | **Imaging Quality** | **Temporal Flickering** |
| Wan2.2-5B | Random | 98.65±0.04 | 52.89±1.20 | 64.21±0.09 | 70.96±0.10 | 97.47±0.04 |
| | +Ours | **99.07** | **61.38** | **64.53** | 70.57 | **98.35** |

Table 4: Additional comparison and efficiency analysis of PSR. Left: comparison with NPNet on SDXL. Right: quality-efficiency trade-off on Wan2.1-1.3B, reporting VMBench performance and inference latency under different anchor pool sizes. Best quality results are highlighted in bold.

| Method | Model | HPSv2 | ImageReward | CLIPScore |
|---|---|---|---|---|
| Random | SDXL | 29.71 | 96.63 | 0.8734 |
| +NPNet | SDXL | 29.88 | 98.81 | 0.8813 |
| +Ours | SDXL | **33.23** | **108.53** | **0.8991** |

| Method | $K$ | Avg. | Base | Extra | Total |
|---|---|---|---|---|---|
| Random | – | 60.82 | 91 s | 0 s | 91 s |
| +Ours | 25 | 62.51 | 91 s | 13 s | 104 s |
| +Ours | 50 | **62.55** | 91 s | 24 s | 115 s |

**Comparison with NPNet.** Our cross-task comparisons with random initialization assess the framework-level applicability of PSR. To complement this evaluation, we additionally instantiate PSR on SDXL, using the same generative backbone and evaluation protocol as NPNet (Golden Noise) (Zhou et al., 2025). NPNet is a relevant learned initialization baseline because it also constructs prompt-conditioned noise without modifying the backbone. PSR uses only HPSv2 (Wu et al., 2023) for anchor discovery and ranking supervision. Table 4 (left) reports better performance than NPNet on all three metrics. In particular, although only HPSv2 is used to induce the anchor subspace and supervise the routing process, PSR also improves ImageReward (Xu et al., 2023) and CLIPScore (Hessel et al., 2021). This result provides two complementary observations: first, PSR is competitive with a representative golden-noise method; second, the induced prior subspace exhibits measurable transfer to evaluation dimensions that are not used during optimization. Comparisons with fixed-anchor selection and semantic-similarity retrieval are provided in Appendix A.5.

### 4.2 Qualitative analysis

**Text-to-Video Generation on VMBench.** We first present qualitative comparisons on VMBench using Wan2.1-1.3B and Wan2.2-5B. For each example, the last column shows the result produced by our method, while the preceding columns are generated from randomly initialized noise. As shown in Figure 2, random noise often results in unstable object localization, distorted motion trajectories, and inconsistent foreground appearance. In contrast, our method produces more reliable samples with improved spatial alignment and more coherent motion.

**Robustness across Different Video Backbones.** The results on Wan series models demonstrate that the proposed noise prior initialization strategy is not tied to a specific video generation backbone. Across different resolutions and model variants, our method consistently generates samples with fewer visual artifacts and stronger prompt alignment. This advantage is evident in the highlighted regions, where random noise may introduce unrealistic deformations, or fail to maintain temporal consistency, whereas our PSR better preserves the intended scene structure and motion semantics.

**Text-to-Image Generation on HPSv3.** We further evaluate the qualitative behavior on HPSv3 using FLUX.1-dev. Similar to the video cases, the last column corresponds to our method, while the preceding columns correspond to samples generated from random noise. Randomly initialized candidates often suffer from incomplete details, or weaker semantic correspondence to the prompt. By initializing generation with the golden noise provided by PSR, our method achieves better fine-grained visual fidelity and stronger prompt consistency.

### 4.3 Ablation study

**Ablation on Ensemble Coefficient $\alpha$.** We further analyze the effect of the ensemble coefficient $\alpha$ in the top-2 golden-anchor recomposition. The ablation is conducted on Wan2.1-1.3B under the VMBench benchmark, and Figure 3 reports the improvement over the random baseline across all VMBench metrics. Several observations can be made. First, using the top-ranked anchor alone




Figure 2: Qualitative comparisons across T2V and T2I benchmarks. Wan2.1 and Wan2.2 are evaluated on VMBench, while FLUX.1-dev is evaluated on HPSv3. Red boxes indicate representative regions where random initialization results in noticeable artifacts or quality degradation.

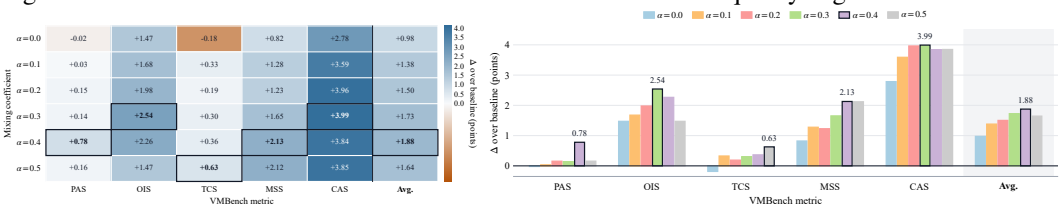


Figure 3: Ablation on the blending coefficient $\alpha$ for top-2 golden-anchor recomposition on Wan2.1-1.3B evaluated by VMBench. Both figures report the improvement over the random baseline across VMBench metrics. The best setting for each metric is highlighted.

($\alpha = 0$) already improves the average VMBench score over the random baseline, indicating that the globally ranked anchor provides a favorable initialization prior. Second, introducing the second-ranked anchor further improves performance for most choices of $\alpha$, suggesting that top-ranked anchors contain complementary generative priors. Third, the performance remains consistently above the baseline for all tested values of $\alpha$, demonstrating that PSR is not sensitive to this hyperparameter.

**Inference Efficiency and Anchor Pool Size.** At inference time, the routing overhead mainly includes extracting the prompt-conditioned first-step velocity fields. To reduce this cost, the unconditional classifier-free-guidance velocity fields associated with the anchors are precomputed and cached, while the prompt-conditioned first-step velocity fields of all 50 anchors are extracted in parallel. With cached prompt-independent unconditional features and batched candidate evaluation, PSR adds 24 seconds to the 91-second generation baseline at $K = 50$ (115 seconds total). Reducing $K$ to 25 lowers the latency to 13 seconds, while the VMBench average changes only from 62.55 to 62.51 (Table 4, right). The full metric results and cost breakdown are provided in Appendix A.3.

## 5 Conclusion

We introduce Prior Subspace Recomposition (PSR), a unified prior initialization framework for continuous conditional generation. Instead of treating the initial noise as an incidental random variable or optimizing it for a specific prompt, PSR reformulates noise initialization as a metric-induced prior discovery and condition-aware recomposition problem. Given a downstream evaluation suite, PSR first discovers a set of metric-induced prior anchors that provide a discrete approximation to favorable regions of the Gaussian prior space. It then learns semantic-conditioned anchor preferences from normalized multi-dimensional metric responses, using first-step denoising velocity fields as efficient proxies for anchor–condition compatibility. At inference time, local anchor preferences are aggregated into a global ordering, and the top-ranked anchors are recomposed on the Gaussian prior shell through geometry-preserving spherical averaging. Experiments across T2I, T2V, and I2V generation demonstrate that PSR consistently improves generation quality across diverse backbones and benchmarks without modifying the underlying generative models.

### AI USE STATEMENT

We used generative AI tools to assist with manuscript writing and language polishing, including improving grammar, clarity, and readability. All AI-assisted text was reviewed and, where necessary, revised by the authors to ensure accuracy and consistency with the underlying research. We take full responsibility for the final content of this work, including all text, claims, and reported results.

## REFERENCES

Chengyu Bai, Yuming Li, Zhongyu Zhao, Jintao Chen, Peidong Jia, Qi She, Ming Lu, and Shanghang Zhang. FastInit: Fast noise initialization for temporally consistent video generation. *arXiv preprint arXiv:2506.16119*, 2025.

Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 89–96, 2005. doi: 10.1145/1102351.1102363.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 129–136, 2007. doi: 10.1145/1273496.1273513.

Changgu Chen, Libing Yang, Xiaoyan Yang, Lianggangxu Chen, Gaoqi He, Changbo Wang, and Yang Li. FIND: Fine-tuning initial noise distribution with policy optimization for diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6735–6744, 2024. doi: 10.1145/3664647.3681047.

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. InitNO: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2404.04650.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: Discovering interpretable GAN controls. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9841–9850, 2020.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2204.03458.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. doi: 10.1109/TPAMI.2025.3633890.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127(1):203–244, 2011. doi: 10.1007/s10107-010-0419-x.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.10960.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Xinran Ling, Chen Zhu, Meiqi Wu, Hangyu Li, Xiaokun Feng, Cundian Yang, Aiming Hao, Jiashu Zhu, Jiahong Wu, and Xiangxiang Chu. Vmbench: A benchmark for perception-aligned video motion generation. *arXiv preprint arXiv:2503.10076*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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*, 2023. URL https://arxiv.org/abs/2209.03003.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score, 2025. URL https://arxiv.org/abs/2508.03789.

Jiafeng Mao, Xueting Wang, and Kiyoharu Aizawa. The lottery ticket hypothesis in denoising: Towards semantic-driven initialization. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2312.08872.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6038–6047, 2023.

Przemysław Pobrotyn and Radosław Białobrzeski. NeuralNDCG: Direct optimisation of a ranking metric via differentiable relaxation of sorting. *arXiv preprint arXiv:2102.07831*, 2021.

Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1532–1540, 2021.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9243–9252, 2020.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Qingsong Wang, Zhengchao Wan, Mikhail Belkin, and Yusu Wang. Seeds of structure: Patch PCA reveals universal compositional cues in diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=EgH5WYB6my.

Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. FreeInit: Bridging initialization gap in video diffusion models. In *European Conference on Computer Vision*, pp. 378–394, 2024. URL https://arxiv.org/abs/2312.07537.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: Theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1192–1199, 2008. doi: 10.1145/1390156.1390306.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Qinyu Yang, Haoxin Chen, Yong Zhang, Menghan Xia, Xiaodong Cun, Zhixun Su, and Ying Shan. Noise calibration: Plug-and-play content-preserving video enhancement using pre-trained video diffusion models. In *European Conference on Computer Vision*, pp. 307–326, 2024. URL https://arxiv.org/abs/2407.10285.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Zikai Zhou, Shitong Shao, Lichen Bai, Shufei Zhang, Zhiqiang Xu, Bo Han, and Zeke Xie. Golden noise for diffusion models: A learning framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2411.09502.

# A Appendix

## A.1 Common prompts for different benchmarks

**Common Prompt for VMBench**

Best quality, cinematic high-fidelity footage, high resolution. Featuring hyper-realistic intricate micro-textures, complex volumetric lighting, and ray-traced reflections. Demonstrating smooth fluid motion dynamics, high temporal coherence, and structural stability without flickering. Sharp focus, depth of field, physically accurate light and shadow interplay, continuous camera movement.

**Common Prompt for VBench**

Best quality, cinematic high-fidelity footage, high resolution. Featuring hyper-realistic intricate micro-textures, complex volumetric lighting, and ray-traced reflections. A mesmerizing close-up of thick, metallic liquid gold and iridescent silver fluids slowly mixing and swirling together in mid-air. Demonstrating smooth fluid motion dynamics, high temporal coherence, and structural stability without flickering. Sharp focus, depth of field, physically accurate light and shadow interplay, continuous camera slowly panning right.

**Common Prompt for HPSv3**

Best quality, wide-spectrum human-preference-aligned still image, high resolution. Featuring refined aesthetic composition, coherent spatial layout, clear visual hierarchy, balanced color harmony, natural illumination, accurate perspective, rich high-frequency details, clean edges, and realistic texture/material rendering when applicable. Demonstrating strong semantic consistency with the given description, overall coherence, faithful style execution, artifact-free generation, distortion-free geometry, and polished high-quality visual appeal.

Figure 4: Common prompts for different benchmarks.

## A.2 Generalization Analysis of PSR

To clarify the scope of PSR's generality, we first discuss its framework-level adaptability and the role of the common prompt.

- **Framework-level Generality of PSR.** PSR does not claim that the same anchors are universally optimal for all evaluators. Instead, the generality of PSR lies at the framework level: the evaluation objective as well as the corresponding designed common prompt are explicit and replaceable inputs to the same prior-discovery, semantic-routing, and recomposition pipeline, rather than a particular quality criterion hard-coded into the sampling algorithm. Therefore, the discovered anchors are metric-induced, while the overall PSR formulation is metric-adaptive. This is consistent with our formulation in Section 3.1 and 3.2, that the evaluation suite defines the favorable prior subspace and the conditioning signal determines how anchors are selected and recomposed within that subspace.
- **Role of the Common Prompt.** Moreover, the suite-aligned common prompt is used only as an offline calibration condition for discovering a reusable anchor pool; it is not optimized for individual test prompts. In addition, PSR can accept either a single metric or a multi-dimensional evaluation suite, since evaluator outputs are used only to score and rank generated candidates and need not be differentiable or coupled to the generative backbone. Changing the evaluation suite therefore changes the induced prior subspace, while leaving the backbone and the procedure unchanged.

**Cross-metric Generalization.** We further examine whether improvement transfers beyond the metrics used to construct the prior subspace. Specifically, on Wan2.2-5B, we restrict both anchor discovery and semantic supervision to the two metrics, TCS and MSS, while evaluating the resulting generations using the full five-dimensional VMBench. We would like to clarify one important detail: the common prompt is redesigned in this experiment. Specifically, we remove descriptions associated with PAS, OIS, and CAS while retaining descriptions corresponding to temporal coherence and

motion smoothness: "Best quality, cinematic high-fidelity footage, high resolution. Demonstrating smooth fluid motion dynamics, high temporal coherence.".

Table 5: Ablation study on the metric suite for prior subspace discovery on Wan2.2-5B evaluated by VMBench. Better results are highlighted in bold.

| Method | PAS | OIS | TCS | MSS | CAS | Avg. |
|---|---|---|---|---|---|---|
| Random | 14.86 | 68.65 | 95.21 | 79.46 | 58.58 | 63.35 |
| Ours (TCS and MSS only) | 21.30 | 69.87 | 95.72 | 83.11 | **63.55** | 66.71 |
| Ours ($\alpha = 0.3$) | **24.84** | **70.43** | **96.31** | **84.00** | 63.32 | **67.78** |

As shown in Table 5, using only TCS and MSS as optimization signals also improves the non-optimized PAS, OIS, and CAS metrics over random initialization. This cross-metric transfer suggests that the discovered prior region is not merely exploiting the two optimized evaluator outputs, but captures generative properties that benefit several related quality dimensions. Using the complete evaluation suite yields the strongest overall performance, further supporting the motivation for metric-induced prior discovery: a broader evaluation objective induces a prior subspace that is better aligned with a broader range of desired qualities.

**Ablation on Semantic Generalization of the Ranker.** To evaluate whether the learned ranker can generalize beyond the semantic concepts observed during training, we conduct a held-out subject split on VMBench. Specifically, we extract the subject nouns from VMBench prompts and split them into disjoint training and testing subsets with an $8{:}2$ ratio. This setting is more challenging, since the test prompts inevitably contain semantic concepts that are absent from the training subject pool. As shown in Table 6, the proposed method still improves the VMBench score from $59.94$ to $61.95$ under this unseen-subject setting. Notably, the gains are consistent on OIS, TCS, MSS, and CAS, with particularly large improvements on OIS and CAS. These results suggest that the ranker does not merely memorize concept-specific anchor assignments for seen subjects. The reason is that our ranker operates in a continuous conditioning space rather than over discrete subject identities. Although the held-out subject nouns are unseen at the string level, their textual embeddings still lie in a semantically structured representation space, where related concepts share nearby or partially aligned contextual features. Moreover, the ranker predicts anchor preferences from first-step velocity fields conditioned on these embeddings, instead of directly classifying each subject into a fixed anchor index. As a result, it learns a continuous compatibility function between semantic conditions and prior anchors, enabling it to infer anchor preferences for unseen concepts.

Table 6: Ablation on semantic generalization of the ranker under an $8{:}2$ held-out subject.

| Method | PAS | OIS | TCS | MSS | CAS | Avg. |
|---|---|---|---|---|---|---|
| Random | **11.92** | 62.19 | 93.84 | 73.32 | 58.44 | 59.94 |
| +Ours | 11.32 | **65.93** | **94.35** | **75.65** | **62.48** | **61.95** |

**Effect of Semantic-pool Coverage.** We vary the proportion of subjects used for ranker training. The full pool contains approximately 300 subjects, so ratios of 0.2–1.0 correspond to roughly 60–300 subjects.

Table 7: Effect of semantic-pool coverage on Wan2.1-1.3B evaluated by VMBench. The full semantic pool contains approximately 300 subject concepts. Better results are highlighted in bold.

| Training Subjects | PAS | OIS | TCS | MSS | CAS | Avg. |
|---|---|---|---|---|---|---|
| Random | 11.46 | 67.27 | 93.77 | 73.42 | 58.18 | 60.82 |
| 20% ($\sim$60) | 11.31 | 68.28 | 93.74 | 74.29 | 60.20 | 61.56 |
| 40% ($\sim$120) | 11.71 | 68.17 | 93.65 | 74.53 | 61.41 | 61.89 |
| 60% ($\sim$180) | **12.42** | 68.27 | 93.86 | 74.40 | 61.94 | 62.18 |
| 80% ($\sim$240) | 10.59 | 69.71 | 93.69 | **75.25** | **62.70** | 62.39 |
| 100% ($\sim$300) | 11.60 | **69.81** | **94.07** | 75.07 | 62.18 | **62.55** |

Table 7 shows a clear overall improvement as semantic coverage increases: the average VMBench score rises from 61.56 with 20% of the subject pool to 62.55 with the complete pool. Performance begins to saturate between 80% and 100% of the pool, suggesting that 240–300 subjects provide sufficient semantic coverage.

This behavior is consistent with the role of the semantic pool in PSR. The ranker learns a continuous compatibility function between textual representations and prompt-conditioned first-step velocity fields. Increasing the number of training subjects provides denser coverage of the semantic space and exposes the ranker to more diverse anchor-preference patterns. This improves its ability to interpolate from isolated subject concepts to longer test prompts. The held-out-subject experiment in Table 6 further supports this interpretation: the ranker remains effective on subjects excluded from training because routing is performed in a continuous semantic representation space rather than through memorized subject identities.

Semantic-pool size and the evaluation suite control two different aspects of PSR. The evaluation suite defines the quality objective used to supervise anchor fitness, whereas the semantic pool determines the coverage of the conditioning space available to the ranker. The consistent improvements across multiple VMBench dimensions suggest that increasing subject coverage benefits condition-aware routing beyond a single aesthetic criterion.

**Image-conditioned Evaluation Dimensions.** During Stage-1 prior-subspace discovery, PSR is conditioned only on the text-based common prompt $p_c$, without an input image. Consequently, the I2V Subject and I2V Background dimensions, which explicitly measure fidelity to the conditioning image, cannot be used for anchor discovery and are excluded from the discovery objective in our I2V implementation. Nevertheless, after prompt- and image-conditioned routing, I2V Subject increases from 98.52 to 98.63, while I2V Background changes only marginally from 99.44 to 99.39. More broadly, PSR improves seven of the ten reported VBench++ dimensions. The decreases on Camera Motion and Imaging Quality therefore do not establish benchmark overfitting; rather, they are consistent with trade-offs among partially competing dimensions in a multi-objective evaluation suite. The manuscript already discusses that emphasizing dynamic expressiveness and temporal stability may not simultaneously optimize camera control or low-level imaging fidelity.

### A.3 Computational Cost and Inference Efficiency

The offline pipeline consists of three stages: (1) generating and evaluating $N = 1000$ candidate noises and retaining the top $K = 50$ anchors; (2) constructing concept–anchor supervision using a semantic pool of approximately 300 subjects; and (3) training the lightweight local ranker. We want to point out that these stages are performed only once for a given backbone and evaluation suite, after which the resulting anchor pool and ranker are reused across all test prompts.

At inference time, the routing overhead mainly includes extracting the prompt-conditioned first-step velocity fields, local ranking, global preference aggregation, and anchor recomposition. To reduce this cost, the unconditional classifier-free-guidance velocity fields associated with the anchors are precomputed and cached, while the prompt-conditioned first-step velocity fields of all 50 anchors are extracted in parallel. More engineering optimizations could further reduce the overhead. A detailed breakdown of the offline and per-prompt computational costs is provided in Table 8. All wall-clock measurements are obtained using Wan2.1-1.3B on 8 NVIDIA H200 GPUs.

Table 8: Computational cost of PSR on Wan2.1-1.3B. Offline stages are performed once for a given backbone and evaluation suite, whereas inference costs are measured per prompt. Wall-clock time is measured on 8 NVIDIA H200 GPUs.

| Stage | Frequency | Configuration | Cost |
|---|---|---|---|
| Prior subspace discovery | Once | $N = 1000$ | 3 h |
| Concept–anchor supervision | Once | 300 subjects, 50 anchors | 46 h |
| Ranker training | Once | 40 epochs | 13 h |
| PSR initialization overhead | Per prompt | 50 anchors in parallel | 24 s |
| Standard generation | Per prompt | Wan2.1-1.3B | 91 s |

Importantly, this additional computation provides a favorable quality–efficiency trade-off. On VMBench, Wan2.1-1.3B equipped with PSR achieves an average score of 62.55, approaching the 63.35 obtained by the randomly initialized 5B model. Meanwhile, PSR requires approximately 115 seconds per video, compared with approximately 150 seconds for the 5B model. Thus, PSR enables the substantially smaller 1.3B backbone to approach the average benchmark performance of the 5B model while retaining approximately 23% lower end-to-end inference latency.

We also emphasize that the computational cost of PSR is configurable through the anchor-pool size rather than being an unavoidable fixed overhead. We therefore evaluate a reduced setting with only $K = 25$ anchors, while keeping all other components unchanged. As shown in Table 9, the $K = 25$ configuration achieves an average score of 62.51, nearly matching the 62.55 obtained with 50 anchors, while halving the number of prompt-conditioned velocity features and graph nodes. Its routing latency decreases from 24 seconds to 13 seconds, corresponding to an approximately 14% overhead relative to the 91-second standard generation time. This result demonstrates a practical quality–efficiency trade-off: a smaller anchor pool preserves nearly all of the performance gain while substantially reducing test-time cost. **For few-step or latency-sensitive models, the routing budget can likewise be reduced by using fewer anchors, thereby mitigating the relative overhead while retaining most of PSR's performance benefits.**

Table 9: Effect of the anchor pool size on generation quality and inference overhead using Wan2.1-1.3B evaluated by VMBench. Routing latency denotes the additional per-prompt latency introduced by PSR. Better results are highlighted in bold.

| Method | Anchors | PAS | OIS | TCS | MSS | CAS | Avg. | Routing Latency |
|---|---|---|---|---|---|---|---|---|
| Random | – | 11.46 | 67.27 | 93.77 | 73.42 | 58.18 | 60.82 | 0 s |
| PSR | 25 | 11.55 | 69.67 | **94.27** | 75.02 | 62.04 | 62.51 | **13 s** |
| PSR | 50 | **11.60** | **69.81** | 94.07 | **75.07** | **62.18** | **62.55** | 24 s |

**Ranker Architecture and Training.** Specifically, our ranker is a lightweight MMDiT-style dual-stream model comprising a shared 3D patch encoder, a two-layer visual Transformer, and four text–video joint-attention blocks with a hidden dimension of 256 and eight attention heads. Candidate scores are produced by a $256 \to 128 \to 1$ MLP head. The ranker is trained for 40 epochs with $K = 5$ candidates per list, optimizing a combination of pairwise margin loss and NeuralNDCG loss.

In summary, our proposed PSR comprises three clearly separated and well-motivated stages. Each stage addresses a distinct limitation and is supported by controlled ablations. Moreover, PSR is a mechanism-level framework rather than a task-specific heuristic: the same formulation is instantiated across T2I, T2V, and I2V, across different backbones and model scales, without modifying the underlying generator. Its computational budget is also adjustable. Therefore, PSR offers a clear, modular, and broadly applicable initialization framework with a practical and configurable quality–efficiency trade-off.

### A.4 Comparison with Existing Noise Initialization Methods

We first claim that the experimental design follows the specific motivation and scope stated in our Introduction that existing initial-noise methods are generally developed for a particular generation regime and a predefined quality objective. PSR addresses this mechanism-level limitation through a unified metric-induced discovery, semantic-conditioned routing, and recomposition framework that can be instantiated across T2I, T2V, and I2V tasks and under different evaluation objectives. Accordingly, our experiments on heterogeneous backbones and benchmarks are designed to test this framework-level generality. Random Gaussian initialization is the standard inference procedure of the underlying generative model, and therefore provides the natural reference for measuring the benefit introduced by PSR without modifying the generator itself.

To provide a direct and controlled comparison and further demonstrate the effectiveness of PSR, we additionally instantiate PSR on SDXL, the same generative backbone used by NPNet (Golden Noise) (Zhou et al., 2025). In this experiment, both Stage-1 anchor discovery and the subsequent semantic-conditioned supervision are constructed exclusively using HPSv2. We then evaluate the resulting model using HPSv2 as well as two additional metrics, ImageReward and CLIPScore.

Table 10: Comparison with NPNet (Golden Noise) on SDXL. PSR uses only HPSv2 for prior subspace discovery and semantic-conditioned supervision, while ImageReward and CLIPScore serve as additional evaluation metrics. Better results are highlighted in bold.

| Method | Model | HPSv2 | ImageReward | CLIPScore |
|---|---|---|---|---|
| Standard (Random) (Zhou et al., 2025) | SDXL | 29.71 | 96.63 | 0.8734 |
| NPNet (Zhou et al., 2025) | SDXL | 29.88 | 98.81 | 0.8813 |
| PSR ($\alpha = 0.3$) | SDXL | **33.23** | **108.53** | **0.8991** |

Under the same backbone and evaluation setting, PSR consistently outperforms NPNet across all three metrics. In particular, although only HPSv2 is used to induce the anchor subspace and supervise the routing process, PSR also improves ImageReward and CLIPScore. This result provides two complementary observations: first, PSR is competitive with a representative learned golden-noise method under its native generation regime; second, the induced prior subspace exhibits measurable transfer to evaluation dimensions that are not directly used during optimization.

We therefore make the scope of our claim explicit: the current results establish consistent improvements over standard initialization across tasks and backbones, while the SDXL experiment establishes direct competitiveness with a representative golden-noise method in a shared setting.

### A.5 Analysis of Prior Anchor Selection and Recomposition

Table 11: Comparison of different anchor selection and recomposition strategies on Wan2.1-1.3B evaluated by VMBench. Common denotes fixed anchors selected solely by their offline scores under the common prompt; Sim denotes semantic-similarity-based retrieval; Ranked denotes condition-aware selection using the learned ranker. Better results are highlighted in bold.

| Method | PAS | OIS | TCS | MSS | CAS | Avg. |
|---|---|---|---|---|---|---|
| Random | 11.46 | 67.27 | 93.77 | 73.42 | 58.18 | 60.82 |
| Common-Top1 | 9.51 | 68.03 | 92.54 | 73.12 | 59.43 | 60.52 |
| Common-Top2 ($\alpha = 0.3$) | 11.71 | 68.30 | 93.39 | 73.62 | 59.36 | 61.28 |
| Sim-Top1 | 8.45 | 69.00 | 93.95 | 73.36 | 61.12 | 61.17 |
| Sim-Top2 ($\alpha = 0.3$) | 9.07 | 68.26 | **94.34** | 73.86 | 60.57 | 61.22 |
| Ranked-Top1 | 11.44 | 68.74 | 93.58 | 74.24 | 60.97 | 61.79 |
| PSR-Top3 | **11.72** | **69.88** | 93.89 | 74.91 | 61.99 | 62.48 |
| PSR-Top2 ($\alpha = 0.3$) | 11.60 | 69.81 | 94.07 | **75.07** | **62.18** | **62.55** |

**Fixed global anchors are insufficient.** To examine whether the reported gains can be attributed simply to a few globally superior, evaluator-favored initial anchors, we conduct a controlled experiment. Specifically, we directly apply the globally highest-scoring anchor, ranked solely by its offline score under the common prompt, to every test prompt without semantic-conditioned routing. We denote this baseline as Common-Top1. Importantly, Common-Top1 is different from the Ranked-Top1, the $\alpha = 0$ setting in Section 4.3 and Figure 3, which selects a prompt-specific anchor using the learned ranker. We also evaluate a fixed recomposition of the two globally highest-scoring offline anchors, denoted as Common-Top2. Table 11 shows that offline global anchor selection alone does not explain the final improvement. Common-Top1 performs slightly worse than random initialization on average and degrades several dimensions. A fixed recomposition of the two globally highest-scoring anchors provides only a modest gain, increasing the average score from 60.82 to 61.28. In contrast, prompt-conditioned routing over the same anchor pool improves the average score to 61.79, and prompt-specific Top-2 recomposition further increases it to 62.55. These results indicate that prior subspace discovery does not simply identify several universally superior seeds. Instead, it constructs a candidate prior subspace, and the principal gains arise from selecting and recomposing anchors according to the test condition.

**Semantic similarity is insufficient for anchor selection.** To evaluate whether a substantially simpler selection mechanism can replace the proposed ranking pipeline, we introduce a semantic-similarity retrieval baseline. For each test prompt, we retrieve the most semantically similar subject from the offline semantic pool and directly use the highest-scoring anchor associated with that subject. We denote this method as Sim-Top1. Sim-Top2 similarly recomposes the two highest-scoring anchors associated with the retrieved subject. These baselines do not require first-step velocity features, a learned ranker, or graph-based preference aggregation. Table 11 shows that the similarity-based baselines yield only modest gains over Random, improving the average score from 60.82 to 61.17/61.22, while substantially degrading PAS. This suggests that static prompt–subject similarity is insufficient for identifying dynamically compatible anchors. In contrast, Ranked-Top1 reaches 61.79, and the full PSR further improves the score to 62.55, supporting the value of velocity-aware ranking and global preference aggregation. Finally, PSR is not restricted to two-anchor recomposition. The Top-3 variant achieves 62.48, comparable to the Top-2 result of 62.55 and clearly above Ranked-Top1. We use $M = 2$ in the main experiments, making Top-2 a more efficient and robust practical choice.