LATENTWEAVE: ARE SEMANTICS MISSING FROM VAE LATENTS—OR MERELY IMPLICIT?

Anonymous authors
Paper under double-blind review

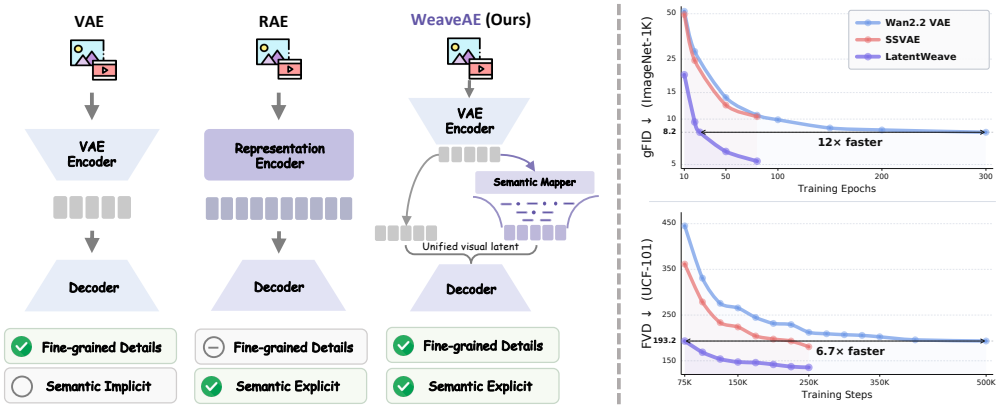


Figure 1: WeaveAE recovers and reorganizes semantics in a pretrained VAE latent, while retaining fine-grained visual information (left) and enabling more efficient visual generative learning (right).

ABSTRACT

Visual encoders define the representation spaces underlying visual reconstruction, understanding, and generation. Reconstruction-oriented VAEs provide compact representations with strong visual fidelity, yet their latents are shaped primarily by reconstruction and exhibit weak semantic organization. Recent approaches address this limitation by replacing VAE encoders with pretrained representation encoders. We revisit whether such replacement is necessary and find that substantial semantic information remains recoverable from pretrained VAE latents despite being poorly exposed in their native organization. Motivated by this observation, we introduce ***LatentWeave***, a unified framework for constructing and generatively modeling semantically organized visual latents from pretrained VAE representations. It comprises ***WeaveAE***, which organizes complementary detail and semantic strands, and ***WeaveDiT***, which explicitly exploits this dual-strand structure during generative modeling. We further introduce *Self-REPA*, which reuses WeaveAE's semantic representation as an internal alignment target to improve generative learning without introducing an external vision encoder during generator training. Across image and video benchmarks, LatentWeave improves semantic accessibility and generative learning efficiency while preserving reconstruction fidelity. On UCF-101, WeaveAE achieves an rFVD of 3.11 for reconstruction, while WeaveDiT achieves an FVD of 135.49 for class-conditional generation. Our results establish latent reorganization from pretrained VAE representations as a practical paradigm for unified visual representation learning and efficient generative modeling.

1 INTRODUCTION

Visual representations determine both what information is retained from images and videos and how that information is organized for subsequent modeling (Bengio et al., 2013). Visual representation learning has largely developed along two complementary directions. Reconstruction-oriented VAEs learn compact latent spaces that preserve fine-grained appearance and spatiotemporal detail for faithful reconstruction (Rombach et al., 2022; Wan et al., 2025; Black Forest Labs, 2025), whereas representation encoders trained with self-supervised or multimodal objectives learn semantically

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.

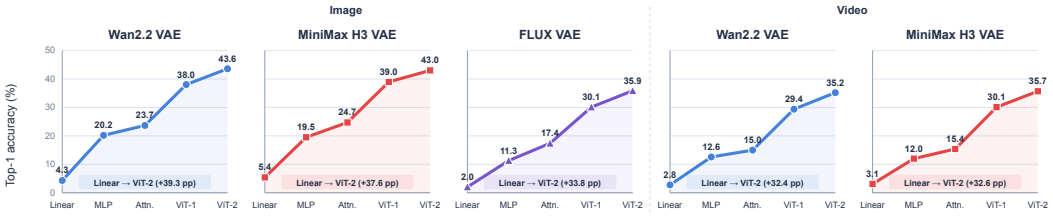


Figure 2: Semantic accessibility of frozen VAE latents under increasingly expressive probes. The probes are linear, MLP, attentive pooling, and ViT probes with one or two Transformer blocks.

structured visual features (Radford et al., 2021; He et al., 2022; Siméoni et al., 2025; Tschannen et al., 2025). The resulting representation spaces emphasize complementary properties: reconstruction fidelity and compactness in VAE latents, and semantic organization in representation features.

Despite their effectiveness for latent generative modeling, reconstruction-trained VAE latents are optimized primarily for faithful reconstruction and often exhibit weaker semantic organization than features learned by pretrained representation encoders (Peebles & Xie, 2023). This gap has motivated recent efforts to introduce semantic structure from external vision foundation models into generative modeling, typically through auxiliary representation alignment applied to VAE latents or intermediate features of the generative model (Yu et al., 2025; Yao et al., 2025; Leng et al., 2025). A more direct approach replaces the reconstruction-trained VAE encoder with a pretrained representation encoder and performs generative modeling directly in its feature space (Zheng et al., 2026; Shi et al., 2026). These approaches further demonstrate that semantically structured representations can improve both generative learning and generation quality. Yet they leave a more fundamental question open: *when VAE latents appear semantically weak, is semantic information largely absent from the representation, or does it remain encoded but poorly exposed by its reconstruction-oriented organization?*

To distinguish semantic absence from limited accessibility, we conduct a controlled probing study on frozen pretrained VAE latents across images and videos, as illustrated in Figure 2. We keep each VAE encoder fixed and vary only how its latent representation is read out for semantic recognition. We progressively increase the expressiveness of the readout, from linear probing to MLP, attentive pooling, and shallow ViTs, while keeping the pretrained VAE encoder frozen (Bardes et al., 2024). Across VAE families on ImageNet-1K and Kinetics-700 (Russakovsky et al., 2015; Smaira et al., 2020), we observe a consistent pattern: although frozen VAE latents appear semantically weak under direct linear readout, substantially more class-discriminative information is already accessible with lightweight nonlinear probes, with further gains from more expressive probes. Taken together, these results suggest that weak linear semantic accessibility in reconstruction-oriented VAE latents should not be conflated with an absence of semantic information.

> **Observation.** Semantic weakness in VAE latents does not necessarily imply semantic absence, yet its accessibility remains limited, requiring expressive nonlinear probes to reveal richer semantic structure.

Motivated by this observation, we introduce **LatentWeave**, a framework that reorganizes information retained in pretrained VAE latents into a semantically accessible representation for reconstruction and generative modeling. To realize this representation, we introduce **WeaveAE**, which constructs complementary *detail* and *semantic* strands from the same frozen VAE latent. The detail strand retains reconstruction-oriented visual information, while a lightweight *Semantic Mapper* recovers semantic structure that is compressed into the semantic strand. WeaveAE then weaves the two strands into a unified visual latent for image and video reconstruction. It achieves an rFID of **0.15** on ImageNet-1K and an rFVD of **1.16** on Kinetics-600 for image and video reconstruction, respectively.

The resulting two-strand representation motivates a generative model that explicitly accounts for its internal structure. We therefore introduce **WeaveDiT**, which first models the detail and semantic strands through strand-specific pathways with cross-strand interaction, and then fuses them for shared joint modeling of the unified latent. Across image and video benchmarks, LatentWeave enables substantially more efficient generative learning, requiring approximately **12**$\times$ fewer training epochs on ImageNet-1K and **6.7**$\times$ fewer training steps on UCF-101 to reach comparable generation quality, as illustrated in Figure 1. It achieves an FID of **5.26** on ImageNet-1K for class-conditional image generation and an FVD of **135.49** on UCF-101 for class-conditional video generation.

Our main contributions are summarized as follows:

- We revisit the semantic weakness of reconstruction-oriented VAE latents through a controlled probing study across image and video VAEs, suggesting that their limitation lies not only in what information is retained, but also in how that information is organized and exposed.
- We introduce **WeaveAE**, which recovers semantic structure from pretrained VAE latents, organizes complementary detail and semantic strands, and weaves them into a unified visual latent supporting reconstruction, semantic accessibility, and generation.
- We introduce **WeaveDiT**, which explicitly exploits the dual-strand structure of the unified visual latent through strand-specific processing, cross-strand interaction, and joint modeling, enabling efficient generative learning across images and videos.

2 Related Work

Visual representations for latent generation. Visual autoencoders provide the representation spaces on which latent generative models operate (Hinton & Salakhutdinov, 2006; Vincent et al., 2008). Reconstruction-oriented VAEs have therefore become a standard choice for scalable image and video generation, with recent systems improving spatiotemporal compression, reconstruction quality, and latent efficiency (Kingma & Welling, 2013; Labs, 2024; Brooks et al., 2024; Yang et al., 2024; Kong et al., 2024; Wan et al., 2025; Cheng & Yuan, 2025). Beyond reconstruction, several works further reshape VAE latents to make them more suitable for generative learning through semantic alignment or additional latent regularization (Yao et al., 2025; Page et al., 2026; Liu et al., 2025). A more direct approach replaces the VAE encoder with pretrained representation encoders to obtain semantically structured generative spaces (Zheng et al., 2026; Xie et al., 2026; Guo et al., 2026). To complement these semantically oriented representations with finer visual detail, recent methods combine reconstruction- and representation-oriented features through complementary encoding pathways (Gong et al., 2026; Shi et al., 2026). **WeaveAE** instead retains the pretrained reconstruction-oriented VAE encoder and recovers semantic structure directly from its latent, organizing it together with reconstruction-oriented information in a unified visual representation.

Diffusion Transformers for visual generation. Diffusion Transformers (DiTs) operate on latent visual tokens and scale effectively to image and video generation (Peebles & Xie, 2023; Ma et al., 2025). Recent work further improves DiT through representation guidance and specialized architectures: REPA aligns intermediate generative features with pretrained visual representations (Yu et al., 2025; Wu et al., 2025), DDT separates representation extraction from velocity prediction (Wang et al., 2026), and MMDiT introduces stream-specific parameterization with joint interaction between heterogeneous representations (Esser et al., 2024). Inspired by the latter design, **WeaveDiT** applies strand-specific and joint modeling to complementary structures within a single visual latent.

3 Methodology

Our probing study in Figure 2 shows that weak linear semantic accessibility does not necessarily imply semantic absence: nonlinear and more expressive probes reveal substantially richer semantic structure from the same frozen VAE latent. This motivates reorganizing, rather than replacing, the pretrained VAE representation. The challenge is twofold: exposing this semantic structure in a compact form without sacrificing reconstruction-oriented information, and effectively modeling the resulting representation for generation. To this end, we introduce **LatentWeave**, a framework for constructing and generatively modeling unified visual representations from pretrained VAE latents. It comprises two core components: **WeaveAE**, which organizes complementary detail and semantic strands into a unified visual representation, and **WeaveDiT**, which explicitly exploits this dual-strand organization during generative modeling. We further introduce *Self-REPA*, which reuses the semantic representation learned by WeaveAE as an internal alignment target to improve generative learning.

3.1 WeaveAE: Unified Visual Latent Construction

WeaveAE addresses the representation-side challenge of exposing recoverable semantic structure while retaining the reconstruction-oriented information of the pretrained VAE. Its design therefore starts from a single frozen VAE representation and organizes complementary detail and semantic information while preserving their spatiotemporal correspondence, as illustrated in Figure 3.

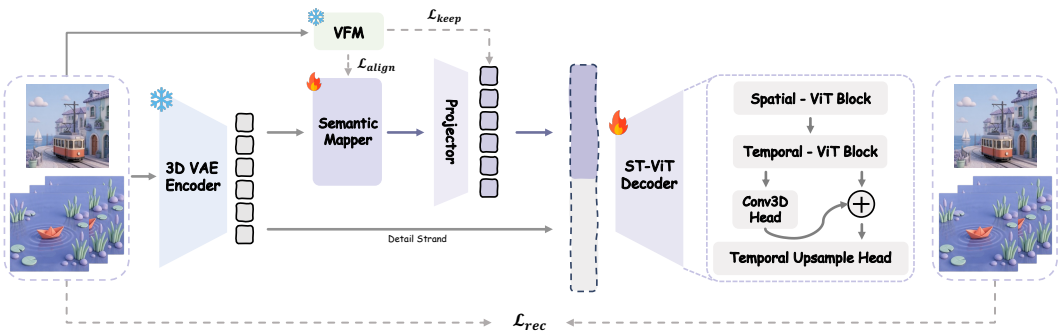


Figure 3: Overview of **WeaveAE**. A frozen VAE latent is organized into complementary detail and semantic strands, which are woven into a unified visual latent for image and video reconstruction.

Native VAE Representation. Given an image or video x, a frozen 3D VAE encoder E_{vae} produces the native latent $z_{\text{vae}} = E_{\text{vae}}(x) \in \mathbb{R}^{C_d \times T' \times H' \times W'}$, where T', H', and W' denote the compressed temporal and spatial dimensions. This representation serves as the basis for constructing the detail strand z_d, which retains the reconstruction-oriented visual information from the pretrained VAE.

Semantic Mapper. Motivated by the recoverable semantics observed in Figure 2, we introduce a lightweight Semantic Mapper M_{sem}. Given the frozen VAE latent z_{vae}, the mapper learns a nonlinear contextual transformation that produces a semantically structured intermediate representation:

$$h_s = M_{\text{sem}}(z_{\text{vae}}), \qquad h_s \in \mathbb{R}^{D_s \times T' \times H' \times W'}. \tag{1}$$

Here, D_s denotes the semantic feature dimension. The Semantic Mapper receives only z_{vae} as its visual input and is trained under the supervision of a frozen vision foundation model.

Latent Weaving. The recovered semantic feature h_s is high-dimensional but already aligned with the spatiotemporal grid of the VAE latent. We therefore introduce a lightweight semantic projector P_{sem} that compresses only the channel dimension:

$$z_s = P_{\text{sem}}(h_s), \qquad z_s \in \mathbb{R}^{C_s \times T' \times H' \times W'}. \tag{2}$$

Here, C_s denotes the compact semantic dimension. Since the projection is applied independently at each spatiotemporal location, z_s preserves the token-wise correspondence with the detail strand. The two strands are then woven along the channel dimension to construct the unified visual latent:

$$z_u = z_d \oplus z_s, \tag{3}$$

where $\oplus$ denotes channel-wise concatenation. Each unified token therefore combines reconstruction-oriented and semantic information at the same spatiotemporal location.

Spatial–Temporal ViT Decoder. Following recent visual autoencoders with ViT-based decoders (Hansen-Estruch et al., 2025; Zheng et al., 2026), we employ a Spatial–Temporal ViT (ST-ViT) decoder to reconstruct images and videos from the unified latent, $\hat{x} = D_{\text{stvit}}(z_u)$, as illustrated in Figure 3. The decoder uses factorized spatial and temporal Transformer blocks (Arnab et al., 2021) to model intra-frame and inter-frame dependencies, followed by a lightweight Conv3D residual branch and causal temporal upsampling for videos. This design reduces decoder computation by 89.4% relative to the original VAE decoder while maintaining high-fidelity reconstruction.

3.2 WeaveAE Training Objective

We train WeaveAE while keeping both the pretrained VAE encoder and the vision foundation model (VFM) frozen. Its objective combines semantic alignment, semantic preservation through compression, and reconstruction from the resulting unified latent across both images and videos.

Spatiotemporal Semantic Alignment. We use a frozen VFM to provide semantic targets for the Semantic Mapper. Because the VFM and VAE operate at different spatiotemporal resolutions, we align the VFM features to the spatiotemporal grid of h_s. For an image x^I, we spatially resize the VFM feature as $\widetilde{\mathbf{G}}^I = \mathcal{R}(G(x^I)) \in \mathbb{R}^{H' \times W' \times D}$, where $\mathcal{R}$ resizes to (H', W').

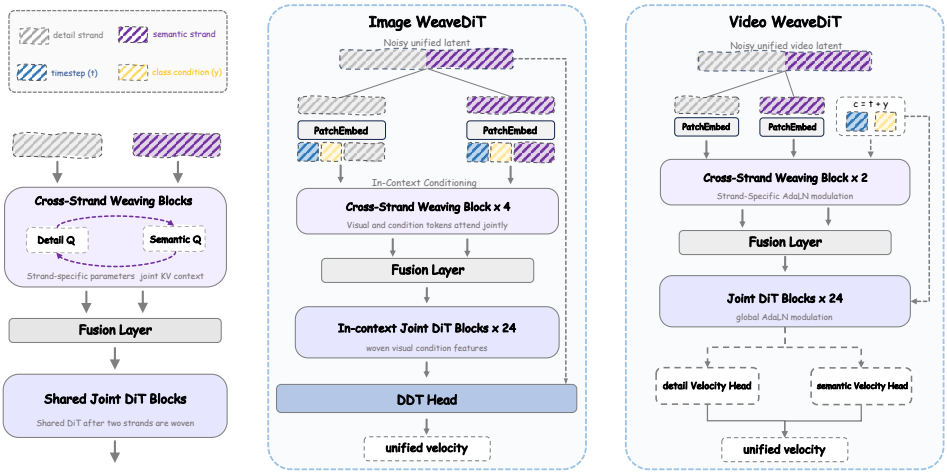


Figure 4: Overview of **WeaveDiT**. The two strands are processed separately with cross-strand interaction, then fused for joint modeling; variants use modality-aware conditioning and prediction.

For a video $x^V = \{x_t^V\}_{t=0}^{T-1}$, we first obtain $\mathbf{G}_t^V = \mathcal{R}(G(x_t^V))$ and temporally aggregate these features to match the causal compression of the VAE:

$$\widetilde{\mathbf{G}}_k^V = \begin{cases} \mathbf{G}_0^V, & k = 0, \\ \dfrac{1}{r}\displaystyle\sum_{j=1}^{r} \mathbf{G}_{r(k-1)+j}^V, & k = 1, \dots, T' - 1, \end{cases} \tag{4}$$

where r denotes the temporal compression factor and $T = 1 + r(T' - 1)$. Following temporal feature averaging (Bolya et al., 2025), this construction aligns the frame-wise VFM features with the causal temporal grid of the VAE. After flattening the aligned teacher representation $\widetilde{\mathbf{G}}^m$ into $N = T'H'W'$ tokens, where $m \in \{I, V\}$ denotes the modality, we optimize

$$\mathcal{L}_{\text{align}} = \frac{1}{N}\sum_{i=1}^{N} \ell_{\cos}\left(h_{s,i}, \widetilde{g}_i^m\right), \tag{5}$$

where $\ell_{\cos}$ denotes the cosine distance for semantic alignment.

Semantic Preservation after Compression. To preserve semantic structure when compressing h_s into the compact strand z_s, we introduce a training-only up-projector $\widehat{h}_s = U_{\text{sem}}(z_s)$. Following similarity-preserving representation alignment (Tung & Mori, 2019; Singh et al., 2026a), we apply an auxiliary objective $\mathcal{L}_{\text{keep}}$ between $\widehat{h}_s$ and the aligned target representation, encouraging z_s to retain semantic consistency and inter-token structure through the channel bottleneck.

Reconstruction and Optimization. We use a reconstruction objective $\mathcal{L}_{\text{rec}}$ combining pixel-wise, LPIPS (Zhang et al., 2018), and adversarial losses (Goodfellow et al., 2014), with adversarial supervision applied only during decoder refinement. The overall WeaveAE objective is

$$\mathcal{L}_{\text{WeaveAE}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{keep}}\mathcal{L}_{\text{keep}}. \tag{6}$$

The corresponding loss weights and optimization details are provided in the Appendix A.1.

3.3 WeaveDiT: Dual-Strand Generative Modeling

We model the unified latent z_u using flow matching (Liu et al., 2022; Ma et al., 2024). Given a clean unified latent z_u and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, we construct the interpolated latent $z_t = (1-t)z_u + t\epsilon$ and train the generative model to predict the corresponding velocity $v_t = \epsilon - z_u$:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{z_u,\epsilon,t}\left[\|v_\theta(z_t, t, c) - v_t\|_2^2\right], \tag{7}$$

where c denotes the optional generation condition. We apply the same timestep to both strands, treating them as complementary components of a single generative trajectory. Standard DiTs (Peebles & Xie, 2023) embed all latent channels into a shared token representation without explicitly modeling their strand-specific structure. Inspired by the stream-specific parameterization and joint interaction of MMDiT (Esser et al., 2024), we introduce WeaveDiT to explicitly model the dual-strand structure of the unified latent produced by WeaveAE throughout the generative process, as illustrated in Figure 4.

Cross-Strand Weaving and Joint Modeling. WeaveDiT embeds the noisy detail and semantic strands separately and processes them with strand-specific pathways. Inspired by MMDiT, each strand retains its own projections while attending jointly to both representations:

$$\mathrm{WeaveAttn}_r(x_d, x_s) = \mathrm{Attn}\left(Q_r,\ [K_d \| K_s],\ [V_d \| V_s]\right), \qquad r \in \{d, s\}, \tag{8}$$

where $[\cdot\|\cdot]$ denotes sequence-wise concatenation. This enables cross-strand interaction while preserving strand-specific transformations. After several weaving blocks, the two representations are fused into x_{joint} and processed by shared Joint DiT blocks for unified velocity prediction.

Modality-Aware Conditioning and Prediction. As illustrated in Figure 4, WeaveDiT uses modality-specific conditioning and prediction mechanisms. For images, timestep and class conditions are introduced as in-context tokens (Geng et al., 2026), with a DDT head for unified velocity prediction (Wang et al., 2026). For videos, whose spatiotemporal sequences are substantially longer, we instead inject global conditions through AdaLN, with strand-specific heads predicting the detail and semantic velocities. The two variants are configured with comparable parameter budgets.

Self-REPresentation Alignment (Self-REPA). We reuse the frozen Semantic Mapper as an internal target for WeaveDiT. Following representation alignment (Yu et al., 2025; Singh et al., 2026b), a lightweight projector P_{rep} aligns intermediate joint tokens $h_\theta^{(\ell)}$ with stop-gradient mapper features h_s via $\mathcal{L}_{\text{Self-REPA}} = N^{-1}\sum_i[1 - \cos(P_{\text{rep}}(h_{\theta,i}^{(\ell)}), \mathrm{sg}(h_{s,i}))]$. This encourages the generative representation to preserve WeaveAE's semantic organization without additional inference-time modules. As shown in Figure 5, Self-REPA further improves generative convergence while using the same frozen WeaveAE representation.

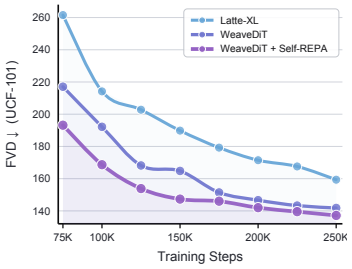


Figure 5: Self-REPA improves generative learning on UCF-101.

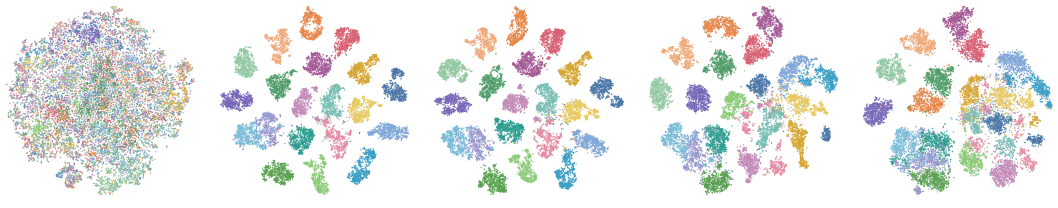
4 Experiments

We evaluate whether LatentWeave recovers accessible semantics from reconstruction-oriented VAE latents without sacrificing reconstruction fidelity, while supporting efficient generative modeling. We study four questions: **(i) Semantic recovery** (Section 4.2), whether semantics remain accessible after compression and weaving; **(ii) Reconstruction** (Section 4.3), whether fidelity is preserved; **(iii) Generative modeling** (Section 4.4), whether WeaveDiT learns the unified latent efficiently; and **(iv) Analysis and Discussion** (Section 4.5), which design principles govern these properties.

4.1 Experimental Setup

Datasets. We train WeaveAE on ImageNet-1K (Russakovsky et al., 2015) and Kinetics-600 (Carreira et al., 2018). Semantic accessibility is evaluated on ImageNet-1K and UCF-101 (Soomro et al., 2012); reconstruction on ImageNet-1K, MS-COCO 2017 (Lin et al., 2014), Kinetics-600, and UCF-101; and generation on ImageNet-1K, UCF-101, and SkyTimelapse (Xiong et al., 2018). All experiments use 256×256 resolution, video experiments use 17-frame clips.

Implementation. WeaveAE uses a frozen Wan2.2 VAE encoder (Wan et al., 2025), with frozen SigLIP2 (Tschannen et al., 2025) providing training-time semantic supervision. For generative modeling, we compare against DiT-XL (Peebles & Xie, 2023) for images and Latte-XL (Ma et al., 2025) for videos, using WeaveDiT configurations with comparable parameter budgets. Full architectural, training, and evaluation details are provided in Appendix A.



(a) Wan2.2 VAE z_{vae} (b) SigLIP2 (c) Semantic Mapper h_{s} (d) Semantic strand z_{s} (e) Unified latent z_{u}

Figure 6: t-SNE visualization of semantic organization across visual representations.

Table 1: Linear-probing accuracy (%) across frozen visual representations.

Probe feature	**Feature dim.**	**Image (ImageNet-1K)**		**Video (UCF-101)**	
		Top-1↑	**Top-5↑**	**Top-1↑**	**Top-5↑**
Wan2.2 VAE	48	4.33	13.13	18.51	45.58
SigLIP2	1152	85.15	97.23	92.07	98.56
Semantic Mapper h_{s}	1152	78.90	94.76	86.24	97.84
Semantic strand z_{s}	48	65.68	85.85	72.69	92.73
Unified latent z_{u}	96	64.97	85.90	72.41	93.42

4.2 Recovering Implicit Semantics from Frozen VAE Latents

Figure 2 shows that semantic information in frozen VAE latents can be exposed by nonlinear and more expressive probes. We next ask whether WeaveAE reorganizes this information into a directly accessible representation. To isolate the representation itself, we freeze all features and apply the same linear-probing protocol to the native VAE latent, SigLIP2 representation, Semantic Mapper output h_s, semantic strand z_s, and unified latent z_u on ImageNet-1K and UCF-101.

As shown in Table 1, the Semantic Mapper raises Top-1 accuracy from 4.33% to 78.90% on ImageNet-1K and from 18.51% to 86.24% on UCF-101, approaching the frozen SigLIP2 representation. After compression, z_s retains 65.68% and 72.69%, while the unified latent z_u achieves 64.97% and 72.41%, showing that semantic accessibility is largely preserved through compression and weaving. Figure 6 shows the same trend geometrically: clear class structure emerges in h_s and remains visible in z_s and z_u. Together, these results show that WeaveAE recovers weakly accessible semantics into a compact, linearly accessible representation across both image and video settings.

4.3 Video Reconstruction

Following standard visual-tokenizer evaluation protocols (Hansen-Estruch et al., 2025; Zheng et al., 2026; Fan et al., 2025; Tan et al., 2025), we evaluate 17-frame clips at 256×256 resolution on Kinetics-600 and UCF-101 (Carreira et al., 2018; Soomro et al., 2012), reporting PSNR, SSIM, and rFVD (Unterthiner et al., 2018). As shown in Table 2, WeaveAE achieves the best rFVD on both Kinetics-600 (1.16) and UCF-101 (3.11), while maintaining PSNR and SSIM comparable to the pretrained Wan2.2 VAE. These results show that organizing a complementary semantic strand preserves the appearance and motion information required for faithful video reconstruction. Image reconstruction exhibits the same trend; full results are provided in Appendix B.2.

Table 2: Video reconstruction on Kinetics-600 and UCF-101.

Method	**Kinetics-600**			**UCF-101**		
	PSNR↑	**SSIM↑**	**rFVD↓**	**PSNR↑**	**SSIM↑**	**rFVD↓**
CogVideoX VAE (Yang et al., 2024)	35.03	0.93	3.09	<u>36.42</u>	0.94	4.77
Wan2.1 VAE (Wan et al., 2025)	34.65	<u>0.94</u>	1.83	35.90	<u>0.95</u>	3.34
HunyuanVideo VAE (Kong et al., 2024)	**35.63**	0.93	<u>1.35</u>	**36.78**	0.93	<u>3.28</u>
LeanVAE (Cheng & Yuan, 2025)	30.04	0.88	4.72	31.15	0.91	9.79
Wan2.2 VAE (Wan et al., 2025)	<u>35.17</u>	**0.95**	1.49	36.28	**0.96**	4.11
SSVAE (Liu et al., 2025)	32.99	0.92	4.29	33.89	0.94	7.91
WeaveAE	33.88	<u>0.94</u>	**1.16**	35.21	**0.96**	**3.11**

Table 3: Class-conditional generation on UCF-101 and unconditional generation on SkyTimelapse.

Method	#Tokens	Config	#Params	UCF-101			SkyTimelapse	
				FVD↓	KVD↓	IS↑	FVD↓	KVD↓
CogVideoX VAE (Yang et al., 2024)	1280	t4s8c16	675M	208.11	19.21	64.55	59.20	2.26
Wan2.1 VAE (Wan et al., 2025)	1280	t4s8c16	675M	179.14	13.31	66.35	40.24	1.24
HunyuanVideo VAE (Kong et al., 2024)	1280	t4s8c16	675M	229.32	50.57	66.40	48.67	1.52
LeanVAE (Cheng & Yuan, 2025)	1280	t4s8c16	675M	175.33	–	–	95.15	–
Wan2.2 VAE (Wan et al., 2025)	1280	t4s16c48	675M	212.19	21.45	64.19	40.86	1.17
SSVAE (Liu et al., 2025)	1280	t4s16c48	675M	180.63	22.21	65.46	41.79	**1.16**
LatentWeave	1280	t4s16c96	685M	**135.49**	**7.34**	**69.89**	**34.28**	1.18

4.4 Video Generative Modeling and Learning Efficiency

We train all generators with flow matching (Liu et al., 2022), with Self-REPA used for LatentWeave unless otherwise specified. Following Latte-XL (Ma et al., 2025), we train the video models for 250K steps and evaluate 2,048 generated 17-frame clips at 256×256 using FVD, KVD, and IS, while keeping model scale, sampling, and evaluation fixed across controlled comparisons. As shown in Table 3, LatentWeave achieves the best FVD (135.49), KVD (7.34), and IS (69.89) on UCF-101, together with the best FVD (34.28) on SkyTimelapse. Figure 1 further shows substantially faster generative learning: LatentWeave requires approximately 6.7× fewer training steps on UCF-101 and 12× fewer training epochs on ImageNet-1K to reach comparable generation quality. Qualitative samples in Figure 7 further illustrate the resulting class-conditional video generation. Full image-generation results, including additional quantitative comparisons, are provided in Appendix B.3.



Figure 7: Qualitative class-conditional video generation on UCF-101 at 256×256 resolution.

4.5 Analysis and Discussion

Can recovered semantics replace reconstruction-oriented information?

Recovering semantic structure from the VAE latent does not necessarily make its reconstruction-oriented information redundant. To examine this, we decode from the detail strand z_d, the semantic strand z_s, or their unified representation z_u while keeping the jointly trained decoder fixed. As shown in Table 4, masking either strand degrades both image and video reconstruction, whereas the unified latent consistently performs best. The qualitative results in Figure 8 further clarify this behavior: when decoded independently, the semantic strand preserves recognizable visual structure but lacks the fine-grained appearance and texture required for faithful reconstruction, while detail-only decoding also degrades reconstruction. These results show that recovered semantic structure complements, rather than replaces, the reconstruction-oriented information inherited from the pretrained VAE.

Decoder input	Image		Video	
	PSNR↑	rFID↓	PSNR↑	rFVD↓
Detail strand z_{d}	22.51	3.52	25.28	68.27
Semantic strand z_{s}	19.45	6.57	23.11	73.01
Unified latent z_{u}	**29.15**	**0.15**	**35.21**	**3.11**

Table 4: Reconstruction from individual and unified strands.

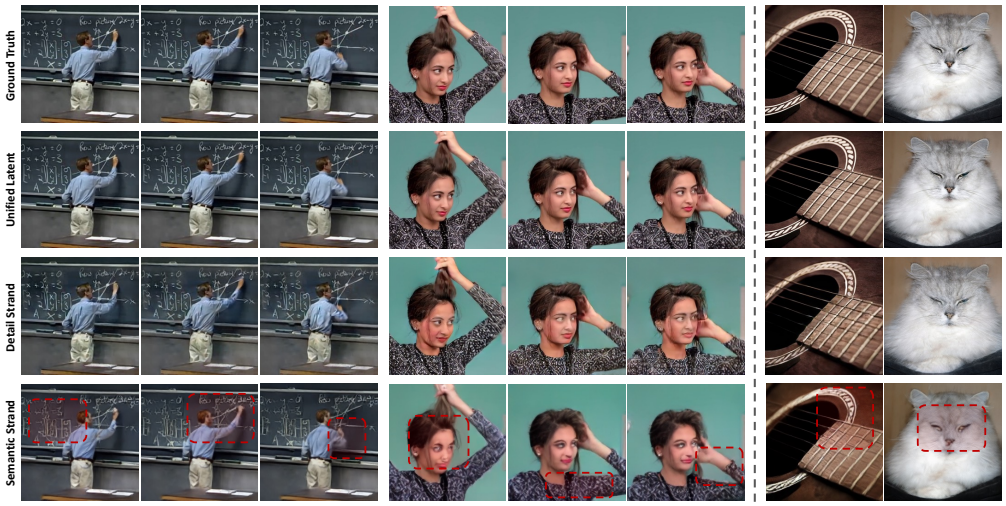


Figure 8: Qualitative strand ablation for image and video reconstruction. The left panel shows video reconstruction, and the right panel shows image reconstruction.

How can semantic structure be recovered and preserved in a compact latent? Table 5 separates the roles of semantic recovery and preservation. Removing $\mathcal{L}_{\text{align}}$ reduces accessibility across h_s, z_s, and z_u, while removing both objectives causes the largest degradation, confirming the importance of explicit semantic supervision. Without $\mathcal{L}_{\text{keep}}$, however, h_s remains strongly semantic, but accessibility drops sharply in z_s and z_u. Figure 9 shows the same effect geometrically, with semantic structure becoming more entangled after compression. Thus, $\mathcal{L}_{\text{align}}$ promotes semantic recovery, whereas $\mathcal{L}_{\text{keep}}$ preserves it through the compression.

$\mathcal{L}_{\text{align}}$	$\mathcal{L}_{\text{keep}}$	h_s	z_s	z_u
Image (ImageNet-1K)				
✓	✓	**78.90**	**65.68**	**64.97**
–	✓	70.94	57.18	56.89
✓	–	77.89	15.20	19.21
–	–	14.89	4.87	5.95
Video (UCF-101)				
✓	✓	**86.24**	**72.69**	**72.41**
–	✓	66.10	43.83	42.56
✓	–	85.39	26.24	28.35
–	–	32.56	22.37	23.10

Table 5: Semantic recovery and preservation

Representation versus strand-aware modeling. Figure 10 isolates the two sources of generative improvement. Under the same baseline generator, replacing the native Wan2.2 latent with WeaveAE improves both image and video convergence. Holding WeaveAE fixed and replacing the baseline generator with WeaveDiT yields a further gain. Latent reorganization and strand-aware modeling therefore make distinct, complementary contributions.

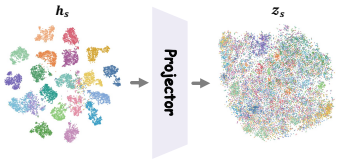


Figure 9: Semantic collapse from h_s to z_s without $\mathcal{L}_{\text{keep}}$.

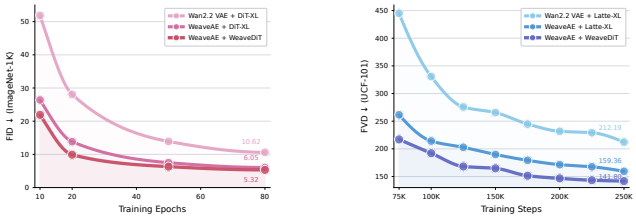


Figure 10: Effects of the latent representation and strand-aware modeling on generative learning. Left: ImageNet-1K; right: UCF-101.

5 Conclusion

In this work, we rethink the common characterization of reconstruction-oriented VAE latents as weak representations. Our probing study shows that weak linear semantic accessibility does not imply semantic absence: substantial semantic structure remains recoverable from frozen VAE latents. Building on this observation, LatentWeave recovers this semantic structure and organizes it together with reconstruction-oriented information into a unified visual representation. Our experiments further show that the recovered semantics can be effectively incorporated into the unified representation while preserving reconstruction quality and enabling efficient image and video generation. More broadly, our findings suggest that the quality of a visual latent depends not only on the information it contains, but also on how that information is organized within the representation and how readily the representation can be exploited by downstream models. Semantic recovery and structured organization therefore offer a new direction for improving generative representations.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript preparation, including language polishing, translation, refinement of wording and sentence structure, LaTeX formatting and paper layout, and suggestions for improving the presentation and organization of figures and diagrams. We also used generative AI tools to assist with coding and implementation, including understanding and navigating the existing codebase, drafting and refining training scripts, and assisting with portions of code implementation. All AI-assisted text, code, and edits were manually reviewed by the authors, and AI-assisted code was tested and validated against the intended implementation before being used in experiments. We take full responsibility for the final content of this work, including all text, code, claims, results, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All authors have read and agree to adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We provide detailed implementation and evaluation information to facilitate the reproduction of our results. The main paper describes the overall framework, methodology, and experimental setup, while the Appendix provides comprehensive details on model architectures, training configurations, optimization hyperparameters, inference settings, and evaluation protocols. Additional experimental results and analyses are also provided in the Appendix.

REFERENCES

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 6816–6826. IEEE, 2021.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=QaCCuDfBk2. Featured Certification.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50.

Black Forest Labs. FLUX.2: Analyzing and Enhancing the Latent Space of FLUX – Representation Comparison. https://bfl.ai/research/representation-comparison, 2025.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=INqBOmwIpG.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about Kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

Yu Cheng and Fajie Yuan. LeanVAE: An ultra-efficient reconstruction VAE for video diffusion models. In *IEEE/CVF International Conference on Computer Vision*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Weichen Fan, Haiwen Diao, Quan Wang, Dahua Lin, and Ziwei Liu. The prism hypothesis: Harmonizing semantic and pixel representations via unified autoencoding. *arXiv preprint arXiv:2512.19693*, 2025.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30467–30476, 2026.

Kaixiong Gong, Xin Cai, Bin Lin, Hao Wang, Yunlong Lin, Mingzhe Zheng, Bohao Li, Jian-Wei Zhang, Miles Yang, Zhao Zhong, et al. Twins: Learn to predict unified representations with focal loss. In *Forty-third International Conference on Machine Learning*, 2026.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Minghui Guo, Shengqiong Wu, and Hao Fei. V-rae: Rethinking video latent spaces for generation. *arXiv preprint arXiv:2608.13556*, 2026.

Philippe Hansen-Estruch, David Yan, Ching-Yao Chuang, Orr Zohar, Jialiang Wang, Tingbo Hou, Tao Xu, Sriram Vishwanath, Peter Vajda, and Xinlei Chen. Learnings from scaling visual tokenizers for reconstruction and generation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=MumOAOs9HY.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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.

Black Forest Labs. FLUX. Technical report, 2024. URL https://blackforestlabs.ai/announcing-black-forest-labs/.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Shizhan Liu, Xinran Deng, Zhuoyi Yang, Jiayan Teng, Xiaotao Gu, and Jie Tang. Delving into latent spectral biasing of video VAEs for superior diffusability. *arXiv preprint arXiv:2512.05394*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

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

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research*, 2025.

John Page, Xuesong Niu, Kai Wu, and Kun Gai. Boosting latent diffusion models via disentangled representation alignment, 2026. URL https://arxiv.org/abs/2601.05823.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. In *International Conference on Learning Representations*, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

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

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

Lucas Smaira, Joao Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the Kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*, 2020.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Zhentao Tan, Ben Xue, Jian Jia, Junhao Wang, Wencai Ye, Shaoyun Shi, Mingjie Sun, Wenjin Wu, Quan Chen, and Peng Jiang. SweetTok: Semantic-aware spatial-temporal tokenizer for compact video discretization. In *IEEE/CVF International Conference on Computer Vision*, pp. 23541–23550, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1365–1374, 2019.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

Team Wan, 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, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 40633–40642, 2026.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, jian Yang, Ming-Ming Cheng, and Xiang Li. Representation entanglement for generation: Training diffusion transformers is much easier than you think. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=koEALFNBj1.

Zhihao Xie, Junfeng Wu, Xinting Hu, Junchao Huang, and Li Jiang. Videorae: Taming video foundation models for generative modeling via representation autoencoders. *arXiv preprint arXiv:2607.14088*, 2026.

Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2364–2373. IEEE, 2018.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Gupta, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

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

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

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.11690.

A LatentWeave Implementation Details

We provide additional implementation details of the final LatentWeave configuration used in our experiments. We first summarize the concrete architectural instantiations of WeaveAE and WeaveDiT, and then describe their training procedures and optimization settings.

WeaveAE implementation. WeaveAE is built on top of a frozen pretrained Wan2.2 causal 3D VAE encoder (Wan et al., 2025). Given an image or video x, the encoder produces

$$z_{\text{vae}} = E_{\text{vae}}(x), \qquad z_{\text{vae}} \in \mathbb{R}^{B\times 48\times T'\times 16\times 16}, \tag{9}$$

where $T' = 1$ for images and $T' = 5$ for 17-frame videos at 256×256 resolution. After flattening the spatiotemporal grid, $N = 256$ and $N = 1280$ tokens are obtained for images and videos, respectively. Specifically, the Semantic Mapper M_{sem} is implemented as a 6-layer Transformer operating on these flattened VAE tokens and produces 1152-dimensional semantic features while preserving the same spatiotemporal token grid. The complete WeaveAE pipeline is

$$x \xrightarrow{E_{\text{vae}}} z_{\text{vae}} \xrightarrow{\text{detail/semantic}} (z_d, z_s) \xrightarrow{\oplus} z_u \xrightarrow{D_{\text{stvit}}} \hat{x}. \tag{10}$$

WeaveDiT implementation. WeaveDiT operates on the 96-channel unified latent $z_t = z_{d,t} \oplus z_{s,t}$, with 48 channels for each strand. Both image and video variants first process the two aligned strands with separate but interacting streams, fuse them position-wise, and then perform joint modeling. For videos, we use 2 dual-stream blocks followed by 24 joint blocks, with AdaLN conditioning and separate detail and semantic velocity heads. For images, we use 4 dual-stream blocks followed by 24 joint blocks, introduce timestep and class conditions as in-context tokens, and employ a 2-block DDT prediction head with hidden dimension 2048 and a shared 96-channel output head.

A.1 Training Details

WeaveAE training. We train WeaveAE in two stages. Stage I jointly learns the unified visual representation and reconstruction, beginning with an image warm-up on ImageNet-1K (Russakovsky et al., 2015) and continuing with joint image–video training on ImageNet-1K and Kinetics-600 (Carreira et al., 2018). The pretrained Wan2.2 VAE encoder (Wan et al., 2025) remains frozen throughout, while the Semantic Mapper, semantic projector, and ST-ViT decoder are optimized. Stage II freezes the complete encoder-side representation and optimizes only the decoder together with modality-specific discriminators. The main optimization settings are summarized in Table 6.

Stage I. The image warm-up jointly optimizes reconstruction and semantic organization. We use a frozen SigLIP2 encoder (Tschannen et al., 2025) as the semantic teacher. The reconstruction loss is

$$\mathcal{L}^{\text{I}}_{\text{rec}} = \mathcal{L}_1 + \mathcal{L}_{\text{LPIPS}}, \tag{11}$$

where $\mathcal{L}_{\text{LPIPS}}$ denotes the perceptual reconstruction loss (Zhang et al., 2018). Together with semantic alignment and preservation, the Stage-I objective is

$$\mathcal{L}_{\text{Stage I}} = \lambda_{\text{rec}}\mathcal{L}^{\text{I}}_{\text{rec}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{keep}}\mathcal{L}_{\text{keep}}. \tag{12}$$

During image warm-up, the two semantic objectives plus reconstruction are applied on ImageNet-1K. We then continue training with 30% image and 70% video batches. This allows the compact semantic strand learned during image training to be preserved while extending the representation to temporally coherent video inputs. We further perturb the unified latent during decoder training with Gaussian noise whose standard deviation is sampled from $\sigma \sim \mathcal{U}(0, \tau)$. We use $\tau = 0.8$ during image warm-up and reduce it to 0.3 during joint training. This latent perturbation is applied only during training and does not modify the learned representation or introduce additional inference-time computation. The complete optimization configurations for different training stages are summarized in Table 6.

Stage II. Starting from the jointly trained Stage-I checkpoint, we freeze the complete encoder-side representation, including the Semantic Mapper and semantic projector. The resulting detail strand z_d, semantic strand z_s, and unified latent z_u therefore remain fixed, while only the ST-ViT decoder and

Table 6: WeaveAE optimization and loss configuration.

Setting	Stage I (Warm-up)	Stage I (Joint)	Stage II
Optimization			
Optimizer	AdamW	AdamW	AdamW
Learning rate	2×10^{-4}	1×10^{-4}	5×10^{-5}
LR schedule	Cosine	Cosine	Cosine
LR Warm-up	2K updates	1K updates	500 updates
Adam (β_1, β_2)	$(0.9, 0.95)$	$(0.9, 0.95)$	$(0.9, 0.95)$
Batch size	256	256 (image) / 8 (video)	256 (image) / 8 (video)
Training steps	200K	120K	80K
Latent noise τ	0.8	0.3	0.3
Objective-specific settings			
λ_{rec}	1.0	1.0	1.0
λ_{LPIPS}	1.0	1.0	1.0
λ_{align}	1.0	0.5	–
λ_{keep}	0.5	0.5	–
Adversarial loss start	–	–	15K (image) / 6K (video)
λ_{adv}	–	–	0.75 (image) / 0.5 (video)

modality-specific discriminators are optimized. We continue training on the same ImageNet-1K and Kinetics-600 mixture used during joint training.

The decoder objective augments pixel-wise and perceptual reconstruction with adversarial supervision:

$$\mathcal{L}_{\text{rec}}^{\text{II}} = \mathcal{L}_1 + \mathcal{L}_{\text{LPIPS}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}, \tag{13}$$

where adversarial training is introduced only at this stage (Goodfellow et al., 2014). For images, we use a DINO-based feature discriminator (Zheng et al., 2026), while videos use a 2D–3D PatchGAN to jointly constrain spatial appearance and local temporal consistency (Ma et al., 2025). We further apply gradient-based adaptive weighting to balance the reconstruction and adversarial terms.

To reduce decoder reliance on the detail strand, we introduce two complementary objectives during Stage II. In the *masked-detail* path, spatial regions of z_d are masked while the remaining detail tokens and the complete z_s are used for full reconstruction, improving local completion when detail evidence is missing. In the *semantic-only* path, z_d is removed entirely and only low-frequency reconstruction is supervised from z_s, ensuring that global content and spatial structure remain decodable without requiring high-frequency appearance. Thus, the two paths target local robustness and coarse structural sufficiency, respectively. Both augment standard full-latent reconstruction, update only the decoder, and leave the latent representation and inference architecture unchanged. We empirically explored several optimization settings and report the final configuration in Table 6.

WeaveDiT training. We train the image and video WeaveDiT variants with flow matching (Liu et al., 2022; Ma et al., 2024) on the frozen unified representation produced by WeaveAE. Both models predict the velocity of the complete 96-channel latent, while using modality-specific optimization schedules to accommodate their different sequence lengths and conditioning mechanisms. For image generation, we train on ImageNet-1K (Russakovsky et al., 2015); for video generation, we train on UCF-101 (Soomro et al., 2012). The corresponding optimization settings are summarized in Table 7.

B Additional Experimental Details

We provide additional experimental details and supplementary results omitted from the main paper for space. We first describe the reconstruction and generation evaluation protocols, with additional details for video experiments, and then report supplementary image reconstruction and generation results that complement the video-focused evaluation in the main text.

B.1 Evaluation Protocols

Reconstruction. All reconstruction experiments are evaluated at 256×256 resolution. For images, WeaveAE is trained only on ImageNet-1K (Russakovsky et al., 2015) and evaluated on both ImageNet-

Table 7: Optimization configuration for WeaveDiT.

Setting	Image	Video
Optimization		
Optimizer	AdamW	AdamW
Learning rate	2×10^{-4}	1×10^{-4}
Adam (β_1, β_2)	(0.9, 0.95)	(0.9, 0.999)
LR schedule	Linear decay	Constant
LR warm-up	40 epochs	–
Precision	BF16	BF16
Trainable parameters	685M	685M
Batch size	1024	64
Training budget	80 epochs	250K steps
Flow matching and conditioning		
Timestep distribution	Logit-normal(0, 1)	Logit-normal(0, 1)
Conditioning	In-context (time/class)	AdaLN (time/class)
Class dropout	0.1	0.1
Prediction target	Velocity	Velocity
Self-REPA		
Alignment weight	0.1	0.1
Attachment depth	6	6
Target dimension	1152	1152

1K and MS-COCO 2017 (Lin et al., 2014); no additional training or adaptation is performed on MS-COCO. We report PSNR and SSIM for distortion-level fidelity, together with rFID computed between reconstructed and real images. For videos, we evaluate 17-frame clips from Kinetics-600 (Carreira et al., 2018) and UCF-101 (Soomro et al., 2012) at 256×256 resolution, reporting PSNR, SSIM, and rFVD. The same preprocessing and metric implementations are used across compared tokenizers whenever applicable.

Generation. For image generation, all models are trained on ImageNet-1K and evaluated using 50K class-conditional samples at 256×256 resolution. We report gFID, Inception Score (IS), precision, and recall, and additionally report classifier-free guidance results where applicable. For video generation, we evaluate class-conditional generation on UCF-101 and unconditional generation on SkyTimelapse (Xiong et al., 2018). Following prior video-generation protocols (Ma et al., 2025), we generate 2,048 clips of 17 frames at 256×256 resolution for each benchmark and report FVD, KVD, and IS. For UCF-101, clips are sampled with a temporal stride of 3. Across controlled comparisons, we keep the sampling procedure and metric implementations fixed.

B.2 Image Reconstruction

As shown in Table 8, WeaveAE achieves the best rFID on ImageNet-1K (**0.15**) and the second-best result on MS-COCO (**1.72**). It also outperforms all compared video tokenizers on both benchmarks, demonstrating that the unified latent remains effective for image reconstruction. Meanwhile, its PSNR and SSIM remain close to those of the pretrained Wan2.2 VAE, indicating that integrating the semantic strand preserves distortion-level reconstruction quality while substantially improving reconstruction distribution fidelity.

B.3 Image Generation

As shown in Table 9, LatentWeave substantially improves generative learning over the native VAE baseline, suggesting that the unified latent with explicitly organized semantic structure is more amenable to generative modeling. After only 20 training epochs, LatentWeave achieves a gFID of **9.57**, already outperforming the 80-epoch results obtained with Wan2.2 VAE (10.63) and SSVAE (10.37). At 80 epochs, it further improves to a gFID of **5.26** and an IS of **161.6**, with precision and recall of **0.73** and **0.65**, respectively. With classifier-free guidance, LatentWeave obtains a gFID of **2.72** and an IS of **258.11**. Beyond final generation quality, Figure 1 further demonstrates substantially faster convergence, with LatentWeave requiring approximately **12× fewer training epochs** than the native Wan2.2 VAE baseline to reach comparable generation quality.

Table 8: Image reconstruction on ImageNet-1K and MS-COCO 2017.

Method	Dim.	ImageNet-1K			MS-COCO 2017		
		PSNR↑	SSIM↑	rFID↓	PSNR↑	SSIM↑	rFID↓
Image tokenizers							
SD-VAE (Rombach et al., 2022)	4	27.37	0.78	0.67	27.08	0.80	3.93
FLUX-VAE (Labs, 2024)	16	**32.74**	**0.92**	0.18	**32.32**	**0.93**	**1.35**
VA-VAE (Yao et al., 2025)	32	27.96	0.79	0.28	27.50	0.81	2.71
SVG (Shi et al., 2026)	768	24.25	0.67	0.78	–	–	–
RAE (Zheng et al., 2026)	1152	22.19	0.58	0.43	18.36	0.47	6.01
Video tokenizers							
CogVideoX VAE (Yang et al., 2024)	16	30.58	0.87	1.68	30.45	0.87	5.23
Wan2.1 VAE (Wan et al., 2025)	16	29.57	0.85	0.83	29.43	0.87	3.44
HunyuanVideo VAE (Kong et al., 2024)	16	30.44	0.89	0.62	30.36	0.88	2.67
LeanVAE (Cheng & Yuan, 2025)	16	30.42	0.88	4.51	29.22	0.84	10.54
Wan2.2 VAE (Wan et al., 2025)	48	29.56	0.84	0.76	29.43	0.86	3.43
SSVAE (Liu et al., 2025)	48	28.23	0.81	1.84	28.05	0.82	6.00
WeaveAE	96	29.15	0.84	**0.15**	29.03	0.85	1.72

Table 9: Class-conditional image generation on ImageNet-1K at 256 × 256 resolution.

Method	Epochs	#Params	Generation@256 w/o guidance				Generation@256 w/ guidance			
			gFID↓	IS↑	Prec.↑	Rec.↑	gFID↓	IS↑	Prec.↑	Rec.↑
Image-based generative systems										
SD-VAE (Rombach et al., 2022)	1400	675M	9.62	121.5	0.67	0.67	2.27	278.2	0.83	0.57
VA-VAE (Yao et al., 2025)	80	675M	5.14	130.2	0.76	0.62	2.11	252.3	**0.81**	0.58
REPA (Yu et al., 2025)	80	675M	7.90	122.6	0.70	0.65	–	–	–	–
DDT (Wang et al., 2026)	80	675M	6.62	135.2	0.69	**0.67**	1.52	263.7	0.78	0.63
SVG (Shi et al., 2026)	80	675M	6.57	137.9	–	–	3.54	207.6	–	–
RAE (Zheng et al., 2026)	80	839M	**2.16**	**214.8**	**0.82**	0.59	–	–	–	–
Video-tokenizer-based generative systems										
SSVAE (Liu et al., 2025)	80	675M	10.37	113.8	0.68	0.63	–	–	–	–
Wan2.2 VAE (Wan et al., 2025)	80	675M	10.63	111.7	0.68	0.64	–	–	–	–
LatentWeave	20	685M	9.57	119.5	0.69	0.61	–	–	–	–
	80		**5.26**	**161.6**	**0.73**	**0.65**	**2.72**	**258.11**	**0.81**	**0.57**

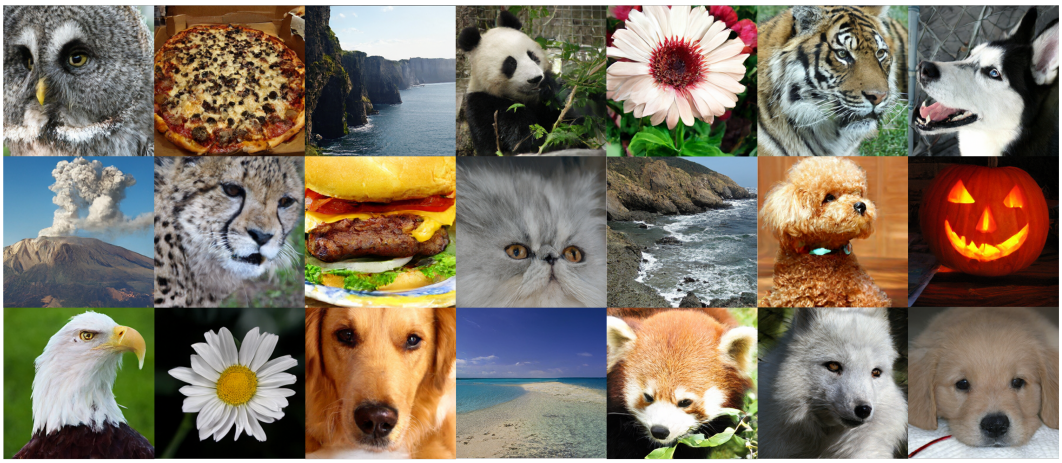


Figure 11: Qualitative class-conditional image generation on ImageNet-1K at 256 × 256 resolution.

C Additional Qualitative Results

Figures 12 and 13 provide additional class-conditional video generation results on UCF-101 across diverse action classes and motion patterns. Figure 14 further presents unconditional generation results on SkyTimelapse under different scenes and sky conditions. These examples complement the qualitative results in the main paper and provide broader coverage of LatentWeave across class-conditional and unconditional video generation settings.

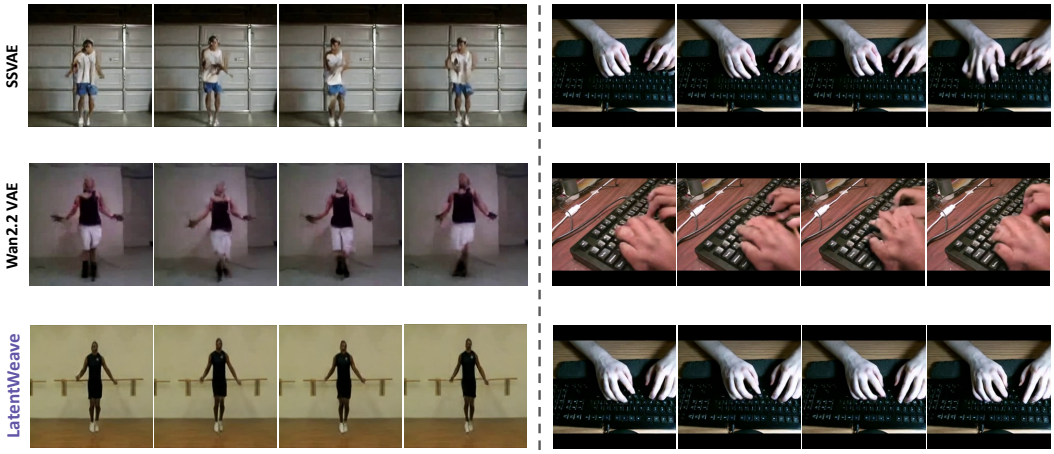


Figure 12: Additional class-conditional video generation results on UCF-101.

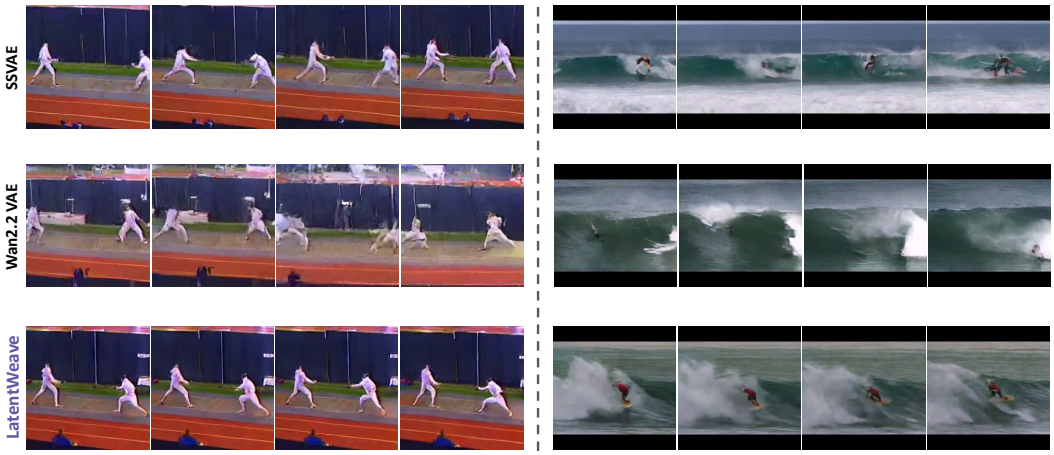


Figure 13: Additional class-conditional video generation results on UCF-101.

D Limitations

LatentWeave shows that improving latent organization can enhance semantic accessibility and generative learning while preserving reconstruction fidelity, suggesting a complementary direction to simply scaling model capacity. Our current experiments focus on moderate-scale DiT backbones and controlled image and video settings; due to computational constraints, we have not yet explored large-scale open-domain T2V or TI2V generation, nor systematically studied the scaling behavior of WeaveDiT with larger models, datasets, resolutions, and longer sequences. Beyond scaling, the resulting representation may also provide a useful basis for broader video intelligence systems. For action-conditioned world or world-action modeling, it would be valuable to investigate whether the complementary semantic and detail strands can jointly support high-level state reasoning, fine-grained visual dynamics, and future prediction within a shared latent space. Similarly, LatentWeave may offer a representation-level bridge for unified video models that combine understanding, generation, prediction, and editing, where semantic accessibility and reconstruction fidelity are both desirable.

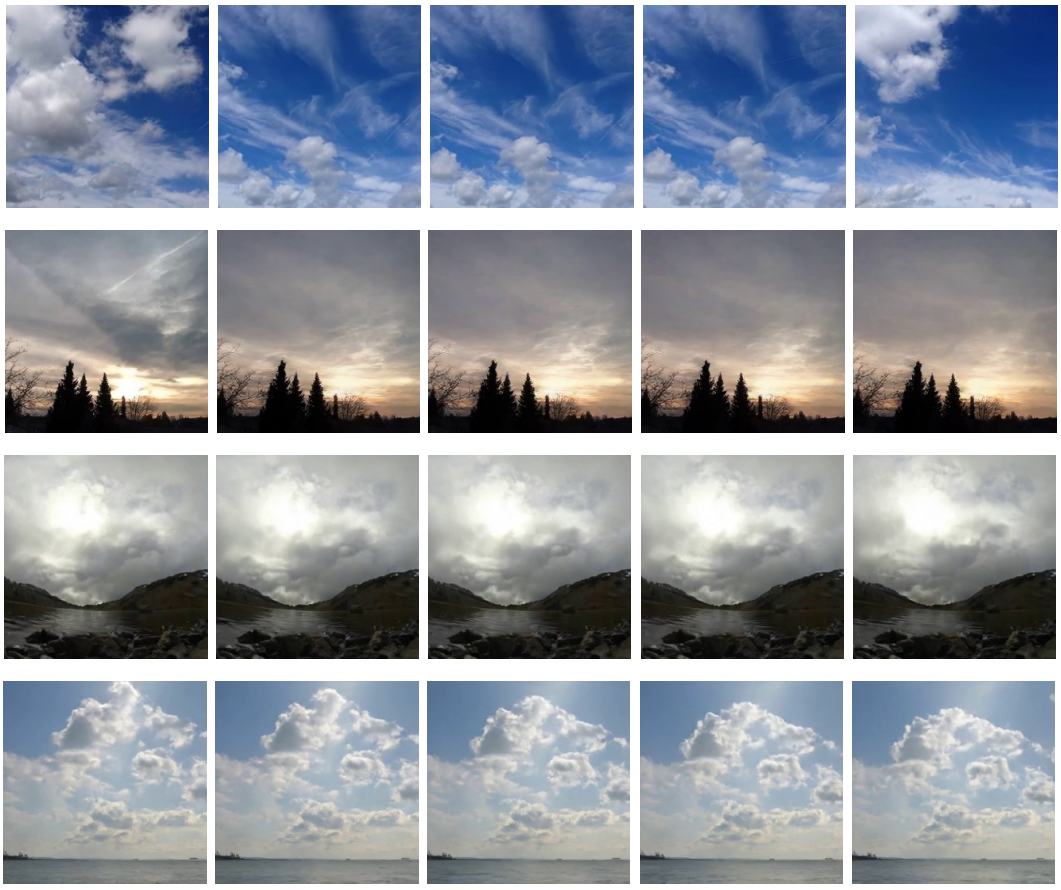


Figure 14: Additional unconditional video generation results on SkyTimelapse.