GENERATIVE UNCERTAINTY AS A SELF-SUPERVISED SIGNAL FOR SEMANTIC SIMILARITY LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating semantic similarity between videos is a fundamental challenge in computer vision, essential for tasks ranging from out-of-distribution (OOD) detection to video retrieval. However, defining and labeling video similarity is notoriously difficult and expensive due to the complex spatio-temporal nature. In this paper, we propose a novel self-supervised approach that leverages generative uncertainty from text-to-video (T2V) diffusion models to learn semantic similarity without human annotations. Our method is based on the observation that T2V models produce consistent outputs for familiar concepts but exhibit high variance and uncertainty when prompted with specialized concepts. We utilize this behavior to identify stable semantic features within existing pretrained representations, such as VideoMAE and V-JEPA. Specifically, we learn a mask over these embeddings using purely generated data, encouraging the model to retain features that remain consistent across generations of general concepts while discarding those associated with generative noise or uncertainty. Experimental results across three key tasks demonstrate that our learned feature subspaces consistently outperform original pretrained features and baseline feature selection methods.

1 INTRODUCTION

Learning semantic similarity between videos is important for many tasks, such as out-of-distribution (OOD) detection (Sim et al., 2023), video retrieval (Dwibedi et al., 2019), and evaluating generative models (Liu et al., 2024). Yet, learning such similarity from human annotations is challenging: videos contain rich spatio-temporal information, and judging which aspects are semantically relevant is inherently difficult (Zhang et al., 2018). Collecting similarity annotations at scale is therefore expensive and often infeasible, particularly for open-domain or generated videos. This motivates the question: *how can we learn video representations that capture semantic similarity without using labeled data?*

Self-supervised video models such as VideoMAEv2 (Wang et al., 2023) and VJEPA 2.1 (Mur-Labadia et al., 2026) provide strong general-purpose representations across many downstream tasks. However, their training objectives are not designed explicitly for measuring semantic similarity, and their feature spaces can contain dimensions that are irrelevant or even detrimental to similarity-based tasks. As we show experimentally, directly comparing these representations can lead to limited alignment with human judgments and unreliable uncertainty estimates for generated videos.

We investigate whether pretrained video representations can instead be adapted for semantic similarity using only synthetic data from text-to-video (T2V) models. Our key observation is that the diversity of a generative model's outputs provides a natural self-supervised signal. When a model is given a prompt describing a general concept, it generates videos that tend to be consistent with each other. In contrast, when the prompt describes a specialized concept to the model, the generated videos are more diverse and less consistent (Fig. 2). We refer to this diversity as *generative uncertainty* and use it to identify which feature dimensions are most useful for measuring semantic similarity.

To achieve this, we rely entirely on synthetic data generated by a T2V model. We generate videos from two types of prompts: (i) *general* concepts that the model can generate reliably, and (ii) *specialized* concepts (i.e., concepts belonging to highly specific or niche domains) that lead to more uncertain, and therefore diverse outputs. Particularly, our method does not require generative variability to be a perfect measure of whether a concept was seen during training. Our hypothesis is weaker: if

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.

some representation dimensions remain stable when repeated generations preserve the same semantic content, but become variable when the generated content becomes less consistent, then this difference in variability can reveal dimensions that are useful for semantic similarity. For each prompt, we generate multiple videos, varying only the initial noise, and learn a mask over pretrained video embeddings, encouraging embeddings of videos generated from the same general prompt to be more similar than videos generated from the same specialized prompt. Importantly, all pairs are therefore obtained from videos generated with the *same* prompt, avoiding human similarity annotations or predefined semantic relations between prompts.

We find that the learned feature subspaces consistently outperform the original features and other subspace methods using different T2V models and across three similarity-based tasks: human alignment, OOD detection for generated samples, and video retrieval. At the same time, the learned masks retain only a fraction of the original feature dimensions, yielding more compact representations. We further evaluate patch-level masking for video object segmentation and observe minimal performance loss despite substantial dimensionality reduction (Appendix C.5).

Overall, our main contributions are as follows:

- We identify generative uncertainty in T2V models as a self-supervised signal for semantic similarity: general concepts produce relatively consistent generations, whereas specialized concepts yield more diverse outputs.
- Using this signal, we introduce a lightweight, model-agnostic method that learns sparse subspaces of pretrained video representations and improves human-similarity alignment, OOD detection, and video retrieval across different feature encoders.

2 Related work

Video similarity. Defining and quantifying perceptual similarity is a fundamental challenge in computer vision. For static images, deep feature distances have proven very effective (Zhang et al., 2018) due to their close alignment with human perception. However, extending this paradigm to videos is inherently more complex, as recently shown by ConViS-Bench (Liberatori et al., 2025). Their benchmark shows that recent powerful models for video understanding, such as V-JEPA 2.1 (Mur-Labadia et al., 2026) and VideoMAE V2 (Wang et al., 2023), extract representations that align only partially with human judgment. In this work, we build upon these backbones, demonstrating that their raw embeddings can be refined into concentrated subspaces via our targeted masking approach to improve performance in video-similarity-based tasks.

Video similarity learning itself is most commonly explored through the proxy tasks of video retrieval and near-duplicate detection. This is done, for instance, by temporal context aggregation via contrastive learning (Shao et al., 2021), temporal alignment for partial video copy detection (Han et al., 2021), and leveraging region attention graphs (Ng et al., 2022). Other works propose unified frameworks for event retrieval and recounting (Gao et al., 2017), suppress irrelevant frames to improve video-to-video retrieval (Jo et al., 2024), or focus explicitly on near-duplicate video retrieval (He et al., 2022). Other works have explored fully self-supervised methods to learn video similarity directly from visual streams (Kordopatis-Zilos et al., 2023). He et al. (2022) learns near-duplicate retrieval from massive pools of unlabeled videos, while Kordopatis-Zilos et al. (2023) relies on synthetic proxy tasks created via heavy data augmentations to learn self-supervised similarity. However, while these methods are effective for retrieval, they often rely on access to large video datasets.

One promising avenue, therefore, is to turn to synthetic data. This is largely unexplored in video similarity, although it has seen significant success in other vision domains, such as perceptual image tasks (Wu et al., 2023) or action recognition (Varol et al., 2021). In our work, we use synthetic data to bypass the need for real-world video collections or manual augmentations by using the zero-shot outputs of modern text-to-video models to learn video similarity.

Generative uncertainty. Previous works on generative uncertainty largely focus on developing methods to measure the uncertainty of samples produced by the model. For image generation, this is typically done by adapting the original model into a sort of ensemble by introducing randomness and measuring the distance between generations. For instance, Berry et al. (2024) obtains multiple fine-tuned versions of the label embedding layer and uses the distances between predicted Gaussians

to measure uncertainty, while Jazbec et al. (2025) draws multiple parameter sets from a Laplace approximation of the final layer and uses the CLIP embedding space to measure distances. Other works make use of the inherent randomness in the diffusion denoising process to obtain uncertainty estimates (De Vita & Belagiannis, 2025). In scenarios where text is available, such as a prompt for text-to-image methods, vision-language models can be used to assess similarity to the prompt and measure uncertainty (Franchi et al., 2025). In this work, we do not propose a new method for measuring generative uncertainty. Rather, to our knowledge, we are the first to exploit this inherent characteristic of generative models to define a self-supervised objective to learn a notion of similarity.

3 METHOD

Our method, summarized in Figure 1, is driven solely by synthetic data, comprising both the generated videos and their source prompts. Human involvement is limited strictly to the initial domain assumptions and prompt curation, which serves as the foundation for our dataset.

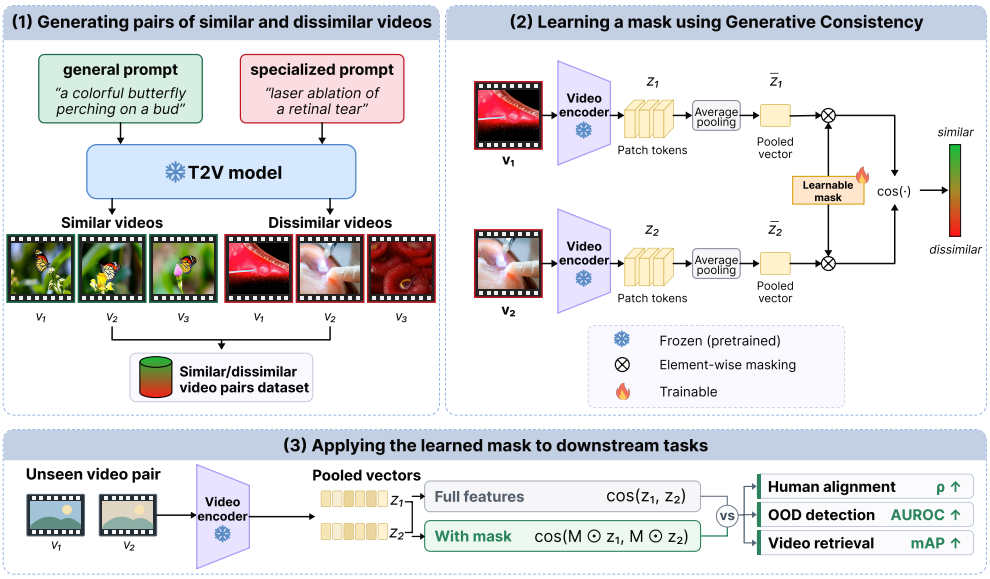


Figure 1: **Proposed framework overview. (1)** A text-to-video model generates videos from *general* and *specialized* prompts with different random seeds; videos from the same prompt are paired. **(2)** A frozen video encoder extracts global representations, and a learnable mask is trained to yield high/low cosine similarity for *similar/dissimilar* pairs. **(3)** For downstream tasks, we mask unseen video representations before cosine similarity. In doing so, our strategy improves alignment with human judgment, out-of-distribution detection, and video retrieval.

3.1 SYNTHETIC DATASET GENERATION

Defining general concepts. We construct our synthetic dataset using the Wan2.1 (T2V-1.3B) text-to-video model (Wan et al., 2025). A key challenge lies in partitioning the prompt space into *general* and *specialized* concepts relative to the model's underlying training distribution. While the exact training corpus of Wan2.1 is proprietary, it is documented to include broad domains such as natural landscapes, daily human activities, and general scenery. Furthermore, Wan2.1 demonstrates state-of-the-art performance on the VBench benchmark (Huang et al., 2024), which evaluates video generation across a wide array of common semantic categories. Consequently, we define our *general* set by sampling $N = 100$ prompts directly from VBench, representing concepts the model can synthesize with high fidelity and consistency.

Constructing specialized concepts. Conversely, we define *specialized* prompts as those belonging to highly uncommon or niche domains, such as medical surgery or microscopy, for which current T2V models tend to produce more variable generations in our experiments (Fig.3). To construct this set,

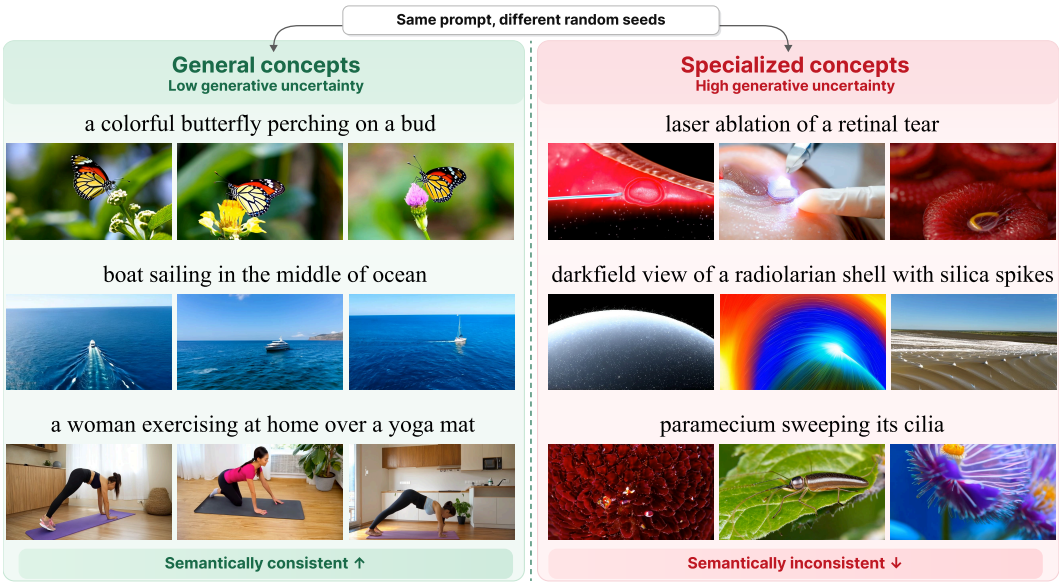


Figure 2: **Example videos used in our generative consistency objective.** Videos generated by Wan 2.1 with *general* and *specialized* text prompts. Videos from *general* prompts serve as proxies for semantic consistency, while those from *specialized* prompts represent generative divergence.

we leveraged Google's Gemini 3.1 Pro to generate $N = 100$ unique prompts from these specialized domains. To ensure that our downstream evaluation captures semantic domain shifts rather than superficial linguistic variances, we constrained the language model to match the *general* prompts in terms of length, syntactic complexity, and descriptive granularity. More details in Appendix B.1.

Generating the synthetic video dataset. For each of the 200 total prompts, we synthesized three distinct videos by varying the initial random seed. This process yielded a dataset of **600** videos. In our framework, videos generated from *general* prompts serve as proxies for semantic consistency (high intra-prompt similarity), while those from *specialized* prompts provide a signal for generative divergence, allowing us to learn a similarity subspace without human supervision. Examples of these synthesized general and specialized concepts are presented in Figure 2 while Figure 3 quantitatively shows how general prompts lead to closer embeddings in the chosen feature space. In Appendix C.1 we present similar findings for other generative models: CogVideoX1.5-5B and Mochi 1.

3.2 Feature Learning via Generative Consistency

Extracting global video features. For a given video v, we utilize a pretrained ViT-based encoder $\mathcal{E}$ to extract spatio-temporal features. While these architectures produce a sequence of patch tokens of shape $[P, D]$, where P denotes the number of tokens and D the embedding dimension, our objective is to capture global semantic similarity rather than localized cues. Thus, we aggregate tokens into a single global representation $\mathbf{z} = \mathcal{E}(v) \in \mathbb{R}^D$ via global average pooling across the token dimension.

Learning a soft feature mask. To identify a subspace that prioritizes stable semantic content, we optimize a learnable soft mask $M \in [0, 1]^D$, which acts as a differentiable feature weighting mechanism. During the training phase, this mask remains continuous to facilitate gradient-based optimization. We utilize the synthetic dataset described in the previous section as a supervisory proxy, optimizing M via the AdamW (Loshchilov & Hutter, 2019) optimizer to minimize the Mean Squared Error (MSE) between the masked cosine similarity of a video pair and a generative stability label y. Specifically, we set $y = 1$ for pairs synthesized from the same *general* prompt (characterizing high semantic consistency) and $y = 0$ for those from the same *specialized* prompt (characterizing generative divergence). The training objective is then defined as:

$$s_M(\mathbf{z}_1, \mathbf{z}_2) = \frac{1 + \cos\left(\mathbf{z}_1 \odot \mathbf{M}, \mathbf{z}_2 \odot \mathbf{M}\right)}{2}, \qquad \mathcal{L} = \left(s_M(\mathbf{z}_1, \mathbf{z}_2) - y\right)^2, \quad y \in \{0, 1\}. \tag{1}$$

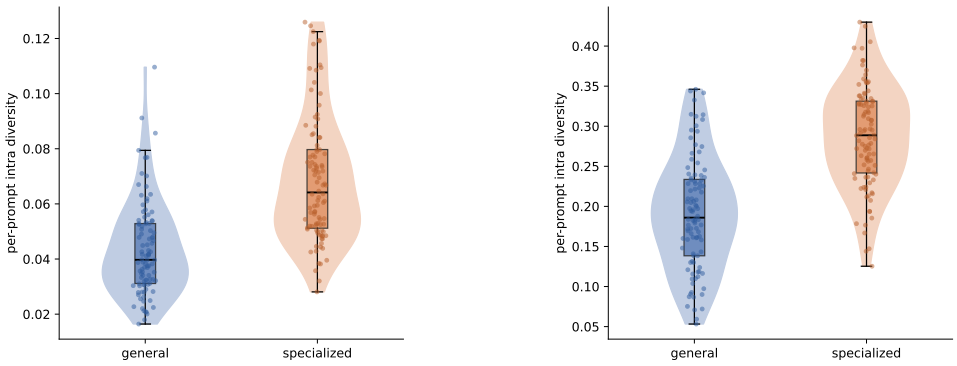


Figure 3: **Per-prompt intra-set feature-space diversity for videos generated by Wan2.1, measured using V-JEPA 2.1 Large (left) and VideoMAEv2 Base (right).** Each point represents one prompt and reports the average pairwise normalized cosine dissimilarity among three videos generated with different random seeds.

From soft weights to sparse feature selection. To prevent the mask from overfitting to the specific synthetic noise of the T2V model, we employ an early stopping mechanism based on the margin between the average similarities of the general and specialized groups. Once optimization is complete, the mask can be binarized using a threshold of 0.5 to produce a binary mask $\hat{M} = \mathbb{I}(M > 0.5) \in \{0, 1\}^D$. We evaluate both the learned soft mask and its binarized version for all downstream tasks.

4 Experiments

We verify the effectiveness of our masked features on three similarity-driven tasks using V-JEPA 2.1 and VideoMAEv2 backbones. To determine whether the uncertainty signal is tied to a specific video generator, we repeat all primary experiments with two additional text-to-video models, CogVideoX1.5-5B (Yang et al., 2024) and Mochi 1 (Team, 2024). While this section presents results produced with Wan2.1 as the generator, consistent analysis and results for the other two models are provided in Appendix C.4. We additionally provide a theoretical interpretation of our method in Appendix A.

4.1 Alignment with human perception

To validate that our learned feature masks select a subspace in which similarity aligns with human judgment, we evaluate our approach on ConViS-Bench (Liberatori et al., 2025). This benchmark proposes a dataset of video pairs meant to measure concept-based video similarity. Each video pair is annotated by humans with a similarity score across different semantic dimensions (e.g., the main action, subjects, objects, location, and the order of actions). Following Liberatori et al. (2025), we compute cosine similarity between videos in the embedding space of different encoders and measure alignment using Spearman's ρ and Kendall's τ rank correlation coefficients, evaluating how well a model's latent representations align with human judgment across fine-grained semantic dimensions.

Learned masks improve human similarity alignment. As shown in Table 1, applying our learned masks yields a consistent improvement in human alignment across the evaluated models. For instance, with V-JEPA 2.1-L, both the discrete binary mask and the continuous soft mask strictly outperform the unmasked baseline across all five conceptual categories. We observe notable increases in the ρ correlation for "Main Subjects" (+6.0 for the binary mask, +6.1 for the soft mask) and "Main Objects", with the soft mask achieving the best overall performance. A similar trend of uniform improvement across all categories is evident for VideoMAEv2-B, where binary and soft mask perform similarly. These results indicate that our training objective successfully forces the models to prioritize features that align with human judgment. These significant gains in human alignment can be achieved with high feature sparsity when using the binary mask: the V-JEPA 2.1-L and VideoMAEv2-B models achieve superior semantic alignment using only 32.2% and 42.0% of their original embedding dimensions, respectively. This demonstrates that our learning approach effectively isolates a concentrated subspace of features that aligns with human perception.

Table 1: **Alignment with human judgement on ConViS-Bench.** Spearman's ρ and Kendall's τ correlations with human judgement ($\times 100$). *Sp.* denotes the sparsity of the learned mask; soft masks are dense by construction. Our masked features better match the human notion of global video similarity. Best results are **bolded**, second-best underlined.

Encoder	Method	Sp. (%)	Main Action		Main Subjects		Main Objects		Location		Actions Order		Average	
			ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
V-JEPA 2.1-L	Baseline (no mask)	–	24.0	16.30	33.4	22.84	24.7	16.83	49.3	35.04	24.3	16.65	31.14	21.53
	Binary mask	67.8	26.9	18.25	39.4	27.14	30.3	20.57	50.5	35.74	30.1	20.75	35.44	24.49
	Soft mask	–	**27.7**	**18.78**	**39.5**	**27.24**	**30.8**	**20.89**	**51.3**	**36.41**	**30.9**	**21.23**	**36.04**	**24.91**
VideoMAEv2-B	Baseline (no mask)	–	47.0	33.02	44.6	31.46	42.4	29.96	45.8	32.26	44.7	31.37	44.90	31.61
	Binary mask	58.0	47.9	33.62	**46.3**	**32.63**	**44.1**	**30.98**	**46.9**	**33.09**	45.3	31.67	**46.10**	**32.40**
	Soft mask	–	**48.0**	**33.74**	46.1	32.47	44.0	30.97	46.7	32.92	**45.4**	**31.79**	46.04	32.38

In contrast, unsupervised feature-selection baselines generally do not provide comparable gains. Detailed comparison to various baselines and per-category results are provided in Appendix C.3.

4.2 Video generation OOD detection

Transfer to generative OOD detection. We focus on OOD detection for image-to-video generation, where a single input image is used as context to generate future frames. When this context image lies far from the distribution expected by the generative model, the generated video may become unreliable, unstable, or semantically inconsistent. Detecting such OOD inputs is therefore important for estimating when future-frame generation can be trusted. Since OOD detection for generative models often relies on measuring similarity between multiple generated outputs, this experiment tests whether a mask learned from synthetic text-to-video uncertainty can also improve similarity-based uncertainty estimation in another generative task.

Evaluation setup. We employ NVIDIA Cosmos-Predict2 (NVIDIA, 2025) as our base Image-to-Video (I2V) generative model. To establish a controlled ID baseline, we perform fine-tuning on the BDD100k dataset (Yu et al., 2020); consequently, videos from the BDD100k validation set are treated as ID samples. For OOD evaluation, we select four diverse datasets: UVEB (underwater footage) (Xie et al., 2024), MedVidBench (medical surgery) (Su et al., 2026), HAM10000 (dermatoscopic images) (Tschandl et al., 2018), and WikiArt (artistic images)[1]. While the exact training corpus of Cosmos is proprietary, its design focus on real-world physics and robotics provides high confidence that these four datasets represent significant distributional shifts. To quantify OOD scores, we adopt two procedures inspired by Franchi et al. (2025):

- **2XDM** (Berry et al., 2024; Franchi et al., 2025): We generate two distinct videos from the same context image using different initial random noise. Then, both videos are mapped to an embedding space via a video encoder, and the cosine dissimilarity between their embeddings serves as the OOD score.
- **DDPM-OOD** (Graham et al., 2023; Franchi et al., 2025): We first generate a base video v_0 conditioned on a context image using $K = 35$ diffusion steps. We then apply $k \in [1, K)$ steps of the forward diffusion process to this video and repeat the denoising steps. This procedure is performed three times with $k \in \{10, 20, 30\}$ to generate three new videos, denoted as v_1, v_2, and v_3, respectively. The final score is computed as the average of the cosine dissimilarities between v_0 and each of the three generated videos (v_1, v_2, v_3) in the embedding space.

Masked features improve OOD detection. This evaluation pipeline is conducted independently for each OOD dataset and repeated with both detection methods (2XDM, DDPM-OOD) using our learned feature mask. As shown in Table 2, applying the feature mask yields consistent improvements across all datasets, feature spaces, and detection methods. The binary masks show the highest gains, with improvements in AUROC up to 34% and FPR@95 up to 51%, while retaining a fraction of the feature dimensions. We again compare our learned mask against alternative masking strategies

[1] https://huggingface.co/datasets/huggan/wikiart

Table 2: **Video generation OOD detection across four diverse datasets.** Videos are generated using Cosmos-Predict2, with BDD100k as the in-distribution (ID) baseline. Applying binary and soft feature masks consistently improves OOD detection performance. Best results are **bolded** and second-best results are underlined.

OOD Dataset	Method	V-JEPA 2.1-L						VideoMAEv2-B					
		2XDM			DDPM-OOD			2XDM			DDPM-OOD		
		AUROC↑	AUPR↑	FPR@95↓	AUROC↑	AUPR↑	FPR@95↓	AUROC↑	AUPR↑	FPR@95↓	AUROC↑	AUPR↑	FPR@95↓
UVEB	Unmasked	78.42	73.77	57.20	76.37	70.31	63.60	71.08	68.50	77.60	75.50	73.33	**83.20**
	Binary mask	**86.77**	**85.72**	**51.20**	**89.52**	**88.04**	**42.80**	**72.31**	**69.27**	73.60	**77.47**	**75.10**	84.80
	Soft mask	85.79	84.48	52.00	89.11	87.35	**42.80**	72.21	68.90	**73.20**	77.37	74.81	**83.20**
MedVidBench	Unmasked	74.66	74.18	76.00	68.11	68.47	86.00	78.59	80.73	82.40	77.01	79.46	**79.60**
	Binary mask	**88.80**	**89.65**	**61.20**	**91.45**	**92.40**	**48.40**	**80.43**	**82.08**	**79.60**	**79.32**	**81.58**	83.20
	Soft mask	88.09	89.00	62.80	91.15	91.97	49.20	80.14	81.49	80.00	78.71	80.91	81.20
HAM10000	Unmasked	70.80	61.21	63.20	87.67	87.95	51.60	68.38	62.42	78.40	86.23	88.37	**67.60**
	Binary mask	**80.09**	**74.44**	**55.60**	**95.57**	**96.03**	**25.20**	**70.16**	**63.64**	75.20	**88.02**	**89.87**	71.60
	Soft mask	79.03	72.74	58.00	95.40	95.87	26.00	69.95	63.35	**74.80**	87.96	89.72	68.80
WikiArt	Unmasked	63.10	55.82	65.20	69.80	65.03	82.40	57.78	51.05	82.00	63.88	63.16	89.20
	Binary mask	**71.25**	**66.30**	**64.80**	**82.06**	**78.11**	**66.00**	**58.89**	**51.52**	79.20	**65.95**	**64.37**	90.00
	Soft mask	69.88	64.56	64.80	81.53	77.27	68.00	58.60	51.28	**77.60**	65.55	63.59	**88.40**

Table 3: Class-level **video retrieval mAP (%)** on UCF101 and HMDB51. Best results are **bolded** and second-best results are underlined.

Dataset	Encoder	Retrieval mAP (%)		
		Unmasked	Binary mask	Soft mask
UCF101	V-JEPA 2.1 Large	40.82	41.34	**42.37**
	VideoMAEv2-Base	95.52	95.61	**95.65**
HMDB51	V-JEPA 2.1 Large	17.84	18.40	**18.99**
	VideoMAEv2-Base	**46.94**	46.39	46.66

and unsupervised feature-selection methods, as well as training with a standard similarity objective, which we further discuss in the ablation studies and Table 4.

4.3 Video retrieval

Measuring semantic similarity between high-dimensional video embeddings is fundamental for video retrieval. We focus on UCF101 (Soomro et al., 2012) and HMDB51 (Kuehne et al., 2011), where given a query video, the objective is to retrieve videos belonging to the same class as the query, thus showing high semantic similarity. We thus benchmark our masked features on both datasets using both V-JEPA 2.1 and VideoMAEv2 feature spaces.

Masked features improve video retrieval. Table 3 presents the mAP results for the standard features and the soft and binary masked ones. Our results indicate that our learned masks consistently improve retrieval performance on both datasets and features spaces, with a slight decrease for VideoMAEv2 on HMDB51. Additional results with different video generators are reported in Appendix C.4 Table C.4, with the strongest performance coming from masks trained on CogVideoX.

4.4 Ablations

To validate the design choices of our proposed feature masking approach, we conduct extensive ablation studies. We aim to answer three primary questions: **(Q1)** Is our *generative consistency* objective necessary, or would a standard similarity-based objective suffice? **(Q2)** Given the high sparsity of our learned masks, is the performance gain simply a byproduct of dimensionality reduction, or are we identifying a highly specific subset of features? And **(Q3)** Do our highly sparse masks destroy essential visual and structural information?

The necessity of generative consistency (Q1). A naive alternative for learning the feature mask is to optimize for a standard similarity objective. Under this setup, we follow the same training strategy but fundamentally change the definition of positive (similar) and negative (dissimilar) pairs to learn

Table 4: **Ablation study on feature masking strategies and training objectives.** Results are averaged across four OOD datasets (UVEB, MedVidBench, HAM10000, and WikiArt). We compare our masking strategy against the standard similarity objective (**Q1**) and unsupervised feature-selection baselines (**Q2**) evaluated at equal sparsity ($k = 329$ for V-JEPA 2.1-L, $k = 322$ for VideoMAEv2-B).

Method	V-JEPA 2.1-L						VideoMAEv2-B					
	2XDM			DDPM-OOD			2XDM			DDPM-OOD		
	AUROC↑	AUPR↑	FPR@95↓	AUROC↑	AUPR↑	FPR@95↓	AUROC↑	AUPR↑	FPR@95↓	AUROC↑	AUPR↑	FPR@95↓
Baseline (unmasked)	71.75	66.25	65.40	75.49	72.94	70.90	68.96	65.68	80.10	75.66	76.08	79.90
Similarity objective	74.39	70.30	65.50	80.97	78.30	67.90	68.90	65.81	82.80	76.06	77.04	81.50
Ours (binary mask)	**81.73**	**79.03**	**58.20**	**89.65**	**88.65**	**45.60**	**70.45**	66.63	**76.90**	**77.69**	**77.73**	82.40
PCA projection	70.33	65.08	69.80	70.52	67.91	77.50	63.07	59.18	87.80	67.89	66.36	81.20
Random mask	70.91	65.62	66.95	72.68	70.54	74.30	67.84	63.80	81.03	74.73	74.59	80.70
Self-attention	70.29	64.78	64.80	69.68	67.63	81.10	70.03	**67.02**	80.50	76.18	76.57	78.60
Variance (top-k)	70.27	64.80	65.40	69.61	67.50	80.90	69.49	66.39	80.50	75.46	75.86	78.60
PCA loading	70.22	64.72	65.90	69.70	67.56	80.30	69.54	66.48	82.30	75.49	75.82	**77.40**
Laplacian score	69.14	63.74	69.50	67.54	66.11	83.50	69.03	65.80	80.60	75.13	75.13	81.00
Correlation filter	73.20	68.04	65.10	78.89	76.92	65.60	69.30	66.38	79.30	75.67	76.18	80.60

our mask. In particular, we ignore the idea behind general and specialized concepts and label a pair of videos as semantically similar ($y = 1$) if generated from the same prompt and build negative pairs ($y = 0$) using videos that come from different prompts. Once again, given that we have generated our dataset with a T2V model, there is no need for manual labeling. As shown in Table 4, while this similarity objective (row 2) provides a marginal improvement over the baseline (unmasked) embeddings, falling significantly short of our proposed method. This suggests that video pairs labeled based on generative consistency, i.e., belonging to *general* or *specialized* concepts, capture more significant semantic differences compared to just considering two videos similar because they have the same description (source prompt).

Learned masks vs dimensionality reduction (Q2). Our learned binary masks exhibit high sparsity. For instance, with V-JEPA 2.1-L, the mask zeros out 67.8% of the features using a threshold $\theta = 0.5$, effectively utilizing only 329 of the original 1024 embedding dimensions. Similar patterns emerge for VideoMAEv2 (58.0% sparsity). To ensure this improvement is not an artifact of simply reducing the feature space, we compare our method against several dimensionality reduction and feature selection baselines, including random masking, variance-based feature selection, PCA-based methods, Laplacian Score, correlation-based filtering, and a self-attention-derived mask, which we further describe in the Appendix C.2. The results in the second half of Tab. 4 validate our approach. Across both models and OOD methods, none of the other unsupervised feature selection or dimensionality reduction techniques perform as well as the subspace identified by our learning strategy. This indicates that the features most critical for similarity are not trivial to define. Notably, our method drastically reduces the False Positive Rate (FPR@95). For instance, in the V-JEPA DDPM-OOD setup, our method reduces the FPR@95 from the unmasked baseline of 70.90% down to 45.60%. These results confirm that these models possess highly capable, latent subspaces for OOD detection, but that they must be isolated using a targeted objective rather than generic dimensionality reduction. Additional results in Table C.1 (Appendix C.3) confirm this trend on ConViS-Bench: our generative-consistency masks outperform all the aforementioned baselines at matched sparsity.

4.4.1 Preservation of visual information

To verify that our highly sparse masks do not destroy essential visual information (**Q3**), we evaluate the learned representations both quantitatively and qualitatively.

PCA feature visualization. To qualitatively illustrate this preserved spatial structure, we visualize the feature space directly using PCA. For a given video frame, we extract patch-level features using the V-JEPA 2.1-L encoder, compute PCA to extract the top three principal components, and map them directly to RGB color channels. The visual comparisons are presented in Figure 4, displaying the original input frames (top row), the PCA visualizations of the unmasked features (middle row), and our masked representations (bottom row). The masked features exhibit remarkably similar structural outlines, object boundaries, and background separation compared to the full baseline, demonstrating their retention of core visual features despite the high sparsity.

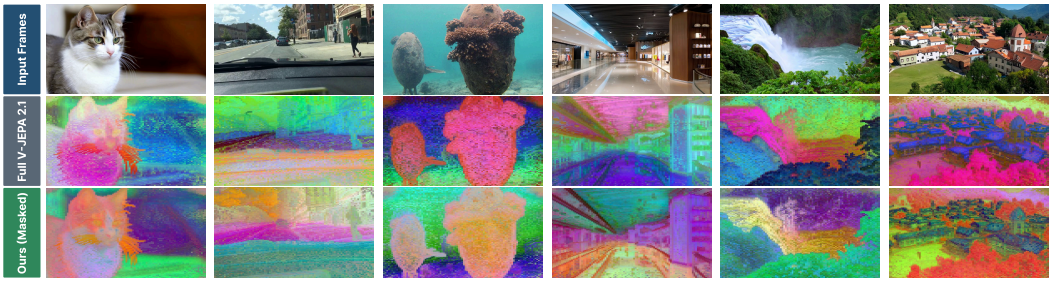


Figure 4: **PCA visualization of V-JEPA 2.1-L patch-level features mapped to RGB channels.** (Top) Original input frames. (Middle) Full V-JEPA features. (Bottom) Our masked representations.

Patch clustering for spatial reconstruction. Following prior work on analyzing the spatial structure encoded in visual representations (Dhanachandra et al., 2015), we cluster both masked and unmasked patch features of one frame per video to obtain a piecewise segmentation, map the segments to their mean pixel value, and compute the MSE_k to the original image, where k denotes the number of clusters. We report the final MSE_k averaged across all evaluated frames in the BDD100k dataset across varying cluster counts showing that even after masking out a substantial portion of the embedding dimensions, the representations still retain much of the original visual and structural information. Details in Appendix C.6.

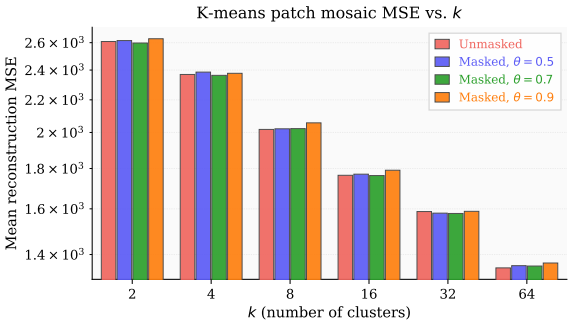


Figure 5: **Quantitative information loss evaluation via K-means mosaic reconstruction.** The plot compares the Mean Squared Error (MSE) between the original RGB frames and the mosaic reconstructions across varying cluster counts (k). The V-JEPA 2.1-L masked representations evaluate multiple binarization thresholds (θ).

5 Conclusion

In this paper, we introduced a self-supervised framework for learning semantic video similarity by leveraging generative uncertainty in text-to-video diffusion models. Rather than relying on human similarity annotations or predefined similarity metrics, our approach uses the consistency of generated videos as a supervisory signal: general concepts tend to produce stable generations, while specialized concepts induce higher generative divergence. We exploit this behavior to learn sparse masks over pretrained video representations, isolating feature subspaces that better capture semantic similarity. Across multiple backbones and tasks, the learned masked representations consistently improve alignment with human similarity judgments, enhance out-of-distribution detection for generated videos, and maintain or improve video retrieval performance. Importantly, these gains are achieved while discarding a large fraction of the original embedding dimensions, leading to more compact and efficient representations. Our dense-task experiments further show that the learned global masks preserve meaningful local structure, enabling efficient patch-level matching with minimal loss in segmentation quality. Finally, we reproduce the experiments across three different text-to-video models, validating our results on independent sets of videos. These results suggest that generative models can provide more than synthetic data: their uncertainty can serve as a useful signal for learning semantically useful subspaces.

AI ASSISTANCE

Generative AI tools were used in different ways. Assisting with polishing the writing of the manuscript and supporting the implementation of the code of reported experiments, which however were carefully verified by the authors. Synthetic data generation: Google Gemini 3.1 Pro was used to generate the specialized text prompts employed in constructing the synthetic training dataset, as described in Appendix B.1. Text-to-video generative models (Wan2.1, CogVideoX1.5-5B, Mochi1) were used to generate the video datasets for the proposed method (Sec. 3.1).

We therefore selected the following options on OpenReview:

- Yes, to aid or polish writing.
- Yes, for research ideation or execution. Referring to the usage for assisted coding.
- Yes, for generating synthetic datasets. As described in the paper.

We have not used generative AI tools for other tasks with required disclosure.

REFERENCES

Lucas Berry, Axel Brando, and David Meger. Shedding light on large generative networks: Estimating epistemic uncertainty in diffusion models. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024. URL https://openreview.net/forum?id=512IkGDqA8.

Michele De Vita and Vasileios Belagiannis. Diffusion model guided sampling with pixel-wise aleatoric uncertainty estimation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3844–3854. IEEE, 2025.

Nameirakpam Dhanachandra, Khumanthem Manglem, and Yambem Jina Chanu. Image segmentation using k -means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54:764–771, 2015. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2015.06.090. URL https://www.sciencedirect.com/science/article/pii/S1877050915014143. Eleventh International Conference on Communication Networks, ICCN 2015, August 21-23, 2015, Bangalore, India Eleventh International Conference on Data Mining and Warehousing, ICDMW 2015, August 21-23, 2015, Bangalore, India Eleventh International Conference on Image and Signal Processing, ICISP 2015, August 21-23, 2015, Bangalore, India.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1801–1810, 2019.

Gianni Franchi, Nacim Belkhir, Dat Nguyen Trong, Guoxuan Xia, and Andrea Pilzer. Towards understanding and quantifying uncertainty for text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8062–8072, 2025.

Zhanning Gao, Gang Hua, Dongqing Zhang, Nebojsa Jojic, Le Wang, Jianru Xue, and Nanning Zheng. Er3: A unified framework for event retrieval, recognition and recounting. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2253–2262, 2017.

Mark S. Graham, Walter H.L. Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2947–2956, June 2023.

Mark A Hall. Correlation-based feature selection of discrete and numeric class machine learning. 2000.

Zhen Han, Xiangteng He, Mingqian Tang, and Yiliang Lv. Video similarity and alignment learning on partial video copy detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pp. 4165–4173, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386517. doi: 10.1145/3474085.3475549. URL https://doi.org/10.1145/3474085.3475549.

Xiangteng He, Yulin Pan, Mingqian Tang, Yiliang Lv, and Yuxin Peng. Learn from unlabeled videos for near-duplicate video retrieval. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 1002–1011, 2022.

Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In Y. Weiss, B. Schölkopf, and J. Platt (eds.), *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005. URL https://proceedings.neurips.cc/paper_files/paper/2005/file/b5b03f06271f8917685d14cea7c6c50a-Paper.pdf.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Metod Jazbec, Eliot Wong-Toi, Guoxuan Xia, Dan Zhang, Eric Nalisnick, and Stephan Mandt. Generative uncertainty in diffusion models. *arXiv preprint arXiv:2502.20946*, 2025.

Won Jo, Geuntaek Lim, Gwangjin Lee, Hyunwoo Kim, Byungsoo Ko, and Yukyung Choi. Vvs: Video-to-video retrieval with irrelevant frame suppression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 2679–2687, 2024.

Ian Jolliffe. Principal component analysis. *Encyclopedia of statistics in behavioral science*, 2005.

Giorgos Kordopatis-Zilos, Giorgos Tolias, Christos Tzelepis, Ioannis Kompatsiaris, Ioannis Patras, and Symeon Papadopoulos. Self-supervised video similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pp. 2556–2563. IEEE, 2011.

Benedetta Liberatori, Alessandro Conti, Lorenzo Vaquero, Yiming Wang, Elisa Ricci, and Paolo Rota. Convis-bench: Estimating video similarity through semantic concepts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=NoIWLerNKH.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22139–22149, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, and Nicolas Ballas. V-jepa 2.1: Unlocking dense features in video self-supervised learning. In *European Conference on Computer Vision*, pp. 671–689. Springer, 2026.

Kennard Jun Hao Ng, Ser Nam Lim, and Gim Hee Lee. Vrag: Region attention graphs for content-based video retrieval. *ArXiv*, abs/2205.09068, 2022. URL https://api.semanticscholar.org/CorpusID:248863060.

NVIDIA. Cosmos-predict2: World simulation model for physical ai, 2025. URL https://github.com/nvidia-cosmos/cosmos-predict2.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

Jie Shao, Xin Wen, Bingchen Zhao, and Xiangyang Xue. Temporal context aggregation for video retrieval with contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3268–3278, January 2021.

Minho Sim, Young-Jun Lee, Dongkun Lee, Jongwhoa Lee, and Ho-Jin Choi. A simple debiasing framework for out-of-distribution detection in human action recognition. In *ECAI 2023: 26th European Conference on Artificial Intelligence, September 30–October 4, 2023, Kraków, Poland–Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, pp. 2154–2161. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2023.

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.

Yuhao Su, Anwesa Choudhuri, Zhongpai Gao, Benjamin Planche, Van Nguyen Nguyen, Meng Zheng, Yuhan Shen, Arun Innanje, Terrence Chen, Ehsan Elhamifar, and Ziyan Wu. MedGRPO: Multi-task reinforcement learning for heterogeneous medical video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Genmo Team. Mochi 1. `https://github.com/genmoai/models`, 2024.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5, 08 2018. doi: 10.1038/sdata.2018.161. URL `https://www.nature.com/articles/sdata2018161`.

Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. *International Journal of Computer Vision*, 129(7):2264–2287, 2021.

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.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14560, June 2023.

Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. *Advances in Neural Information Processing Systems*, 36:54683–54695, 2023.

Yaofeng Xie, Lingwei Kong, Kai Chen, Ziqiang Zheng, Xiao Yu, Zhibin Yu, and Bing Zheng. Uveb: A large-scale benchmark and baseline towards real-world underwater video enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22358–22367, 2024.

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

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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

A Theoretical Analysis of Generative Variability

A.1 Why is generative variability a useful learning signal?

Our method relies on a simple observation: when the same text prompt is sampled several times, the generated videos can be either very consistent or highly variable. We use this variability as a self-supervised signal to identify which dimensions of a pretrained video representation are useful for semantic similarity.

Importantly, our analysis does not assume that a specialized prompt is necessarily absent from the training set of the video generator. It also does not require all generative variability to be epistemic uncertainty. Instead, we only study the variability of repeated generations conditioned on the same prompt, and show why differences in this variability can reveal useful feature dimensions.

Variability of repeated video generations. Let p denote a text prompt and let

$$\mathbf{v} \sim p_\theta(\cdot \mid p) \tag{A.1}$$

be a video generated by a stochastic text-to-video model. For a fixed prompt p, we consider two independent generations

$$\mathbf{v}, \mathbf{v}' \overset{\text{i.i.d.}}{\sim} p_\theta(\cdot \mid p). \tag{A.2}$$

Intuitively, if the generator represents the content of p in a stable way, V and V' should contain similar semantic information. If several different semantic realizations are plausible, repeated generations can instead be substantially different.

To formalize this idea, suppose that the possible videos generated from p can be grouped into K_p semantic outcomes,

$$\mathcal{V}_p = \bigcup_{k=1}^{K_p} \mathcal{C}_{p,k}, \tag{A.3}$$

where videos in the same set $\mathcal{C}_{p,k}$ express the same main semantic content. Let

$$g_p(\mathbf{v}) \in \{1, \dots, K_p\} \tag{A.4}$$

denote the semantic outcome of $\mathbf{v}$, and define

$$\pi_k(p) = \Pr\left(g_p(\mathbf{v}) = k \mid p\right). \tag{A.5}$$

We can then define the semantic variability of the video generator as

$$\mathcal{V}_{\text{sem}}(p) = \Pr\left(g_p(\mathbf{v}) \neq g_p(\mathbf{v}') \mid p\right). \tag{A.6}$$

Proposition 1 (Pairwise disagreement measures semantic variability). For two independent generations from the same prompt,

$$\boxed{\mathcal{V}_{\text{sem}}(p) = 1 - \sum_{k=1}^{K_p} \pi_k(p)^2.} \tag{A.7}$$

Proof. Since V and V' are independent conditioned on p,

$$\Pr\left(g_p(\mathbf{v}) = g_p(\mathbf{v}') \mid p\right) = \sum_{k=1}^{K_p} \Pr\left(g_p(\mathbf{v}) = k \mid p\right) \Pr\left(g_p(\mathbf{v}') = k \mid p\right) \tag{A.8}$$

$$= \sum_{k=1}^{K_p} \pi_k(p)^2. \tag{A.9}$$

Taking the complementary event gives equation A.7. □

Equation A.7 gives a simple interpretation of repeated generation. If most of the probability mass is concentrated on one semantic outcome, two generations are likely to agree and $\mathcal{V}_{\text{sem}}(p)$ is small. If the probability mass is spread over several possible outcomes, $\mathcal{V}_{\text{sem}}(p)$ becomes larger.

As above, let $g_p(\mathbf{v})$ denote the semantic content of $\mathbf{v}$. We define the semantic variability of the generator as

$$\mathcal{V}_{\text{sem}}(p) = \Pr\left(g_p(\mathbf{v}) \neq g_p(\mathbf{v}') \mid p\right). \tag{A.10}$$

We now define the ideal semantic similarity between two videos as

$$s^\star(\mathbf{v}, \mathbf{v}') = 1\left[g_p(\mathbf{v}) = g_p(\mathbf{v}')\right]. \tag{A.11}$$

Thus, $s^\star = 1$ when the two videos have the same semantic content and $s^\star = 0$ otherwise.

Proposition 2 (Uncertainty and semantic similarity). For two independent videos generated from the same prompt,

$$\boxed{\mathbb{E}\left[s^\star(\mathbf{v}, \mathbf{v}') \mid p\right] = 1 - \mathcal{V}_{\text{sem}}(p).} \tag{A.12}$$

Proof. By definition,

$$\begin{aligned}
\mathbb{E}\left[s^\star(\mathbf{v}, \mathbf{v}') \mid p\right] &= \Pr\left(g_p(\mathbf{v}) = g_p(\mathbf{v}') \mid p\right) && \text{(A.13)}\\
&= 1 - \Pr\left(g_p(\mathbf{v}) \neq g_p(\mathbf{v}') \mid p\right) && \text{(A.14)}\\
&= 1 - \mathcal{V}_{\text{sem}}(p). && \text{(A.15)}
\end{aligned}$$

□

Interpretation. Proposition A.12 gives a direct connection between generative uncertainty and semantic similarity. If the uncertainty of the generator is low, two independent generations are likely to have the same semantic content, and their expected semantic similarity is high. Conversely, if the uncertainty is high, two generations are more likely to have different semantic contents, and their expected semantic similarity is lower. We want to insist on the assumption of the ideal semantic similarity of equation A.11, which is linked with the fact that two videos are similar if they have similar concepts. Under this assumption, uncertainty can be linked to the ideal semantic similarity and provide a natural supervision signal for learning video similarity, In our work, we exploit this uncertainty signal to learn similarity over the feature space defined by a pretrained video encoder $\mathcal{E}$.

B Implementation details

B.1 *Specialized* prompt generation

We ensure that the *specialized* prompts differ from the *general* prompts strictly in their semantic domain, rather than through superficial linguistic artifacts (e.g., varying sequence lengths, different grammatical structures, or stylistic shifts). To achieve this controlled domain shift, we leveraged Google's Gemini 3.1 Pro to generate a parallel dataset of 100 specialized concepts.

We explicitly constrained the language model to synthesize prompts focused on highly specialized, non-everyday domains (specifically, clinical surgery and cellular microscopy) while perfectly mirroring the structural distribution of the general prompts. The exact prompt provided to the language model was as follows:

> *"I have a dataset of 100 general concept prompts. I need an equivalent set of 100 specialized concept prompts that perfectly match the general dataset in detail, tone, exact length distribution, and concise structural style. The specialized prompts must focus exclusively on the domains of clinical surgery (50 prompts) and cellular microscopy (50 prompts). Additionally, they must be entirely devoid of everyday objects, animals, and human subjects (e.g., 'scissors', 'dog', 'person')."*
>
> `'general_prompts.txt' attached`

By enforcing these constraints, we reduce the likelihood that the observed generative divergence is driven by superficial linguistic differences between the general and specialized prompts.

Here are some examples of *general* and *specialized* prompts used for video generation:

General concepts

- "cheetah lying on the grass"
- "scenery of desert landscape"
- "a relaxing scenery of beach view under cloudy sky"
- "a chef holding and checking kitchen utensils"
- "people packing their furniture"

Specialized concepts

- "skin closure with sutures"
- "surgical retractor holding back thick adipose tissue"
- "macrophage engulfing rod bacteria"
- "fluorescent view of actin filaments contracting muscle"
- "sealing a bleeding vessel with bipolar cautery"

We provide a `.zip` file for supplementary material containing the full list of prompts used in this work and some examples of generated videos for both general and specialized domains.

B.2 Learnable mask training

We optimize the continuous soft mask by partitioning our synthetic dataset of general and specialized video pairs into training (80%) and validation (20%) sets. The mask parameters are updated using the AdamW optimizer with a learning rate of 1×10^{-3} and a batch size of 128. To prevent overfitting, we employ an early stopping mechanism on the validation set. Features from the used video encoder were extracted and stored before training to speed up the process. Given that we use $y = 1$ for positive pairs and $y = 0$ for negative pairs, during training the computed cosine similarity is rescaled from $[-1, 1]$ to $[0, 1]$ range to match the target y. All training procedures were executed on a single NVIDIA H100 GPU, with an average training time of approximately 1 minute.

B.3 Cosmos-Predict2 finetuning

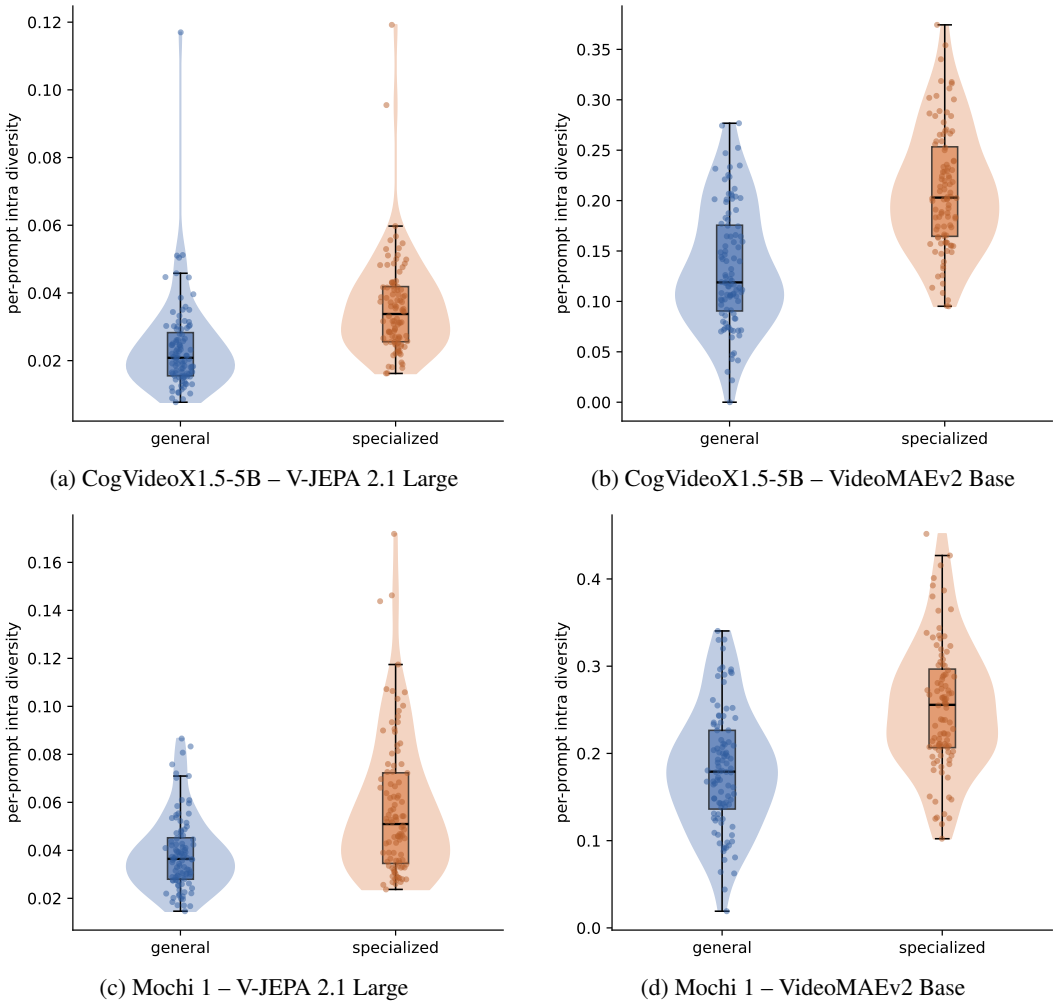
To ensure that driving scenarios from the BDD100k dataset could indeed be considered as In-Distribution (ID) for the Cosmos model, we performed a short fine-tuning. For our experiments, we used `cosmos-predict2` 2B model, and fine-tuned it on BDD100k with batch size 32 and learning rate 1.1×10^{-6} for 5000 iterations, approximately 12 hours using 16 NVIDIA H100 GPUs. The resulting model was then used to perform all the image-to-video generation needed for the experiment as described in sec. 4.2.

C Additional analysis

C.1 Additional video generators

We repeat the intra-prompt diversity analysis using videos generated by CogVideoX1.5-5B and Mochi 1. Figure C.1 shows the results for CogVideoX1.5-5B and Mochi 1. Consistent with the Wan2.1 results in the main paper, specialized prompts exhibit higher intra-prompt feature-space diversity than general prompts in both V-JEPA 2.1-L and VideoMAEv2-B feature spaces. Together with the downstream results obtained using separately learned Wan2.1-, CogVideoX-, and Mochi-based masks, these results show that the proposed learning signal is not specific to a particular video diffusion generator.

We further provide qualitative examples from three different text-to-video generators. Figure C.2 compares generations obtained from general and specialized prompts using three random seeds. Across all three generators, general prompts tend to produce more consistent content across seeds, while specialized prompts exhibit greater visual and semantic variation. This suggests that the proposed generative uncertainty signal is not specific to a particular video generator.



(a) CogVideoX1.5-5B – V-JEPA 2.1 Large

(b) CogVideoX1.5-5B – VideoMAEv2 Base

(c) Mochi 1 – V-JEPA 2.1 Large

(d) Mochi 1 – VideoMAEv2 Base

Figure C.1: **Per-prompt intra-set feature-space diversity for videos generated by CogVideoX1.5-5B (top row) and Mochi 1 (bottom row), measured using V-JEPA 2.1 Large (left column) and VideoMAEv2 Base (right column).** Each point represents one prompt and reports the average pairwise normalized cosine dissimilarity among three videos generated with different random seeds. Specialized prompts exhibit higher intra-prompt diversity than general prompts in both feature spaces, indicating lower consistency across generations.

C.2 Feature selection and Dimensionality reduction Baselines

We compare our generative-consistency mask with several unsupervised feature-selection and dimensionality-reduction methods. The goal is to test whether the gains of our method can be explained simply by reducing the representation size, keeping high-variance features, removing redundant dimensions, or preserving the geometry of the original embedding space. For all masking baselines, we retain the same number of feature dimensions as our learned binary mask: $k = 329/1024$ for V-JEPA 2.1-L and $k = 322/768$ for VideoMAEv2-B.

Random mask. We randomly select k feature dimensions from the original representation. This serves as a sparsity control and tests whether the improvements arise simply from reducing the representation dimensionality.

Variance (top-k). We rank feature dimensions by their empirical variance across videos and retain the k highest-variance dimensions. This baseline favors statistically active channels but does not explicitly account for semantic relevance.

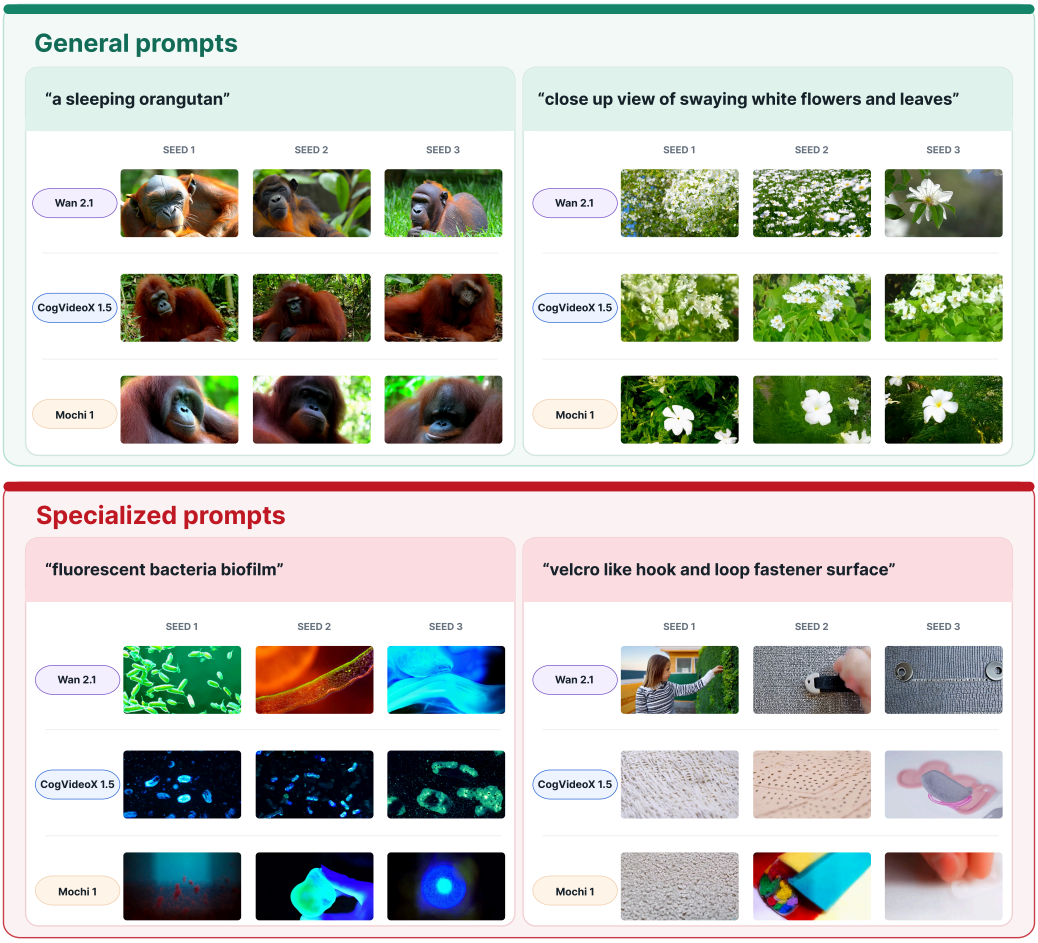


Figure C.2: **Cross-generator qualitative comparison of intra-prompt consistency.** For each general and specialized prompt, we show three videos generated with different random seeds using Wan 2.1, CogVideoX1.5, and Mochi 1. Each tile contains sampled frames from one generated video. General prompts produce comparatively consistent generations across seeds, whereas specialized prompts exhibit greater visual and semantic variation across all three generators.

PCA loading. We perform PCA (Jolliffe, 2005) on the video representations and rank the original feature dimensions according to their contribution to the dominant principal components. The top-k dimensions are retained, allowing us to test whether our mask primarily captures the dominant variance structure of the pretrained representation.

Laplacian score. We use Laplacian Score (He et al., 2005) as an unsupervised feature-selection method that favors dimensions that preserve the local neighborhood structure of the embedding space. Features with the lowest scores are retained. This baseline evaluates whether preserving the representation's local manifold structure is sufficient to obtain similar improvements.

Correlation filter. We rank features according to their redundancy with other channels using pairwise feature correlations. Dimensions that provide comparatively non-redundant information are preferred. (Hall, 2000) This baseline tests whether the benefits of our method can be explained by removing redundant channels.

Self-attention mask. We additionally construct a feature mask from the encoder's final-block self-attention activations. Since self-attention is defined over tokens rather than embedding dimensions,

we combine token-level attention weights with the corresponding patch features to obtain a channel-importance score, average these scores over the generated videos, and retain the top-k channels.

This baseline tests whether the encoder's internal attention saliency alone identifies a subspace suitable for semantic similarity. As shown in Tables C.1 and 4, the attention-derived mask does not provide improvements comparable to our generative-consistency objective.

C.3 Full Results on ConViS-Bench

Table C.1 reports the complete results on ConViS-Bench across the five semantic dimensions: main action, main subjects, main objects, location, and action order. We compare the original pretrained representations with the feature-selection baselines introduced in Appendix C.2, as well as our binary and soft masks. For all sparse methods, we keep exactly the same number of dimensions as our binary mask. Therefore, the comparison isolates the effect of *which* dimensions are selected, rather than the effect of sparsity itself.

Results with V-JEPA 2.1-L. For V-JEPA 2.1-L, generic feature-selection methods do not improve the alignment with human judgements. Random selection, variance-based selection, PCA loading, Laplacian Score, correlation filtering, and the self-attention mask all obtain lower average correlations than the original representation. For instance, the unmasked features obtain an average Spearman correlation of 31.14, while the strongest non-learned baseline reaches 29.74.

In contrast, our binary mask increases the average Spearman correlation to 35.44 and the average Kendall correlation from 21.53 to 24.49. The improvement is also consistent across all five semantic dimensions. This result is obtained while discarding 67.8% of the original feature dimensions. The soft mask performs slightly better, reaching an average Spearman correlation of 36.04. This suggests that the generative-consistency objective identifies dimensions that are particularly useful for semantic similarity, rather than simply removing low-variance or redundant features.

Results with VideoMAEv2-B. The same general trend is observed with VideoMAEv2-B, although the gains are smaller. Unlike V-JEPA, some generic feature-selection methods can slightly improve the unmasked representation. For example, the random mask increases the average Spearman correlation from 44.90 to 45.42. This suggests that the VideoMAEv2 representation contains some redundancy and that reducing its dimensionality can already be beneficial. However, our learned masks still provide the strongest overall alignment.

C.4 Downstream Robustness to the Video Generator

The previous analysis shows that the difference in generative consistency between general and specialized concepts is observed across multiple text-to-video generators. We next investigate whether this property also results in useful feature masks when the source generator used to construct the synthetic training set is changed.

To this end, we independently train masks using the synthetic datasets generated by CogVideoX1.5-5B and Mochi 1 and evaluate them using the same downstream protocols as the Wan2.1-based masks in the main experiments. Importantly, only the source of the synthetic mask-training data changes; the downstream datasets, feature encoders, and evaluation procedures remain unchanged.

Human similarity alignment. Table C.2 evaluates masks learned from CogVideoX1.5-5B and Mochi 1 on ConViS-Bench. For both V-JEPA 2.1-L and VideoMAEv2-B, masks learned from either generator improve the average correlation with human similarity judgements compared with the corresponding unmasked representations. For V-JEPA 2.1-L, for example, the average Spearman correlation increases from 31.14 for the unmasked representation to 37.20 and 35.66 using binary masks derived from soft masks learned on CogVideoX1.5-5B and Mochi 1, respectively. The corresponding soft masks achieve average correlations of 38.14 and 36.78. A similar trend is observed with VideoMAEv2-B. These results show that the semantic subspaces recovered by our objective remain useful when the synthetic training videos are produced by different generators.

OOD detection. We next evaluate whether the generator-specific masks transfer to video-generation OOD detection. Recall that these masks are learned solely from synthetic text-to-video generations,

Table C.1: **Alignment with human judgement on ConViS-Bench.** Spearman's ρ and Kendall's τ correlations with human judgement ($\times 100$). *Sp.* denotes the sparsity of the mask; soft masks are dense by construction. Unsupervised feature selectors are fit on the video embeddings alone and matched to the sparsity of our learned mask. Best results are **bolded**, second-best underlined.

Encoder	Method	Sp. (%)	Main Action		Main Subjects		Main Objects		Location		Actions Order		Average	
			ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
V-JEPA 2.1-L	Baseline (no mask)	–	24.0	16.30	33.4	22.84	24.7	16.83	49.3	35.04	24.3	16.65	31.14	21.53
	Random mask	67.8	22.8	15.42	31.8	21.77	23.3	15.84	47.3	33.53	22.8	15.61	29.60	20.43
	Variance (top-k)	67.8	21.7	14.67	31.4	21.51	21.3	14.51	47.8	33.75	22.1	15.16	28.86	19.92
	PCA loading	67.8	21.5	14.54	31.3	21.45	21.1	14.43	47.8	33.76	21.9	15.10	28.72	19.86
	Laplacian score	67.8	21.7	14.66	30.9	21.16	20.7	14.14	47.0	33.26	22.0	15.10	28.46	19.66
	Correlation filter	67.8	23.9	16.19	31.1	21.34	23.0	15.55	46.2	32.66	24.5	16.84	29.74	20.52
	Self-attention mask	67.8	21.6	14.54	31.1	21.23	21.0	14.32	47.6	33.67	21.8	14.94	28.62	19.74
	Binary mask (Ours)	67.8	26.9	18.25	39.4	27.14	30.3	20.57	50.5	35.74	30.1	20.75	35.44	24.49
	Soft mask (Ours)	–	**27.7**	**18.78**	**39.5**	**27.24**	**30.8**	**20.89**	**51.3**	**36.41**	**30.9**	**21.23**	**36.04**	**24.91**
VideoMAEv2-B	Baseline (no mask)	–	47.0	33.02	44.6	31.46	42.4	29.96	45.8	32.26	44.7	31.37	44.90	31.61
	Random mask	58.0	47.3	33.31	45.6	32.11	43.0	30.29	46.3	32.68	44.9	31.47	45.42	31.97
	Variance (top-k)	58.0	45.9	32.16	43.0	30.09	40.4	28.43	44.4	31.33	43.9	30.69	43.52	30.54
	PCA loading	58.0	45.8	32.13	43.1	30.19	40.4	28.44	44.7	31.47	43.7	30.60	43.54	30.57
	Laplacian score	58.0	46.4	32.55	45.6	32.17	42.6	30.05	45.9	32.45	44.6	31.23	45.02	31.69
	Correlation filter	58.0	45.1	31.70	42.6	29.83	40.2	28.28	42.7	29.94	42.6	29.77	42.64	29.90
	Self-attention mask	58.0	44.4	31.04	43.0	30.18	39.4	27.79	44.8	31.57	42.3	29.61	42.78	30.04
	Binary mask (Ours)	58.0	47.9	33.62	**46.3**	**32.63**	**44.1**	**30.98**	**46.9**	**33.09**	45.3	31.67	**46.10**	**32.40**
	Soft mask (Ours)	–	**48.0**	**33.74**	46.1	32.47	44.0	30.97	46.7	32.92	**45.4**	**31.79**	46.04	32.38

Table C.2: **Effect of the mask source generator on ConViS-Bench.** Spearman's ρ and Kendall's τ correlations with human judgement ($\times 100$). *Sp.* denotes the sparsity of the learned mask; soft masks are dense by construction. Masks learned from synthetic datasets generated by either CogVideoX1.5-5B or Mochi 1 improve average alignment with human judgement over the corresponding unmasked representations, demonstrating robustness to the choice of source video generator. Best results are **bolded**, second-best underlined.

Encoder	Method	Sp. (%)	Main Action		Main Subjects		Main Objects		Location		Actions Order		Average	
			ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
V-JEPA 2.1-L	Baseline (no mask)	–	24.0	16.30	33.4	22.84	24.7	16.83	49.3	35.04	24.3	16.65	31.14	21.53
	Binary mask (CogVideoX)	75.3	30.8	21.02	37.9	26.12	31.4	21.46	51.7	36.94	34.2	23.67	37.20	25.84
	Binary mask (Mochi)	72.1	28.2	19.30	37.5	25.91	30.1	20.48	52.3	37.44	30.2	20.77	35.66	24.78
	Soft mask (CogVideoX)	–	**31.8**	**21.76**	**38.9**	**26.94**	**32.3**	**22.08**	**53.0**	37.88	**34.7**	**24.04**	**38.14**	**26.54**
	Soft mask (Mochi)	–	29.4	20.04	38.0	26.26	31.6	21.47	53.0	**37.97**	31.9	22.04	36.78	25.56
VideoMAEv2-B	Baseline (no mask)	–	47.0	33.02	44.6	31.46	42.4	29.96	45.8	32.26	**44.7**	**31.37**	44.90	31.61
	Binary mask (CogVideoX)	57.8	46.8	32.91	**45.9**	**32.26**	44.0	30.88	**47.0**	33.17	43.8	30.59	45.50	31.96
	Binary mask (Mochi)	51.9	**47.2**	**33.19**	45.6	32.18	**44.5**	**31.25**	**47.0**	**33.20**	**44.7**	31.27	**45.80**	**32.22**
	Soft mask (CogVideoX)	–	46.9	32.99	45.7	32.22	43.8	30.71	46.7	32.94	44.0	30.68	45.42	31.91
	Soft mask (Mochi)	–	46.9	33.04	45.7	32.19	44.2	31.08	46.9	33.19	44.3	31.03	45.60	32.11

whereas the downstream OOD experiment uses Cosmos-Predict2 for image-to-video generation. This therefore evaluates transfer not only across datasets but also across generative tasks and model families.

Table C.3 reports the results obtained using masks learned from CogVideoX1.5-5B's generated videos. Across the four OOD datasets, both feature encoders, and the 2XDM and DDPM-OOD scoring procedures, the learned masks provide substantial improvements over the unmasked representations in most evaluation settings. Particularly large gains are observed for MedVidBench and HAM10000, demonstrating that the generative-consistency signal extracted from CogVideoX generations transfers effectively to OOD detection with a different generative model.

Table C.5 repeats the same evaluation using masks learned from Mochi 1. The overall behavior is consistent with the CogVideoX1.5-5B results: masking improves OOD discrimination across a broad range of datasets, feature encoders, and scoring procedures, with especially clear improvements on MedVidBench and HAM10000. Although the magnitude of the gains varies with the source generator and downstream configuration, the masks learned from both alternative generators remain

Table C.3: **OOD detection using masks learned from CogVideoX1.5-5B.** Video generation OOD detection across four datasets using Cosmos-Predict2, with BDD100k as the in-distribution (ID) baseline. The soft masks are learned exclusively from synthetic videos generated by CogVideoX1.5-5B, while the corresponding binary masks are obtained by post-hoc thresholding at $\theta = 0.5$. The resulting improvements over the unmasked representations demonstrate transfer across source video generators and generative tasks. Best results are **bolded** and second-best results are underlined.

OOD Dataset	Method	V-JEPA 2.1-L						VideoMAEv2-B					
		2XDM			DDPM-OOD			2XDM			DDPM-OOD		
		AUROC↑	AUPR↑	FPR@95↓	AUROC↑	AUPR↑	FPR@95↓	AUROC↑	AUPR↑	FPR@95↓	AUROC↑	AUPR↑	FPR@95↓
UVEB	Unmasked	78.42	73.77	57.20	76.37	70.31	63.60	71.08	68.50	77.60	75.50	73.33	**83.20**
	Binary mask	84.19	82.29	**52.80**	**89.43**	88.51	42.40	**73.18**	**70.24**	**71.20**	**78.97**	**76.73**	84.40
	Soft mask	**84.63**	**82.98**	**52.80**	**89.43**	**88.54**	**42.00**	72.98	70.03	**71.20**	78.55	76.38	84.40
MedVidBench	Unmasked	74.66	74.18	76.00	68.11	68.47	86.00	78.59	80.73	82.40	77.01	79.46	79.60
	Binary mask	89.47	90.36	**58.80**	92.99	93.23	**34.00**	**83.00**	**84.40**	**73.60**	**81.85**	83.53	78.40
	Soft mask	**89.81**	**90.64**	59.20	**93.10**	**93.31**	**34.00**	82.87	84.30	74.40	81.72	**83.59**	**77.20**
HAM10000	Unmasked	70.80	61.21	63.20	87.67	87.95	51.60	68.38	62.42	78.40	86.23	88.37	**67.60**
	Binary mask	79.52	72.17	**55.20**	**96.43**	**96.90**	19.60	**70.41**	**63.77**	**74.40**	**88.66**	**90.32**	68.80
	Soft mask	**80.01**	**73.22**	**55.20**	**96.43**	**96.90**	**19.20**	70.27	63.74	**74.40**	88.51	90.22	69.20
WikiArt	Unmasked	63.10	55.82	**65.20**	69.80	65.03	82.40	57.78	51.05	82.00	63.88	63.16	89.20
	Binary mask	66.22	59.83	66.80	80.75	76.70	68.80	**59.11**	**51.66**	**78.40**	**66.78**	**64.97**	**88.00**
	Soft mask	**67.21**	**61.13**	66.80	**81.00**	**76.73**	**67.60**	58.91	51.54	**78.40**	66.32	64.52	**88.00**

Table C.4: Class-level **video retrieval mAP (%)** on UCF101 and HMDB51, without and with the masks learned from Mochi 1 and CogVideoX generated datasets. *Sp.* denotes mask sparsity in %. Best results are **bolded** and second-best results are underlined per generator.

			Mochi 1			CogVideoX1.5-5B		
Dataset	**Encoder**	**Unmasked**	Sp. (%)	Binary	Soft	Sp. (%)	Binary	Soft
UCF101	V-JEPA 2.1 Large	40.82	72.2	42.84	**43.48**	75.4	44.02	**44.42**
	VideoMAEv2-Base	95.52	52.0	95.56	**95.62**	57.8	95.55	**95.61**
HMDB51	V-JEPA 2.1 Large	17.84	72.2	18.99	**19.59**	75.4	19.91	**20.19**
	VideoMAEv2-Base	**46.94**	52.0	46.59	46.88	57.8	46.51	46.74

substantially more effective than the corresponding unmasked representations in the majority of settings.

Retrieval. We extend video retrieval experiments on UCF101 and HMDB51 with the two new video generators. Results in Table C.4 are consistent with those reported in Table 3 for Wan2.1, with actually stronger performances given by masks learned on CogVideoX videos (+3.6 and +2.3 mAP for V-JEPA features).

Overall robustness results. Together with the Wan2.1 results in the main paper, these experiments show that the effectiveness of our learned feature subspaces does not depend on using a particular T2V model to construct the synthetic training dataset. Wan2.1, CogVideoX1.5-5B, and Mochi 1 differ in architecture, training data and objective, yet the same general-versus-specialized consistency signal produces masks that transfer to downstream semantic-similarity and OOD-detection tasks.

C.5 Extending global masks to dense local tasks

While our mask is optimized using global representations, we evaluate its transferability to dense, patch-level tasks where similarity is computed at a finer granularity. Following Mur-Labadia et al. (2026), we assess performance on Video Object Segmentation (VOS) using the DAVIS-2017 dataset (Pont-Tuset et al., 2017). In this task, an initial ground-truth object mask must be propagated across subsequent frames. We adopt the non-parametric label propagation approach detailed in Mur-Labadia et al. (2026), which matches local patch features between frames via cosine similarity in the V-JEPA 2.1 embedding space. To ensure a fair comparison, we maintain the hyperparameters reported in the original study.

Table C.5: **OOD detection using masks learned from Mochi 1.** Video generation OOD detection across four datasets using Cosmos-Predict2, with BDD100k as the in-distribution (ID) baseline. The soft masks are learned exclusively from synthetic videos generated by Mochi 1, while the corresponding binary masks are obtained by post-hoc thresholding at $\theta = 0.5$. The improvements across diverse OOD domains provide additional evidence that the proposed learning signal transfers across source video generators. Best results are **bolded** and second-best results are underlined.

OOD Dataset	Method	V-JEPA2.1-L						VideoMAEv2-B					
		2XDM			DDPM-OOD			2XDM			DDPM-OOD		
		AUROC↑	AUPR↑	FPR@95↓	AUROC↑	AUPR↑	FPR@95↓	AUROC↑	AUPR↑	FPR@95↓	AUROC↑	AUPR↑	FPR@95↓
UVEB	Unmasked	78.42	73.77	57.20	76.37	70.31	63.60	71.08	68.50	77.60	75.50	73.33	**83.20**
	Binary mask	81.84	78.98	54.00	84.85	81.37	51.20	**72.32**	**69.96**	75.20	**78.01**	**76.01**	**83.20**
	Soft mask	**82.61**	**80.33**	**52.80**	**85.99**	**83.20**	**48.80**	72.19	69.92	**74.80**	77.66	75.76	83.60
MedVidBench	Unmasked	74.66	74.18	76.00	68.11	68.47	86.00	78.59	80.73	82.40	77.01	79.46	**79.60**
	Binary mask	85.93	86.71	64.40	88.74	88.54	55.60	**80.69**	**82.01**	82.00	**79.97**	**82.25**	80.80
	Soft mask	**86.74**	**87.56**	**62.80**	**90.12**	**90.26**	**52.00**	80.45	81.85	**80.40**	79.54	81.97	80.00
HAM10000	Unmasked	70.80	61.21	63.20	87.67	87.95	51.60	68.38	62.42	78.40	86.23	88.37	67.60
	Binary mask	76.42	68.06	60.40	93.72	94.16	32.00	**71.00**	**64.98**	76.40	**89.46**	**91.01**	**60.40**
	Soft mask	**77.18**	**69.06**	**58.40**	**94.32**	**94.79**	**30.00**	70.79	64.80	**76.00**	89.00	90.63	62.00
WikiArt	Unmasked	63.10	55.82	65.20	69.80	65.03	82.40	57.78	51.05	**82.00**	63.88	63.16	89.20
	Binary mask	66.87	60.39	**64.40**	76.14	70.05	72.00	**58.45**	**51.34**	82.40	**65.88**	**64.35**	**86.80**
	Soft mask	**67.20**	**60.77**	64.80	**77.39**	**71.96**	**69.20**	57.99	51.08	82.80	64.95	63.46	88.40

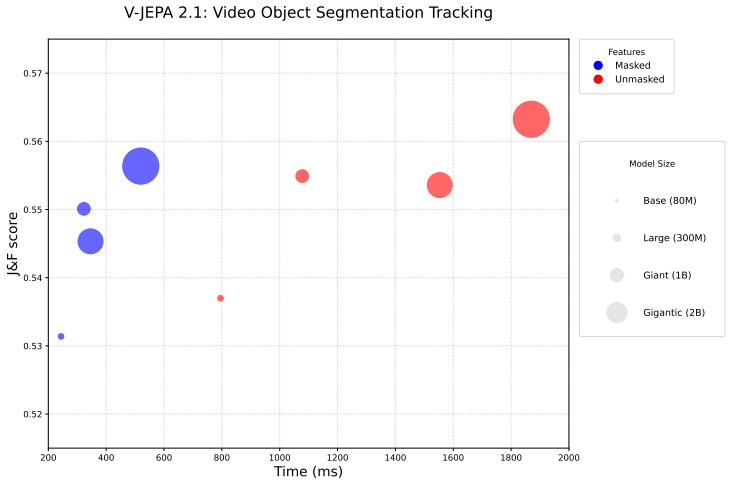


Figure C.3: **Patch-level effectiveness of masked features.** VOS on DAVIS-2017 (Pont-Tuset et al., 2017) via patch features matching.

Efficient dense tracking with masked features. We evaluate all model variants using the full feature set as a baseline, and then repeat the experiments by applying our learned mask to prune the feature dimensions. As illustrated in Fig. C.3, our masked subspace maintains competitive segmentation performance while providing a significant computational advantage. Specifically, by reducing the dimensionality of the search space, our approach accelerates the label propagation algorithm by $\approx$**4**$\times$, demonstrating that the semantic features identified at the global level remain highly effective for localized tracking.

For our experiments, we use V-JEPA 2.1 with all model sizes and follow the best hyperparameters reported by Mur-Labadia et al. (2026), i.e., a context window of 15 frames, a spatial search radius of 12, $k = 5$ nearest neighbors, and a softmax temperature of 0.2. Reported time refers to the inference performed on a single NVIDIA H100 GPU.

C.6 Full details on reconstruction via patch clustering

Following prior work on analyzing the spatial structure encoded in visual representations (Dhanachandra et al., 2015), we assess whether our masked features preserve local visual information. For each

video clip in the dataset, we extract a single RGB frame $I \in \mathbb{R}^{3\times H\times W}$ and compute patch-level feature tokens $z_i \in \mathbb{R}^d$ using the frozen V-JEPA 2.1 encoder with our learned mask applied. We then perform K-means clustering on the set of patch tokens $\{z_i\}$ in the embedding space, partitioning them into k clusters. By assigning each patch token's cluster label back to its original spatial location, we obtain a piecewise-constant segmentation of the image in which each pixel p is associated with a discrete cluster label $\ell(p) \in \{1, \ldots, k\}$.

We then generate a mosaic reconstruction $\hat{I}$ of the original frame by replacing the RGB value of each pixel with the mean color vector of all pixels sharing the same cluster assignment. The reconstruction error for a single frame is quantified by computing the Mean Squared Error (MSE) between the original frame I and the reconstruction $\hat{I}$, averaged over all spatial locations and color channels:

$$\mathrm{MSE}_k(I) = \frac{1}{3HW} \sum_{c,h,w} (I_{c,h,w} - \hat{I}_{c,h,w})^2 \tag{C.1}$$

To qualitatively assess this preservation, Figure C.4 visualizes the mosaic reconstructions across varying cluster counts (k), demonstrating that the spatial and color regions produced by the masked representations closely mirror the unmasked baseline. Quantitatively, Figure 5 reports the final MSE_k averaged across all evaluated frames. We compare the unmasked features against our learned mask at varying binarization thresholds ($\theta \in \{0.5, 0.7, 0.9\}$). The average reconstruction MSE for the masked features remains remarkably close to the full baseline across all evaluated values of k and θ, confirming that the heavily pruned subspaces retain the core visual features.

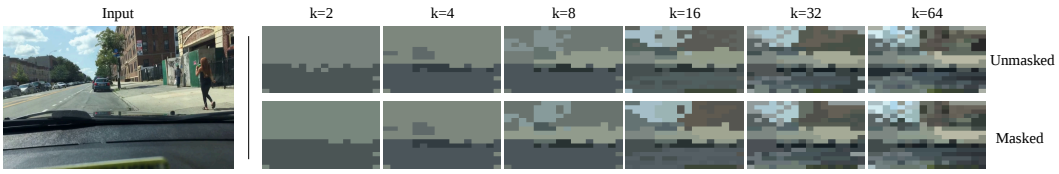


Figure C.4: **Qualitative visualization of K-means mosaic reconstructions.** The original input frame (left) is segmented based on patch-level cluster assignments for $k \in \{2, 4, 8, 16, 32, 64\}$. The visual consistency between the unmasked baseline (top row) and our masked representations (bottom row) illustrates that the masked feature space preserves essential local structural and color information.

C.7 Broader impacts

This research work proposes a lightweight model-agnostic method to learn a semantic similarity score on top of pretrained feature encoders. Such similarity shows improvements in alignment with human judgment and, as shown with experiments over various tasks, can be used for a variety of real-world use cases that rely on the concept of similarity between two videos. Additionally, it achieves such results with higher efficiency than baseline methods by effectively reducing the representation dimensions, and thus the storage required by the extracted features. Negative impacts could result from the limitations and biases inherited from the models used for data synthesis and feature encoding.

D Limitations

While our framework demonstrates the utility of generative uncertainty for representation adaptation, several limitations remain. First, our pipeline is inherently coupled with the performance and inductive biases of the underlying pre-trained components, including Gemini for prompt synthesis, video generative models for video generation, and the foundation models used as encoders. Any systematic biases or domain-specific shortcomings in these models, such as cultural biases in text-to-video synthesis, may propagate into the learned similarity masks. Second, our optimization objective prioritizes global semantic consistency. While this yields highly compact and efficient representations, we observed a marginal performance trade-off in tasks requiring dense, fine-grained local precision (e.g., video object segmentation). Future work could explore multi-scale masking strategies to better bridge this gap.

E LICENSES

List of licenses for existing assets used in this paper.

Models:

- Wan2.1 (T2V-1.3B): Apache License Version 2.0
- CogVideoX1.5-5B: CogVideoX License
- Mochi 1: Apache License Version 2.0
- V-JEPA 2.1: MIT license
- VideoMAEv2: CC-BY-NC-4.0 license
- Cosmos-predict2: Apache License Version 2.0

Data:

- VBench: Apache License Version 2.0
- BDD100k: BSD 3-Clause License
- UVEB: MIT License
- MedVidBench: CC-BY-NC-SA-4.0
- HAM10000: CC-BY-NC-SA-4.0
- WikiArt: CC0: Public Domain
- DAVIS-2017: BSD 3-Clause License
- UCF-101: CC0: Public Domain
- HMDB51: CC BY 4.0