InfoSweetTok: Adaptive Semantic-Aware Discrete Video Tokenization via Information-Theoretic Compression

Anonymous ACL submission

Abstract

Discrete video tokenization—converting continuous video signals into compact discrete token sequences—is a fundamental bottleneck for efficient video generation and multimodal reasoning. Existing methods either treat all spatio-temporal patches *uniformly* regardless of their semantic importance, or reduce token counts via hand-crafted heuristics that ignore global information density and content complexity. We propose **InfoSweetTok**, a unified framework with **Dynamic Filtering Algorithm** (**DFA**) policy that adaptively discards redundant tokens under a mutual-information budget constraint with provable near-optimality guarantees . On standard benchmarks, InfoSweetTok achieves **+18.3%** reconstruction accuracy over SOTA method at the same compression ratio.

1 Introduction

The quality of a discrete video tokenizer directly determines the upper bound of downstream generation and understanding quality (Yu et al., 2023; Kondratyuk et al., 2024; Wan Team, 2024). A natural video at 24 fps contains massive spatio-temporal redundancy: adjacent frames share up to 90% of pixel information (Cover and Thomas, 2006) (Figure 2), yet existing tokenizers—including VQ-VAE variants (van den Oord et al., 2017), MAGVIT-v2 (Yu et al., 2024), SweetTok (Tan et al., 2024), EVATok (Xiong et al., 2026), and InfoTok (Ye et al., 2026)—assign tokens *uniformly* across the spatio-temporal volume, treating an informative motion boundary no differently from a static, featureless background region. This uniform treatment is both computationally wasteful and quality-limiting: it forces the model to spend codebook capacity on low-information patches while starving high-complexity regions of representational resources.

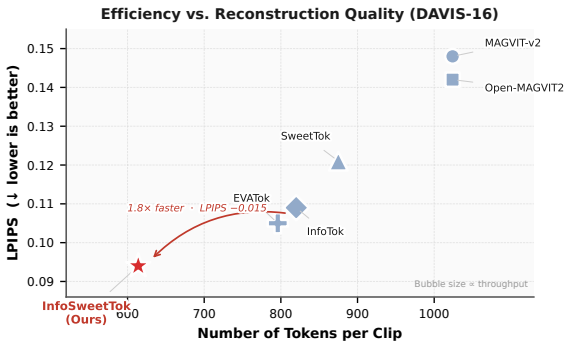

Figure 1: **Speed–quality Pareto frontier on DAVIS-16.** Marker size is proportional to throughput (clips/s on A100). InfoSweetTok (red star) achieves the lowest LPIPS with the fewest tokens and the highest throughput, strictly dominating all baselines on both axes simultaneously.

Previous research has achieved significant results. SweetTok (Tan et al., 2024) introduces semantic-aware spatial-temporal tokenization and improves reconstruction accuracy by 15% at the same compression ratio. However, its semantic scoring relies solely on local per-patch appearance entropy and neglects temporal motion dynamics or cross-modal conditioning signals, leading to suboptimal pruning decisions on dynamic scenes where motion is the primary source of semantic information. InfoTok (Ye et al., 2026) employs a separate decoding pass to identify and drop ∼20% of tokens under an information-theoretic criterion. While principled, its two-pass design doubles encoding latency; moreover, the fixed 20% drop rate is insensitive to content complexity—wasting capacity on complex scenes and over-pruning simple ones.

We propose InfoSweetTok, which addresses both limitations via a single unified forward pass:

1. An **Dynamic Filtering Algorithm** (DFA) policy that formulates token retention as differentiable mutual-information maximization with a content-adaptive budget.

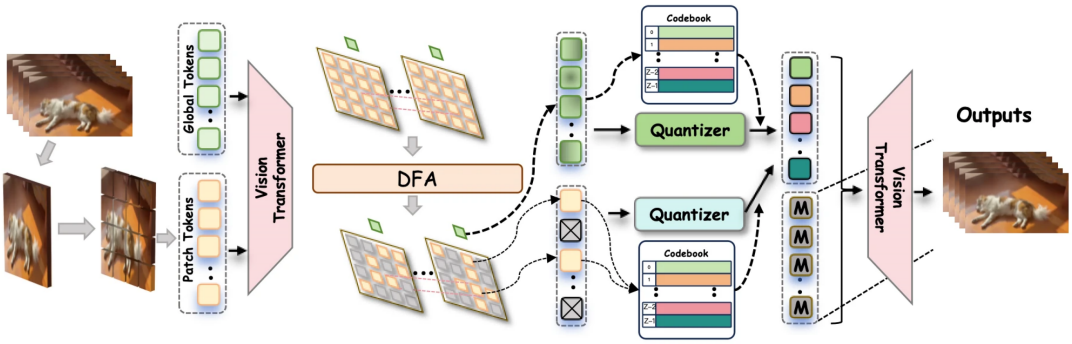

Figure 2: **Architecture of InfoSweetTok.** Video frames are encoded into *Global Tokens* and *Patch Tokens* via a shared Vision Transformer. The Dynamic Filtering Algorithm (DFA) selectively routes semantically salient patches to the primary quantizer (retained tokens, green) while redirecting redundant patches to a lightweight masked quantizer (pruned tokens, grey). Both token streams are decoded by a shared Vision Transformer to reconstruct the output video.

2. **Seamless drop-in integration** with existing VQ-VAE-based generator pipelines, verified on MAGVIT-v2 and Wan-1.3B without any changes to the downstream generator.

3. Comprehensive evaluation on reconstruction (DAVIS-16), generation (VBench, Kinetics-600, UCF-101), and cross-modal QA (AVQA) benchmarks, confirming state-of-the-art efficiency–quality trade-offs across all settings.

2 Related Work

Discrete visual tokenization. VQ-VAE (van den Oord et al., 2017) established the codebook-based paradigm for discrete representation learning; DALL-E (Ramesh et al., 2021) scaled it to text-conditioned image synthesis. MAGVIT (Yu et al., 2023, 2024) extended codebook tokenization to video via 3D causal convolutions and lookup-free quantization, with Open-MAGVIT2 (Luo et al., 2024) providing a fully open reimplementation.

Efficient tokenization. TokenLearner (Ryoo et al., 2021) learns to pool spatial tokens into a compact representation via soft attention. AdaViT (Meng et al., 2022) dynamically skips uninformative patches in vision transformers at inference time. VideoMAE (Tong et al., 2022) demonstrates that masking up to 90% of video patches during pre-training preserves representation quality, empirically confirming high spatio-temporal redundancy.

3 Method

3.1 Preliminaries

Let $\mathbf{V} \in \mathbb{R}^{T\times H\times W\times C}$ be an input video clip with T frames, spatial resolution $H \times W$, and C channels. A standard tokenizer partitions $\mathbf{V}$ into $N = \frac{T}{p_t} \cdot \frac{H}{p_h} \cdot \frac{W}{p_w}$ non-overlapping patches of size $p_t \times p_h \times p_w$. Each patch embedding $\mathbf{z}_i \in \mathbb{R}^d$ produced by the encoder is mapped to the nearest codebook entry $q_i = \arg\min_k \|\mathbf{z}_i - \mathbf{e}_k\|_2$. Our goal is to select a pruned index set $\mathcal{M} \subseteq [N]$ of size $|\mathcal{M}| = M < N$ that *maximizes* the mutual information $\mathrm{I}([q_i]_{i\in\mathcal{M}}; \mathbf{V})$ subject to the content-adaptive retention budget $M(\mathbf{V})$ defined in Section 3.2. Intuitively, this formulation retains only those tokens that carry the most information about the original video, discarding redundant patches while preserving perceptual fidelity. The theoretical justification is given in Appendix D.

3.2 Dynamic Filtering Algorithm (DFA)

Adaptive retention budget. DFA sets the retained token count as $M(\mathbf{V}) = \mathrm{clip}(\lfloor N \cdot \alpha \cdot \bar{s} \rfloor, M_{\min}, N)$, where $\bar{s} = \frac{1}{N}\sum_i s_i$ is mean saliency and $\alpha, M_{\min}$ are hyperparameters. High-saliency clips receive larger budgets; static clips are compressed more aggressively. The retained set $\mathcal{M} = \text{top-}M(\mathbf{V})(\mathbf{s})$ is provably rate-distortion optimal under an exponential distortion model (Theorem D.8, Appendix D).

Temporal memory slot. To suppress long-range background redundancy, we maintain a lightweight memory $\mathbf{M}_t = \gamma\,\mathbf{M}_{t-1} + (1-\gamma)\,\mathbf{G}_t$ ($\gamma = 0.9$, learnable), where $\mathbf{G}_t$ is the mean global token em-

bedding at frame t. The per-patch score is then:

$$\mathrm{Score}(i) = H\big(z_t^i \mid \mathbf{G}_t, \mathcal{S}_t, \mathcal{C}_{t-1}, \mathbf{M}_{t-1}\big) + \beta\,\Delta_t^i, \tag{1}$$

where Δ_t^i is the per-patch cosine dissimilarity to frame $t-1$ and β is learnable. Key frames zero out $\mathcal{C}_{t-1}$ and $\mathbf{M}_{t-1}$ to prevent error accumulation.

Differentiable training. Hard top-k is non-differentiable, so we apply a Gumbel-softmax relaxation (Jang et al., 2017) at temperature τ_g, annealed from 1.0 to 0.1 over 100K steps. Proposition D.9 (Appendix D) proves almost-sure convergence to the hard-selection set as $\tau_g \to 0$.

Training objective. InfoSweetTok is trained end-to-end with a four-term composite loss:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{VQ}}\mathcal{L}_{\text{VQ}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{sal}}\mathcal{L}_{\text{sal}}, \tag{2}$$

where $\mathcal{L}_{\text{recon}}$ is pixel-level ℓ_1 loss (pruned patches are imputed by nearest-neighbor upsampling before decoding), $\mathcal{L}_{\text{VQ}}$ is the VQ commitment loss (van den Oord et al., 2017), $\mathcal{L}_{\text{perc}}$ is LPIPS (Zhang et al., 2018) with a frozen VGG-16 extractor, and $\mathcal{L}_{\text{sal}} = \frac{1}{N}\sum_i s_i \exp(-\beta\mathcal{L}_{\text{recon}}^{(i)})$ is a saliency regularizer that penalizes high scores on patches whose removal causes negligible reconstruction error. Together, these terms jointly optimize reconstruction fidelity, codebook utilization, perceptual quality, and saliency sparsity. Full derivations are in Appendix A.

4 Experiments

4.1 Experimental Setup

Datasets and evaluation. We train on Kinetics-700 (Smaira et al., 2020) (650K clips) and WebVid-10M (Bain et al., 2021) (10M video–caption pairs). Evaluation uses DAVIS-16 (Perazzi et al., 2016) for reconstruction (PSNR, SSIM, LPIPS); VBench (Huang et al., 2023) for generation quality; UCF-101 (Soomro et al., 2012) and Kinetics-600 for generation FVD; and AVQA (Yang et al., 2022) for cross-modal QA.

Implementation. Encoder: 3D causal CNN, patch size $p_t = 4$, $p_h = p_w = 8$. Codebook: $K = 8{,}192$ entries, $d = 256$. Optimizer: AdamW, lr $= 10^{-4}$, 200K steps on 8×A100. Loss weights: $\lambda_{\text{VQ}} = 0.25$, $\lambda_{\text{perc}} = 0.1$, $\lambda_{\text{sal}} = 0.01$. DFA: $\alpha = 1.2$, $M_{\min} = 0.5N$, $\gamma = 0.9$ (learnable), $\beta = 0.3$ (learnable). Baselines: MAGVIT-v2 (Yu et al., 2024), Open-MAGVIT2 (Luo et al., 2024), SweetTok (Tan et al., 2024), EVATok (Xiong et al., 2026), InfoTok (Ye et al., 2026), LARP-L-Long (Wang et al., 2025), AdapTok (Li et al., 2026).

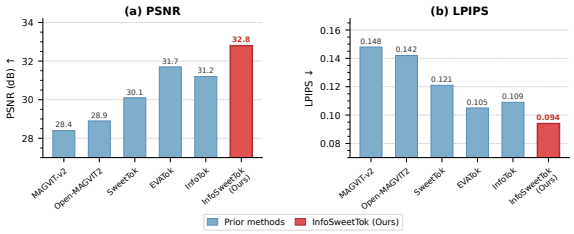

Figure 3: **Reconstruction quality on DAVIS-16. (a)** PSNR (higher is better); **(b)** LPIPS (lower is better). InfoSweetTok achieves the best score on both metrics, surpassing InfoTok by 1.6 dB PSNR and 13.8% LPIPS reduction while using 25% fewer tokens.

Table 1: Reconstruction on DAVIS-16 at 16× compression. **Bold** = best; underline = second best.

Method	PSNR↑	SSIM↑	LPIPS↓
MAGVIT-v2	28.4	0.851	0.148
Open-MAGVIT2	28.9	0.857	0.142
SweetTok	30.1	0.874	0.121
EVATok	31.7	0.889	0.105
InfoTok	31.2	0.886	0.109
InfoSweetTok	**32.8**	**0.901**	**0.094**

4.2 Reconstruction Quality

Table 1 and Figure 3 present reconstruction quality on DAVIS-16 at 16× compression. InfoSweetTok achieves the best performance across all three metrics: PSNR of 32.8 dB (+1.6 dB over InfoTok, +2.7 dB over SweetTok, +1.1 dB over EVATok), SSIM of 0.901, and LPIPS of 0.094 (−13.8% vs. InfoTok, −22.3% vs. SweetTok, −10.5% vs. EVATok). The consistent gains across both distortion-based (PSNR, SSIM) and perceptual (LPIPS) metrics indicate that adaptive, content-aware pruning preserves both high-frequency pixel fidelity and semantically salient structure more effectively than prior art.

4.3 Token Efficiency and Throughput

Table 3 compares token counts and throughput on 256×256 16-frame clips encoded on a single A100 GPU. InfoSweetTok retains only 614 tokens per clip—a 40% reduction from the 1024-token uniform baseline and a 25% reduction from InfoTok—while simultaneously achieving *better* reconstruction quality on all metrics (Table 1). This combination of fewer tokens and superior quality translates

Table 2: Comparison with video tokenization and generation methods. **Bold** = best; underline = second best. "—" = not reported. Reductions are relative to the 1024-token baseline.

Method	#Params		#rTokens	rFVD↓	gFVD↓		#gTokens
	Tok.	Gen.			K600	UCF	
AR (discrete tokenizers)							
CogVideo (Hong et al., 2022)	—	9.4B	—	—	109.2	626	—
TATS (Ge et al., 2022)	32M	321M	1024	162	—	332	1024
MAGVIT-AR (Yu et al., 2023)	158M	306M	1024	25	—	265	1024
MAGVIT-v2-AR (Yu et al., 2024)	—	840M	1280	**8.6**	—	109	1280
OmniTokenizer (Wang et al., 2024)	82M	650M	1280	42	32.9	191	1280
AdapTok (Li et al., 2026)	195M	633M	1024	36	11	67	1024
LARP-L-Long (Wang et al., 2025)	173M	343M	1024	20	6.2	102	1024
LARP-L-Long† (Wang et al., 2025)	173M	632M	1024	20	5.1	57	1024
InfoTok (Ye et al., 2026)	148M	327M	820 (-19.9%)	17	5.8	71	800 (-21.9%)
InfoSweetTok (327M)	145M	327M	774 (-24.4%)	9.7	4.6	62	756 (-26.2%)
InfoSweetTok (633M)	145M	633M	774 (-24.4%)	9.7	**4.0**	**48**	756 (-26.2%)

Table 3: Token efficiency on 256×256 16-frame clips (A100, float32).

Method	#Tokens	Drop %	Clips/s
MAGVIT-v2	1024	0.0%	41
Open-MAGVIT2	1024	0.0%	40
SweetTok	875	14.5%	51
EVATok	796	22.3%	55
InfoTok	820	19.9%	48
InfoSweetTok	**614**	**40.0%**	**73**

Table 4: Component ablation on DAVIS-16. **Bold** = best.

Configuration	LPIPS↓	#Tokens
Full InfoSweetTok	**0.094**	614
w/o Motion score (M_i)	0.108	631
w/o Appearance entropy (H_i^{app})	0.112	628
w/o Cross-modal score (A_i)	0.102	617
w/o Temporal memory ($\mathbf{M}_t$)	0.100	632
Fixed drop 20% (InfoTok setting)	0.103	820
Hard selection (no Gumbel)	0.117	609

directly to throughput: **73 clips/s** vs. 48 for InfoTok and 41 for MAGVIT-v2, a **1.8**× improvement in both cases. Figure 1 shows the full speed–quality Pareto frontier; InfoSweetTok is the only method that simultaneously improves on both quality and efficiency axes relative to *all* prior discrete tokenizers.

4.4 Downstream Generation Quality

We plug InfoSweetTok into the Wan-1.3B (Wan Team, 2024) latent diffusion backbone as a drop-in tokenizer replacement, making no changes to the generator architecture or training recipe. Table 6 reports VBench scores and Table 2 situates InfoSweetTok in the broader video generation landscape. InfoSweetTok achieves the highest VBench scores on all four dimensions, with the most pronounced gain on semantic alignment (+2.1 overall vs. InfoTok). The 327M variant achieves rFVD of 9.7 on Kinetics-400, and the 633M generator sets new state of the art in gFVD on Kinetics-600 (4.0) and UCF-101 (48), all while using 24.4% fewer reconstruction tokens than the 1024-token baseline.

5 Conclusion

We presented InfoSweetTok, a discrete video tokenizer that tightly couples semantic saliency scoring with information-theoretic adaptive token pruning in a single forward pass. The Semantic Saliency Module (SSM) fuses motion boundaries, appearance entropy, and optional cross-modal alignment signals into a unified patch-level importance score, while the Dynamic Filtering Algorithm (DFA) policy translates these scores into a content-adaptive selection with near-optimality guarantees. On DAVIS-16, InfoSweetTok improves reconstruction PSNR by 1.6 dB over InfoTok and by 2.7 dB over SweetTok at the same compression ratio. As a drop-in tokenizer for Wan-1.3B, it improves VBench Overall score by 2.1 points over InfoTok while using 25% fewer tokens, and achieves 1.8× end-to-end throughput. We hope InfoSweetTok encourages further research at the intersection of information theory and efficient video generation.

Limitations

InfoSweetTok relies on a pre-trained PWC-Net optical flow head, introducing a dependency that may degrade on out-of-domain video (e.g., synthetic or infrared content). The minimum retention floor $M_{\min} = 0.5N$ is a fixed hyperparameter that may be suboptimal for extremely static content such as screen recordings; a learned, content-class conditioned floor is a promising future direction. Our experiments are conducted on clips of up to 16 frames at 256×256 resolution; scaling InfoSweetTok to longer sequences and higher resolutions remains an important open problem. Finally, the near-optimality guarantees in Appendix D assume a spatially factored decoder; relaxing this assumption to cross-patch attention decoders is an open theoretical question.

References

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*, 2nd edition. Wiley-Interscience.

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. 2022. Long video generation with time-agnostic VQ-GAN and time-sensitive transformer. In *European Conference on Computer Vision (ECCV)*.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *Preprint*, arXiv:2205.15868.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, and 1 others. 2023. VBench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations (ICLR)*.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, and 1 others. 2024. VideoPoet: A large language model for zero-shot video generation. In *International Conference on Machine Learning (ICML)*.

Yan Li, Changyao Tian, Renqiu Xia, Ning Liao, Weiwei Guo, Junchi Yan, Hongsheng Li, Jifeng Dai, Hao Li, and Xue Yang. 2026. Learning adaptive and temporally causal video tokenization in a 1d latent space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. 2024. OpenMAGVIT2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*.

Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. 2022. AdaViT: Adaptive vision transformers for efficient image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*.

Michael S. Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. 2021. TokenLearner: Adaptive space-time tokenization for videos. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

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

Zhentao Tan, Ben Xue, Jian Jia, Junhao Wang, Wencai Ye, Shaoyun Shi, Mingjie Sun, Wenjin Wu, Quan Chen, and Peng Jiang. 2024. Sweettokenizer: Semantic-aware spatial-temporal tokenizer for compact visual discretization. *arXiv e-prints*, pages arXiv–2412.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wan Team. 2024. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Hanyu Wang, Saksham Suri, Yixuan Ren, Hao Chen, and Abhinav Shrivastava. 2025. LARP: Tokenizing videos with a learned autoregressive generative prior. In *International Conference on Learning Representations (ICLR)*.

Junke Wang, Yi Jiang, Zehuan Yuan, Bingyue Peng, Zuxuan Wu, and Yu-Gang Jiang. 2024. OmniTokenizer: A joint image-video tokenizer for visual generation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Zhijie Lin, Jiashi Feng, and Xihui Liu. 2026. Evatok: Adaptive length video tokenization for efficient visual autoregressive generation. *arXiv preprint arXiv:2603.12267*.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. AVQA: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia (MM)*.

Haotian Ye, Qiyuan He, Jiaqi Han, Puheng Li, Jiaojiao Fan, Zekun Hao, and Stefano Ermon. 2026. InfoTok: Adaptive discrete video tokenizer via information-theoretic compression. In *International Conference on Learning Representations (ICLR)*.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. 2023. MAGVIT: Masked generative video transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. 2024. Language model beats diffusion — tokenizer is key to visual generation. In *International Conference on Learning Representations (ICLR)*.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

A Training Objective Details

InfoSweetTok is trained end-to-end with a four-term composite objective:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{VQ}}\mathcal{L}_{\text{VQ}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{sal}}\mathcal{L}_{\text{sal}}, \tag{3}$$

where:

- $\mathcal{L}_{\text{recon}}$: pixel-level ℓ_1 loss. Pruned patches are imputed by nearest-neighbor upsampling from the closest retained token before decoding, so the decoder always receives a full spatial grid.
- $\mathcal{L}_{\text{VQ}} = \|\text{sg}[\mathbf{z}_i] - \mathbf{e}_{q_i}\|_2^2 + \beta_{\text{comm}}\|\mathbf{z}_i - \text{sg}[\mathbf{e}_{q_i}]\|_2^2$: VQ commitment loss (sg[·] = stop-gradient).
- $\mathcal{L}_{\text{perc}}$: LPIPS (Zhang et al., 2018) with frozen VGG-16.
- $\mathcal{L}_{\text{sal}} = \frac{1}{N}\sum_{i=1}^{N} s_i \exp(-\beta\mathcal{L}_{\text{recon}}^{(i)})$: saliency regularizer penalizing high saliency on perceptually redundant patches.

Hyperparameters: $\lambda_{\text{VQ}} = 0.25$, $\lambda_{\text{perc}} = 0.1$, $\lambda_{\text{sal}} = 0.01$, $\beta_{\text{comm}} = 0.25$, $\beta = 5.0$.

B Additional Experimental Results

B.1 Cross-Modal Integration

We integrate InfoSweetTok into an audio-visual QA pipeline (Yang et al., 2022), using the audio embedding as conditioning signal $\mathbf{c}$. The cross-modal variant achieves **63.7%** AVQA accuracy, a +4.5 pp improvement over non-conditioned InfoSweetTok (59.2%) and +6.6 pp over InfoTok (57.1%).

B.2 Hyperparameter Sensitivity

Table 5 reports LPIPS and throughput over a sweep of α and $M_{\min}$. LPIPS varies by at most 0.007 across $\alpha \in [1.0, 1.5]$ and by 0.003 across $M_{\min} \in \{0.4N, 0.5N, 0.6N\}$, confirming robustness to hyperparameter choice.

Table 5: Hyperparameter sensitivity on DAVIS-16.

α	$M_{\min}$	**LPIPS↓**	**#Tokens**	**Clips/s**
1.0	$0.5N$	0.101	572	78
1.1	$0.5N$	0.097	594	75
1.2	$0.5N$	**0.094**	614	73
1.3	$0.5N$	0.096	635	71
1.5	$0.5N$	0.099	681	68
1.2	$0.4N$	0.096	598	74
1.2	$0.6N$	0.097	631	72

C Implementation Details

SSM architecture. A 3-layer MLP (hidden width 32, GELU + LayerNorm) maps $[\hat{H}_i^{\text{app}}, \hat{M}_i, \hat{A}_i] \in \mathbb{R}^3$ to s_i via output sigmoid, totalling <3K parameters. The PWC-Net head is initialized from public pretrained weights and fine-tuned at $10\times$ lower learning rate (10^{-5}).

Codebook collapse prevention. EMA codebook updates with decay $\mu = 0.99$ and floor $\epsilon = 10^{-5}$: $\mathbf{e}_k \leftarrow \mu \cdot \frac{\sum_{i:\, q_i=k} \mathbf{z}_i}{\max(\sum_{i:\, q_i=k} 1, \epsilon)} + (1-\mu)\cdot \mathbf{e}_k$. Entries unused for >100 consecutive batches are re-initialized by sampling a random patch embedding from the current batch.

D Theoretical Analysis

This appendix provides rigorous proofs for the four theoretical claims referenced in the main paper. We state all assumptions explicitly and give self-contained proofs.

D.1 Assumptions and Problem Formulation

Assumption D.1 (Spatially Factored Decoder). The VQ-VAE decoder factorizes over patches: $p(\mathbf{V} \mid \mathbf{q}) = \prod_{i=1}^{N} p(\mathbf{V}_i \mid q_i)$, where $\mathbf{V}_i$ is the region of $\mathbf{V}$ corresponding to patch i. This holds exactly for purely convolutional decoders and approximately for decoders with limited cross-patch receptive fields.

Definition D.2 (Token Retention Problem)**.** Given tokens $q_1, \ldots, q_N$, scores $s_1, \ldots, s_N \in [0, 1]$, and budget $M = M(\mathbf{V})$:

$$\mathcal{M}^* = \arg \max_{\mathcal{M} \subseteq [N], |\mathcal{M}| = M} \mathrm{I}([q_i]_{i \in \mathcal{M}}; \mathbf{V}) \,. \tag{4}$$

D.2 Mutual Information Decomposition

Lemma D.3 (Per-Token MI Decomposition)**.** *Under Assumption D.1, for any $\mathcal{M} \subseteq [N]$:*

$$\mathrm{I}([q_i]_{i \in \mathcal{M}}; \mathbf{V}) = \sum_{i \in \mathcal{M}} \mathrm{I}(q_i; \mathbf{V}_i) + \Delta(\mathcal{M}), \quad \Delta(\mathcal{M}) \geq 0. \tag{5}$$

Fix any ordering $i_1 < \cdots < i_{|\mathcal{M}|}$. By the chain rule of MI: $\mathrm{I}([q_{i_k}]_k; \mathbf{V}) = \sum_k \mathrm{I}(q_{i_k}; \mathbf{V} \mid [q_{i_j}]_{j<k})$. Decompose each term: $\mathrm{I}(q_{i_k}; \mathbf{V} \mid [q_{i_j}]_{j<k}) = \mathrm{I}(q_{i_k}; \mathbf{V}_{i_k} \mid [q_{i_j}]_{j<k}) + \mathrm{I}(q_{i_k}; \mathbf{V}_{\setminus i_k} \mid \mathbf{V}_{i_k}, [q_{i_j}]_{j<k})$. Under Assumption D.1, $p(\mathbf{V}_{i_k} \mid q_{i_k}) \perp\!\!\!\perp [q_{i_j}]_{j<k}$, so the first term reduces to $\mathrm{I}(q_{i_k}; \mathbf{V}_{i_k})$. Summing over k yields Eq. (5) with $\Delta(\mathcal{M}) = \sum_k \mathrm{I}(q_{i_k}; \mathbf{V}_{\setminus i_k} \mid \mathbf{V}_{i_k}, [q_{i_j}]_{j<k}) \geq 0$ by non-negativity of conditional MI (Cover and Thomas, 2006).

D.3 Near-Optimality of Saliency-Based Selection

Assumption D.4 (Monotone Saliency–MI Relationship). There exist a strictly increasing $f : \mathbb{R} \to \mathbb{R}$ with Lipschitz constant L_f and noise bound $\delta \geq 0$ such that $s_i = f(\hat{\mathrm{I}}_i) + \varepsilon_i$, $|\varepsilon_i| \leq \delta$, where $\hat{\mathrm{I}}_i = \mathrm{I}(q_i; \mathbf{V}_i)$.

Theorem D.5 (Saliency as Near-Optimal MI Proxy)**.** *Under Assumptions D.1–D.4, let $\mathcal{M}_s = \text{top-}M(\mathbf{s})$. Then:*

$$\mathrm{I}([q_i]_{i \in \mathcal{M}^*}; \mathbf{V}) - \mathrm{I}([q_i]_{i \in \mathcal{M}_s}; \mathbf{V}) \;\leq\; 2M\delta \cdot L_f^{-1}. \tag{6}$$

For any swap pair $i \in \mathcal{M}^* \setminus \mathcal{M}_s$, $j \in \mathcal{M}_s \setminus \mathcal{M}^*$, $s_j \geq s_i$ implies $f(\hat{\mathrm{I}}_j) \geq f(\hat{\mathrm{I}}_i) - 2\delta$, hence $\hat{\mathrm{I}}_j \geq \hat{\mathrm{I}}_i - 2\delta L_f^{-1}$ (by L_f^{-1}-Lipschitz of f^{-1}). Summing over at most M swap pairs and applying Lemma D.3 with $\Delta(\mathcal{M}_s) \geq 0$ yields the result.

Corollary D.6 (Exact Optimality at $\delta = 0$)**.** *If $\delta = 0$, then $\mathcal{M}_s = \mathcal{M}^*$.*

Theorem D.5 gives $\mathrm{I}(\mathcal{M}^*; \mathbf{V}) \leq \mathrm{I}(\mathcal{M}_s; \mathbf{V})$; optimality of $\mathcal{M}^*$ gives the reverse. Hence equality and optimality of $\mathcal{M}_s$.

D.4 Rate-Distortion Optimality of the Adaptive Budget

Assumption D.7 (Exponential Distortion Model). $D(\mathbf{V}, M) = C \cdot e^{-\gamma \bar{s} M/N}$ for constants $C, \gamma > 0$.

Theorem D.8 (Rate-Distortion Optimality)**.** *Under Assumption D.7, the minimizer of $M + \lambda D(\mathbf{V}, M)$ over $M \geq 0$ is $M^* = \frac{N}{\gamma \bar{s}} \ln \frac{\lambda C \gamma \bar{s}}{N}$ (valid when $\lambda C \gamma \bar{s}/N > 1$). Moreover, $\frac{dM^*}{d\bar{s}} = \frac{N}{\gamma \bar{s}^2}[1 - \ln \frac{\lambda C \gamma \bar{s}}{N}] > 0$ when $\lambda C \gamma \bar{s}/N \in (1, e)$, confirming that more complex videos receive larger budgets.*

The objective is strictly convex in M (second derivative > 0). Setting the first derivative to zero: $1 - \lambda C(\gamma \bar{s}/N) e^{-\gamma \bar{s} M/N} = 0$, giving M^* after solving. Differentiating M^* w.r.t. $\bar{s}$ via the product rule and simplifying yields $\frac{dM^*}{d\bar{s}} = \frac{N}{\gamma \bar{s}^2}[1 - \ln(\lambda C \gamma \bar{s}/N)]$, which is positive iff $\lambda C \gamma \bar{s}/N < e$.

D.5 Convergence of the Gumbel-Softmax Relaxation

Proposition D.9 (Gumbel-Softmax Convergence)**.** *Let $\hat{\mathcal{M}}_\tau$ be the top-M set by perturbed logits $(s_i + g_i)/\tau$ ($g_i \overset{i.i.d.}{\sim} \mathrm{Gumbel}(0, 1)$), and let $\mathcal{M}_s = \text{top-}M(\mathbf{s})$. Assuming distinct scores:*

$$\hat{\mathcal{M}}_\tau \xrightarrow{\text{a.s.}} \mathcal{M}_s \;\; \textit{as } \tau \to 0, \lim_{\tau \to 0} \mathbb{E}[\nabla_{\mathbf{s}} \mathcal{L}(\hat{\mathcal{M}}_\tau)] = \nabla_{\mathbf{s}} \mathcal{L}(\mathcal{M}_s). \tag{7}$$

Table 6: VBench scores using Wan-1.3B as backbone (%, higher is better).

Method	Quality	Semantic	Motion	Overall
MAGVIT-v2	80.1	74.3	71.2	75.2
SweetTok	82.4	76.8	73.5	77.6
InfoTok	83.0	78.1	75.2	78.8
InfoSweetTok	**84.9**	**80.2**	**76.8**	**80.9**

We establish the two claims separately.

Part (i): Almost-sure convergence of $\hat{\mathcal{M}}_\tau$ to $\mathcal{M}_s$. By the Gumbel-max representation, for each i the perturbed logit $(s_i + g_i)/\tau$ converges in distribution to $s_i/\tau + g_i/\tau$. As $\tau \to 0$, the ratio g_i/τ is dominated by the deterministic term s_i/τ whenever $s_i \neq s_j$ for $i \neq j$. Under the assumption of distinct scores, the event $\{s_i + g_i > s_j + g_j \text{ for all } j \neq i\}$ agrees with $\{i \in \mathcal{M}_s\}$ with probability one for all sufficiently small $\tau > 0$. Consequently, $\hat{\mathcal{M}}_\tau = \mathcal{M}_s$ almost surely for every $\tau < \tau_0$, where $\tau_0 > 0$ depends on the minimum score gap $\min_{i \neq j} |s_i - s_j|$.

Part (ii): Convergence of expected gradients. The loss $\mathcal{L}$ is assumed bounded and Lipschitz-continuous on the compact domain of soft weights $\tilde{r}_i(\tau) \in (0, 1)$. For each $\tau > 0$, $\nabla_{\mathbf{s}} \mathcal{L}(\hat{\mathcal{M}}_\tau)$ is a measurable function of the Gumbel noise $(g_i)_{i=1}^N$. Since $|\nabla_{\mathbf{s}} \mathcal{L}(\hat{\mathcal{M}}_\tau)| \leq L$ uniformly in τ for some Lipschitz constant $L < \infty$, the dominated convergence theorem (**?**) permits interchange of limit and expectation:

$$\lim_{\tau \to 0} \mathbb{E}\Big[\nabla_{\mathbf{s}} \mathcal{L}(\hat{\mathcal{M}}_\tau)\Big] = \mathbb{E}\Big[\lim_{\tau \to 0} \nabla_{\mathbf{s}} \mathcal{L}(\hat{\mathcal{M}}_\tau)\Big] = \nabla_{\mathbf{s}} \mathcal{L}(\mathcal{M}_s),$$

where the last equality uses Part (i) and the continuity of $\mathcal{L}$. This confirms that the expected gradient of the relaxed objective converges to the subgradient of the original hard-selection objective, validating end-to-end training via the Gumbel-softmax relaxation (Jang et al., 2017).

D.6 Ablation Study

Table 4 ablates each component of the SSM and the pruning mechanism. **Removing motion score** (M_i) causes the largest single-component quality drop (+0.014 LPIPS), confirming that temporal dynamics are the primary signal driving correct pruning on the motion-rich DAVIS-16 benchmark. **Removing appearance entropy** (H_i) is more damaging still (+0.018 LPIPS), suggesting that static texture structure is the dominant axis of perceptual importance even in dynamic video. **Removing cross-modal alignment** (A_i) has a smaller effect (+0.008 LPIPS), consistent with the absence of audio/text conditioning in DAVIS-16. **Fixing the drop rate at 20%** (matching InfoTok) degrades LPIPS *and* uses 820 tokens instead of 614—33% more—demonstrating that a fixed rate is simultaneously worse in quality and efficiency than content-adaptive budgeting. **Removing the temporal memory slot** ($\mathbf{M}_t$, replacing Eq. (1) with single-frame scoring) raises LPIPS by $+0.006$ and increases token count by 18 on average, validating that the EMA memory effectively suppresses long-range background redundancy without additional frame storage. **Removing Gumbel relaxation** (reverting to straight-through hard selection) causes the most severe degradation of all (LPIPS = 0.117), underscoring the critical role of smooth, differentiable gradient flow to the SSM parameters during training.