CODECTOK: CODEC-INFORMED ADAPTIVE TOKEN ALLOCATION FOR EFFICIENT VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

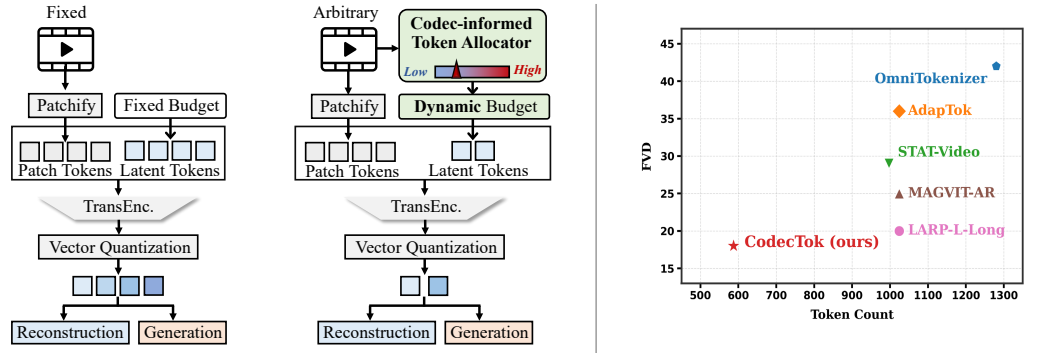
Discrete video tokenization is a central efficiency bottleneck in autoregressive video generation, as each clip must be compressed into a token sequence whose length strongly affects downstream decoding cost. Fixed-budget video tokenizers allocate the same number of tokens to every clip, which can use capacity inefficiently across videos with different spatiotemporal complexity. We draw inspiration from rate control in classical video codecs: inexpensive measures of spatial and temporal complexity can guide content-adaptive token allocation. Based on this view, we introduce CodecTok, a codec-informed adaptive video tokenizer that adapts the idea of content-dependent capacity allocation to discrete video representations. CodecTok uses a training-free allocator to estimate spatial complexity from block-wise gradient statistics and temporal complexity from Hadamard-domain inter-frame residuals, producing a content-adaptive token budget for each clip without training an auxiliary allocation network or performing per-clip budget search. To support variable budgets in practice, CodecTok further employs a padding-free flexible backbone with sequence packing and 3D rotary positional encoding, enabling arbitrary resolutions and variable-length token sequences. A semantic alignment objective distills structured representations from a frozen video foundation model, encouraging highly compressed tokens to preserve meaningful spatiotemporal semantics. Experiments on UCF-101 demonstrate competitive reconstruction and downstream AR generation at reduced token budgets. With 743 reconstruction tokens and 730 generated tokens on average, CodecTok achieves an rFVD of 16 and a gFVD of 44, supporting a favorable balance between visual quality and token efficiency. The code and model will be released.

1 INTRODUCTION

Driven by advances in Large Language Models (LLMs) and their potential to build unified multimodal systems (Wu et al., 2024; Team, 2024), autoregressive (AR) visual generation (Van Den Oord et al., 2017; Sun et al., 2024; Esser et al., 2021; Zhao et al., 2024; Mentzer et al., 2023; Jia et al., 2025) has rapidly advanced. This paradigm has also been naturally extended to video generation (Wang et al., 2024a; Chen et al., 2026; Ye et al., 2025; Yan et al., 2024; Li et al., 2025; Guo et al., 2025; Tan et al., 2025; Ge et al., 2022; Hong et al., 2022), where high-dimensional videos must first be compressed into discrete token sequences before autoregressive modeling. At the center of this pipeline lies the video tokenizer, typically consisting of a visual encoder, a quantizer, and a decoder. It serves as a lossy compressor that maps continuous video signals into a compact discrete latent space where AR generation is performed. However, compared with images, videos introduce substantially greater tokenization challenges due to their high dimensionality, strong temporal redundancy, and large variation in duration, resolution, and aspect ratio. Without effective compression, video token sequences can easily grow to prohibitive lengths, substantially increasing the computational cost of Transformer architectures (Vaswani et al., 2017; Dosovitskiy et al., 2020).

Despite recent progress, efficient video tokenization requires balancing three design considerations. First, token budgets should reflect differences in spatiotemporal complexity: a uniform budget can devote unnecessary capacity to simple clips while limiting the representation of more complex ones. Second, flexible input handling is important for supporting videos with different shapes and token budgets; fixed-shape pipelines can require cropping, resizing, or padding. Third, under aggressive compression, reconstruction objectives can benefit from semantic regularization to preserve structures

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.



(a) Previous video tokenizers (b) Our CodecTok (c) Video reconstruction quality *v.s.* Token efficiency

Figure 1: Comparison of video tokenization paradigms. (a) A fixed-shape, fixed-budget tokenizer. (b) CodecTok supports arbitrary-shaped input, while adaptively allocating token budgets according to content complexity. 'TransEnc.' denotes Transformer Encoder. (c) CodecTok achieves an effective tradeoff between reconstruction fidelity and token efficiency.

useful for downstream generation. These considerations motivate a unified design that combines adaptive allocation, flexible tokenization, and semantic alignment.

A central question therefore arises: *how should a video tokenizer determine the number of tokens required by each clip?* Recent adaptive tokenizers use reconstruction-quality criteria, information-based objectives, or learned allocation policies (Yan et al., 2024; Li et al., 2025; Ye et al., 2025; Xiong et al., 2026), demonstrating the value of variable-length representations. Their allocation procedures differ in the evaluations, search, or auxiliary training they require. This motivates a complementary design goal: assigning useful token budgets through an explicit, inexpensive content-based rule. We explore codec-derived spatial and temporal statistics as a practical prior for this decision.

In this work, we take a codec-informed rate-control view of adaptive video tokenization. Classical video codecs motivate allocating capacity according to coding difficulty; we use spatial texture and temporal residual activity as inexpensive indicators for token allocation. Guided by this perspective, we propose **CodecTok**, an efficient adaptive video tokenizer that brings content-dependent capacity allocation to discrete video representations. CodecTok unifies three components: **First**, a training-free *Adaptive Token Allocator* estimates spatial texture complexity from block-wise gradient statistics and temporal complexity from Hadamard-domain inter-frame residuals, directly producing content-adaptive token budgets without an auxiliary prediction network or per-clip budget search. Training-free refers to this allocator; the tokenizer and downstream AR model are trained normally. **Second**, a padding-free *Flexible Tokenizer* uses sequence packing and 3D rotary positional encoding (Su et al., 2024) to natively process arbitrary resolutions and variable-length token sequences. **Third**, a *Semantic Alignment* objective (Yu et al., 2025) distills structured representations from a frozen video foundation model, encouraging highly compressed tokens to preserve meaningful spatiotemporal semantics alongside reconstruction fidelity.

Experiments on UCF-101 (Soomro et al., 2012) demonstrate competitive quality–token-budget trade-offs for video reconstruction and downstream AR generation. At an average of 743 reconstruction tokens and 730 generated tokens, CodecTok achieves an rFVD of 16 and a gFVD of 44. A more compact configuration uses 585 reconstruction tokens and 560 generated tokens, reducing the respective lengths by approximately 43% and 45% relative to 1024 tokens, with an rFVD of 19 and a gFVD of 53. These operating points, together with allocation and alignment ablations, support the effectiveness of the integrated framework for efficient video tokenization and generation.

In summary, our main contributions are three-fold:

- We propose **CodecTok**, a unified content-adaptive video tokenizer for efficient autoregressive generation. Our framework integrates a padding-free backbone capable of natively processing arbitrary-resolution videos, a semantic alignment module to preserve meaningful spatiotemporal structure, and a codec-informed mechanism for dynamic token allocation.

- We introduce a training-free codec-informed token allocation strategy. By combining explicit spatial texture and temporal residual statistics, it assigns content-adaptive token budgets without an auxiliary prediction network or per-clip budget search, providing a low-overhead allocation mechanism within the framework.
- Experiments on UCF-101 demonstrate competitive reconstruction and autoregressive generation across different token budgets. Comparisons of allocation policies, component ablations, and generation timing characterize the quality–efficiency trade-offs of CodecTok.

2 Related Work

Discrete Image and Video Tokenizers. The discretization of visual signals has fundamentally advanced generative modeling. Pioneered by VQ-VAE (Van Den Oord et al., 2017), representing images as discrete tokens propelled the development of the Autoregressive paradigm (Esser et al., 2021; Wu et al., 2024; Team, 2024; Ramesh et al., 2021; Sun et al., 2024). Naturally, this paradigm was extended to the video domain (Yan et al., 2024; Li et al., 2025; Wang et al., 2024a; Guo et al., 2025; Ge et al., 2022; Tan et al., 2025; Hong et al., 2022; Xie et al., 2026). Early research (Ge et al., 2022; Wang et al., 2024c) primarily used 3D convolutions, but this tends to result in overly long token sequences, exacerbating computational bottlenecks and hindering the efficiency of subsequent AR models. Recently, inspired by TiTok (Yu et al., 2024), modern video tokenizers (Yan et al., 2024; Wang et al., 2024b) increasingly favor the Transformer architecture (Vaswani et al., 2017; Dosovitskiy et al., 2020). While the Transformer effectively improves scalability, increasing the compression ratio to address the generation efficiency bottleneck remains an unresolved challenge.

Adaptive Tokenization. Adaptive tokenization (Bachmann et al., 2025; Wang et al., 2025; Duggal et al., 2025; Mao et al., 2025) has garnered significant attention in recent years. In the realm of image tokenization, ALIT (Duggal et al., 2025) achieves dynamic representation by recursively distilling 2D tokens into 1D latent sequences. One-D-piece (Miwa et al., 2025) adopts a stochastic nested dropout strategy, while STAT (Chen et al., 2026) builds upon this by employing perceptual loss as a proxy for image complexity. In contrast, CAT (Shen et al., 2025) utilizes Large Language Models to evaluate image complexity. As the focus extends to the video domain, ElasticTok (Yan et al., 2024) and AdapTok (Li et al., 2025) similarly employ tail-dropping strategies and rely on heuristics for adaptive token allocation; however, this reliance can lead to suboptimal results. More recently, InfoTok (Ye et al., 2025) adopts an approach based on the Evidence Lower Bound to mask out less informative tokens.Differently, we propose to leverage video coding principles. By directly evaluating spatial and temporal complexity through codec-informed entropy metrics, our method achieves a highly accurate and efficient token allocation scheme.

Representation Alignment. The efficacy of knowledge distillation from Vision Foundation Models (Oquab et al., 2023; Siméoni et al., 2025; Assran et al., 2025; Mur-Labadia et al., 2026; Tschannen et al., 2025; Wang et al., 2022) has been well-established. In the image domain, GigaTok (Xiong et al., 2025) employs semantic regularization to enhance the generation quality of downstream autoregressive models, while UniTok (Ma et al., 2025) introduces an image-text contrastive loss to unify visual understanding and generation tasks. As representation alignment (Yu et al., 2025) extends into the video generation domain, VideoREPA (Zhang et al., 2025) incorporates a soft alignment loss to distill the understanding of temporal dynamics and physical laws. Inspired by them, we leverage a video foundation model for semantic alignment, aiming to restore conceptual structures and enhance the robust representation capability of the tokenizer.

3 Method

Our goal is to develop an efficient video tokenizer that adaptively reduces redundant token usage while preserving high-fidelity reconstruction and downstream generation capability. Motivated by the rate-control principle of classical video coding, we propose **CodecTok**, a codec-informed video tokenizer that transfers this principle from bit allocation to token allocation. As illustrated in Fig. 2, CodecTok is built around a **Codec-Informed Token Allocator** (Sec. 3.1), which estimates spatial texture complexity and temporal residual complexity to assign a content-adaptive token budget for each clip. A padding-free flexible tokenizer then uses sequence packing and 3D rotary positional encoding to natively process arbitrary resolutions and variable-length token sequences. Finally,

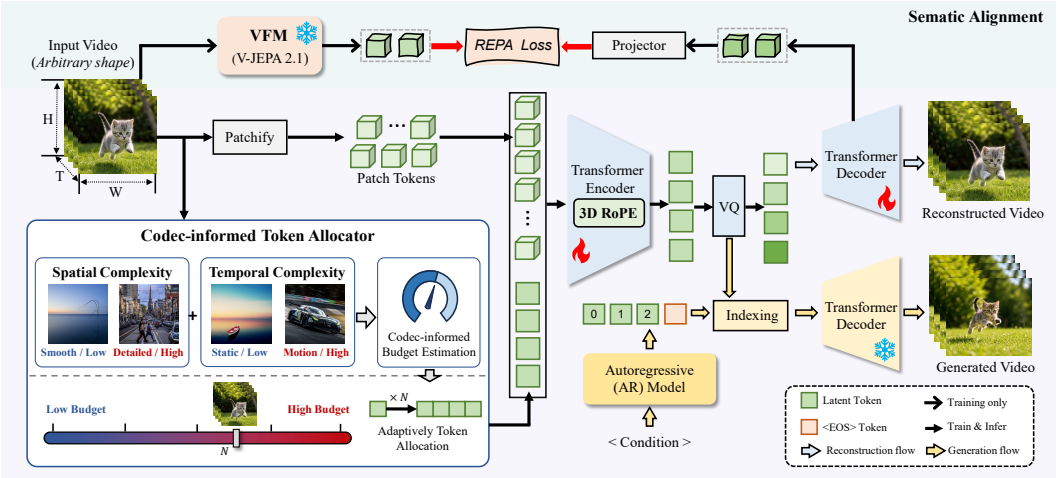


Figure 2: Overall architecture of CodecTok. An arbitrary-resolution video is first evaluated by the Codec-informed Token Allocator to assign an optimal token budget (N). During training, the framework is jointly optimized via standard reconstruction and semantic alignment distillation. For downstream generation, an autoregressive model adaptively predicts the sequence termination position, passing the generated tokens through the decoder to efficiently synthesize novel videos.

a **Semantic Alignment** module (Sec. 3.3) distills structured representations from a frozen video foundation model, encouraging highly compressed tokens to preserve meaningful spatiotemporal semantics.

3.1 Codec-informed Token Allocator

A fundamental challenge in efficient video tokenization is determining the optimal sequence length for diverse clips. Existing tokenizers typically adopt a rigid constant-budget paradigm, inefficiently assigning identical token counts to both static and highly dynamic scenes. From a rate-distortion perspective, a tokenizer acts as a lossy codec mapping a video $V_i \in \mathbb{R}^{C\times T_i\times H_i\times W_i}$ to N_i discrete symbols from a codebook Q, yielding an effective bitrate $R_i = N_i \log_2 |Q|$. Since higher information content necessitates higher coding rates for a fixed fidelity, token budgets must adapt to intrinsic spatiotemporal information densities.

Motivated by this, we formulate token allocation as a variable-bitrate problem. Mimicking classical video codecs that allocate more bits to complex content, we introduce a lightweight, training-free *Codec-Informed Token Allocator*. It explicitly approximates coding difficulty by estimating spatial complexity via block-wise gradient statistics and temporal complexity via Hadamard-domain inter-frame residuals. These explicit signals are directly mapped to a content-adaptive token budget, completely bypassing the need for auxiliary networks or inference-time search.

Spatial Complexity via Non-stationary Block Analysis.

Modern video codecs such as HEVC do not treat a frame as a homogeneous signal: they recursively partition it into Coding Tree Units so that coding parameters adapt to local content complexity—a direct response to the heavy-tailed statistics of natural images. Following this principle, we partition each frame into non-overlapping $B\times B$ blocks $\mathcal{B}$ and characterize each block with a classical operator. Under a local Gaussian model, the differential entropy of an image patch is monotonic in its gradient variance, so gradient energy serves as a principled, training-free surrogate for $h(L_t)$. We extract the luminance $L \in \mathbb{R}^{T\times H\times W}$ and estimate the gradient field via 3×3 Sobel operators S_x, S_y:

$$G_t = \sqrt{(L_t * S_x)^2 + (L_t * S_y)^2 + \epsilon}. \tag{1}$$

Within each block we compute the mean μ_b and standard deviation σ_b, and form a dispersion-corrected frame-level score:

$$\mathcal{C}^{(t)}_{\text{spa}} = \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} \mu_b \left(1 + \frac{\sigma_b}{\mu_b + \epsilon}\right). \tag{2}$$

The multiplicative correction is a (scaled) coefficient of variation that up-weights blocks whose gradient energy is concentrated in a few hotspots. Taking the median across a uniformly sampled temporal subset yields $\mathcal{C}_{\text{spa}}$. From a compression perspective, a higher $\mathcal{C}_{\text{spa}}$ provides a lightweight proxy for spatial information density, thereby demanding the router to allocate a larger spatial token budget to prevent detail degradation.

Temporal Complexity via Transform-domain Residuals. The decomposition reduces temporal entropy to $h(R_t)$, the entropy of the inter-frame residual. Naive pixel-wise absolute differences fail to capture the true frequency-domain sparsity of motion. Adopting the theoretically grounded rate estimation strategy of classical motion compensation, we quantify temporal entropy via the *Sum of Absolute Transformed Differences* (SATD) under an orthogonal basis. Specifically, we extract the inter-frame residual $R_t = L_t - L_{t-1}$, partition it into 4×4 blocks P, and project each block into the spectral domain via the 2D Hadamard transform:

$$\mathcal{T} = H_4\, P\, H_4^{\top}, \tag{3}$$

$$\mathrm{SATD}(P) = \sum_{u,v} |\mathcal{T}_{u,v}|, \tag{4}$$

where H_4 is the normalized Hadamard matrix. As an orthogonal transform, SATD serves as a mathematically sound and monotonic surrogate for temporal entropy. Furthermore, since $H_4 \in \{\pm1\}^{4\times4}$, the projection requires only highly efficient integer additions and sign flips. Averaging SATD across all blocks and frame pairs yields the temporal complexity $\mathcal{C}_{\text{tem}}$. Acting as a practical indicator of temporal complexity, a higher $\mathcal{C}_{\text{tem}}$ signals sparse inter-frame correlation, dictating the router to allocate a larger temporal token budget to encode complex motion.

Efficient Training-Free Allocation. The final information density is defined as $\mathcal{S}_i = \alpha\,\mathcal{C}_{\text{spa},i} + \beta\,\mathcal{C}_{\text{tem},i}$. To seamlessly accommodate videos of arbitrary resolutions and durations, we introduce a sub-linear spatiotemporal scaling factor $\rho_i = (\frac{T_i H_i W_i}{T_{\text{ref}} H_{\text{ref}} W_{\text{ref}}})^{\gamma}$. Here, $T_{\text{ref}}, H_{\text{ref}}, W_{\text{ref}}$ denote the reference dimensions (e.g., $16 \times 128 \times 128$), and the exponent $\gamma \in (0, 1)$ dampens the proportional growth to account for inherent spatial redundancies at higher resolutions. Based on our **Codec-informed Token Allocator**, the discrete token budget N_i is thus deterministically mapped via:

$$N_i = \mathrm{Clip}(\lfloor (k \cdot \mathcal{S}_i + b) \cdot \rho_i \rceil, N_{\min}, N_{\max}) \tag{5}$$

where k and b act as base bitrate anchors pre-calibrated at the reference resolution.

3.2 Flexible Tokenization for Arbitrary-shaped Input

To seamlessly accommodate the dynamic token budgets allocated by the Codec-informed Token Allocator and natively process videos of arbitrary resolutions, we formulate a sequence packing strategy. Specifically, the i-th heterogeneous video is partitioned into visual patches Z_i. We dynamically instantiate N_i latent queries M_i, where the exact quantity N_i is strictly governed by the router. The unified sequence for each video is formed by concatenation $X_i = [M_i \parallel Z_i]$ with a total length of L_i. These varying-length sequences are then directly flattened into a continuous 1D stream $X_{\text{packed}} \in \mathbb{R}^{(\sum L_i)\times D}$. By tracking the cumulative sequence boundaries, we leverage memory-efficient variable-length attention mechanisms to ensure exact, isolated computations within each video instance. To explicitly inject structural priors and maintain spatiotemporal coherence across diverse input dimensions, we incorporate 3D Rotary Position Embedding (RoPE), grounding visual patches with 3D axial coordinates. Architecturally, the data flow operates as an information bottleneck: the encoder aggregates global spatiotemporal contexts into the query tokens M_i while discarding the raw visual patches Z_i. Following standard quantization, a symmetric flexible decoder reverses this process, utilizing the highly compressed discrete tokens to attend and expand back to the original pixel space for high-fidelity reconstruction.

3.3 Semantic Alignment via Video Foundation Model

While our allocator achieves extreme token compression, relying exclusively on pixel-level objectives forces the encoder to cram high-frequency textures into limited surviving tokens. Consequently, the resulting latent space becomes heavily entangled, obscuring semantic structures and hindering downstream generative modeling. To counteract this, we introduce a *Semantic Alignment* module

that regularizes the latent manifold by distilling structured conceptual knowledge from a frozen foundation model (e.g., V-JEPA 2.1 Mur-Labadia et al. (2026)).

Let $F_{\mathcal{T}}$ denote the reference feature grid extracted from the frozen teacher. For the student representation, we extract the feature $F_{\mathcal{S}}$ from the shallow layers of the decoder, where the compressed token sequence has already been expanded back into a dense spatiotemporal grid, inherently possessing the 3D structural geometry of $\mathbb{R}^{T' \times H' \times W' \times D}$. Because CodecTok operates on arbitrary resolutions, we apply dynamic trilinear interpolation to upsample the student features to the teacher's grid. Formally, this upward alignment uses interpolation followed by a lightweight linear adapter Φ: $\hat{F}_{\mathcal{S}} = \Phi(\text{Interpolate}(F_{\mathcal{S}}))$. The alignment objective maximizes the cosine similarity between the aligned student and the teacher:

$$\mathcal{L}_{\text{align}} = 1 - \frac{1}{|\Omega|} \sum_{x \in \Omega} \frac{\hat{F}_{\mathcal{S},x} \cdot F_{\mathcal{T},x}}{\|\hat{F}_{\mathcal{S},x}\|_2 \|F_{\mathcal{T},x}\|_2}, \tag{6}$$

where Ω denotes the set of spatiotemporal coordinates. Optimized alongside reconstruction losses, this distillation ensures that highly compressed tokens prioritize meaningful motion semantics over perceptual noise, bridging the gap between extreme compression and high-fidelity generation.

3.4 Training Objectives

Our CodecTok is optimized with a composite objective:

$$\mathcal{L} = \mathcal{L}_R + \mathcal{L}_{VQ} + \mathcal{L}_P + \mathcal{L}_G + \mathcal{L}_{align}, \tag{7}$$

where $\mathcal{L}_R$ denotes the L_1 reconstruction loss, $\mathcal{L}_{VQ}$ the quantizer loss, $\mathcal{L}_P$ the perceptual loss (Zhang et al., 2018), $\mathcal{L}_G$ the adversarial loss to enhance visual quality, and $\mathcal{L}_{align}$ the semantic alignment loss for incorporating foundational priors.

4 Experiments

4.1 Experimental Settings

Datasets: We train CodecTok on the UCF-101 (Soomro et al., 2012) and Kinetics-600 (Carreira et al., 2018) datasets and evaluate it on UCF-101. During training, we use video clips with dimensions ranging from $12 \times 128 \times 128$ to $24 \times 256 \times 256$. During evaluation, we fix the clip dimensions to $16 \times 128 \times 128$.

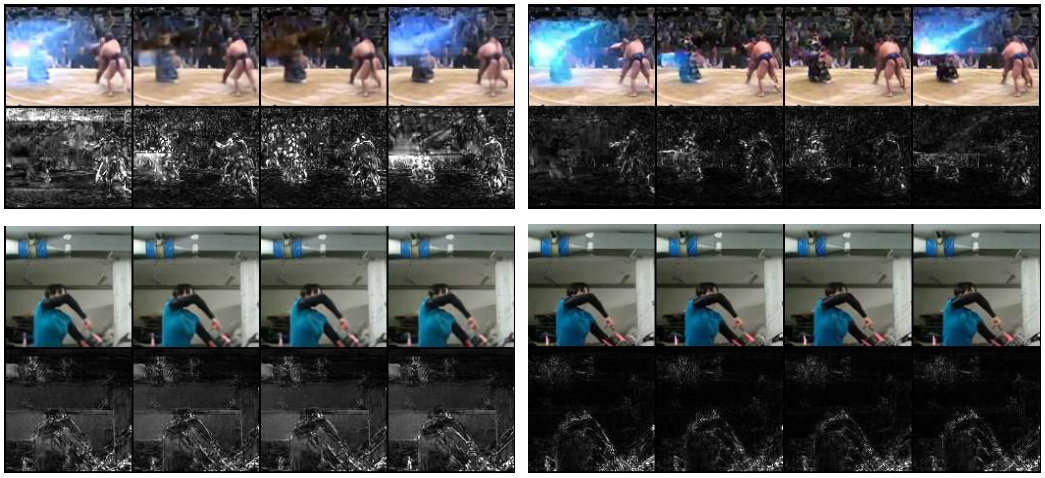
Implementation details: When patchifying videos in CodecTok, the spatial and temporal downsample ratios are set to 8 and 4, respectively. CodecTok adopts SLQ (Zhao et al., 2025) subset as the quantization method, and V-JEPA 2.1 (Mur-Labadia et al., 2026) is employed as the frozen teacher for representation alignment. For our main results, the models were trained for 150 epochs. Due to computational resource constraints, the models used in the ablation studies and experiments across varying resolutions were trained for 75 epochs. For downstream generative modeling, we train LLaMA-like GPT models (Sun et al., 2024) natively on these variable-length token sequences. Detailed hyper-parameters are provided in the supplementary material.

4.2 Main Results

Video Reconstruction and Generation. We evaluate CodecTok on the UCF-101 (Soomro et al., 2012) dataset in terms of video reconstruction quality and downstream autoregressive (AR) generation. As shown in Tab. 1, CodecTok achieves a competitive quality–efficiency trade-off at two adaptive token budgets. With an average of 743 reconstruction tokens and 730 generation tokens, CodecTok achieves a rFVD of 16, and a gFVD of 44. Compared with LARP-L-Long (Wang et al., 2024a), this variant improves both reconstruction and generation quality while reducing reconstruction and generation token counts by approximately 27% and 29%, respectively. Compared with CodecTok, EVATok achieves a lower rFVD (9.7 vs. 16) but exhibits lower pixel-level fidelity, as reflected by its PSNR (25.75 vs. 28.85). CodecTok also achieves a lower gFVD (44 vs. 48), while using fewer tokens for both reconstruction and generation. At a more compact budget, CodecTok uses only 585 reconstruction tokens and 560 generation tokens on average, achieving an rFVD of 19 and a

Table 1: System-level comparison on the UCF-101 dataset. CodecTok achieves comparable or better results in both video reconstruction and generation, while reducing token consumption by over 40%.

Method	#Params		#rTokens	PSNR↑	rFVD↓	gFVD↓	#gTokens
	Tokenizer	Generator					
Diffusion-based generative models with continuous video tokenizers							
VideoFusion	-	2B	-	-	-	173	-
HPDM	-	725M	-	-	-	66	-
W.A.L.T-L	-	313M	-	-	-	46	-
MLM generative models with discrete video tokenizers							
MAGVIT-MLM	158M	306M	1024	22	25	76	1024
MAGVIT-v2-MLM	-	307M	1280	-	**8.6**	58	1280
AR generative models with discrete video tokenizers							
CogVideo	-	9.4B	-	-	-	626	-
TATS	32M	321M	1024	-	162	332	1024
MAGVIT-AR	158M	306M	1024	22	25	265	1024
MAGVIT-v2-AR	-	840M	1280	-	**8.6**	109	1280
OmniTokenizer	82.2M	650M	1280	-	42	191	1280
ElasticTok	356M	-	2048	-	93	-	-
AdapTok-512	195M	-	512	24.06	60	-	-
AdapTok	195M	633M	1024	25.72	36	67	1024
STAT-Video	195M	-	997	<u>28.77</u>	29	-	-
LARP-L-Long	173M	343M	1024	28.70	20	102	1024
LARP-L-Long	173M	632M	1024	28.70	20	57	1024
EVATok	145M	633M	774	25.75	<u>9.7</u>	<u>48</u>	756
CodecTok (ours)	157M	632M	743 (-27%)	**28.85**	16	**44**	730 (-29%)
CodecTok (ours)	157M	632M	585 (-43%)	27.95	19	53	560 (-45%)



(a) OmniTokenizer (b) CodecTok

Figure 3: Visualization comparison of the reconstruction results on the UCF-101 dataset.We visualize the reconstruction and GT error map, where brighter areas indicate larger errors.

gFVD of 53 while reducing the corresponding token counts by approximately 43% and 45% relative to a 1024-token budget. These results demonstrate that CodecTok substantially reduces the token requirements of video reconstruction and AR generation while maintaining competitive visual quality.

Accelerated Autoregressive Generation. A primary advantage of CodecTok is its ability to significantly accelerate downstream generation. By achieving an exceptionally high spatiotemporal



(a) LARP (b) CodecTok

Figure 4: Visualization comparison of the generation results on the UCF-101 dataset.

Table 2: Comparison of inference latency in autoregressive generation.

Method	Time (s/video)
LARP	22
CodecTok (Ours)	**14**

Table 3: Reconstruction performance and token efficiency at higher resolution (480p).

Setting (480p)	Avg. Tokens ↓	PSNR ↑	rFVD ↓
CodecTok-Fix	4096	31.73	5.93
CodecTok	2362	30.94	7.10

compression ratio, our tokenizer represents videos with highly compact token sequences. This directly alleviates the computational bottleneck of sequential autoregressive decoding. As shown in Tab. 2, compared to LARP, CodecTok accelerates the video generation process by over 36%, demonstrating its superior efficiency for practical deployment.

Model Scaling. To further investigate the impact of model scale on performance, we trained three models of different scales: CodecTok-S, CodecTok-B, and CodecTok-L. As shown in Tab. 4, reconstruction metrics consistently improve as the model size increases. This demonstrates that our model benefits from increased model capacity, highlighting the scalability of our approach.

Table 4: Reconstruction performance across different model scales (1024 tokens).

Settings	Params	PSNR↑	rFVD↓
CodecTok-S	50M	28.36	31
CodecTok-B	107M	28.68	21
CodecTok-L	157M	**29.20**	**14**

Scalability to Higher Resolutions. To further validate the scalability of our approach for modern high-resolution video generation, we extend our evaluation to the 480p spatial resolution. As shown in Table 3, after fine-tuning the model at 480p, we compare our adaptive CodecTok against a constant-budget baseline (CodecTok-Fix). Notably, our dynamic allocation strategy slashes the average token consumption from 4096 down to 2362 (an approximately 42% reduction). Meanwhile, the model maintains highly competitive reconstruction fidelity, yielding a PSNR of 30.94 and an rFVD of 7.10. This demonstrates that our method generalizes well to higher resolutions, successfully mitigating the quadratic sequence-length explosion while preserving essential visual quality.

4.3 Ablation Study

Comparison of Different Token Allocation Strategies. We compare our Codec-informed strategy with two distinct baselines: fixed allocation and LPIPS-based strategy. We assess these approaches across different average token budgets for robust evaluation. As illustrated in Fig. 6, our method consistently outperforms others across all perceptual metrics, further proving that our spatiotemporal entropy estimation captures genuine information density better than purely perceptual proxies. This more efficient information representation is thereby beneficial for subsequent generation tasks.

Correlation with Traditional Video Coding. To validate the physical rationality of our dynamic allocation, we visualize the allocated token budgets against the raw MP4 file sizes of the videos. We use MP4 file size under a fixed encoding configuration as an empirical proxy for codec-based compression difficulty. As shown in Fig. 5, a highly positive correlation is observed between our token budgets and traditional video coding outcomes. Videos with complex motion or dense textures (larger MP4 sizes) are naturally assigned more tokens, while static sequences are aggressively compressed. This strong alignment confirms that our Codec-informed Allocator strictly adheres to the Variable Bitrate philosophy, effectively allocating capacity based on true spatiotemporal redundancy.



Figure 5: Correlation between allocated token budgets and intrinsic video complexity.

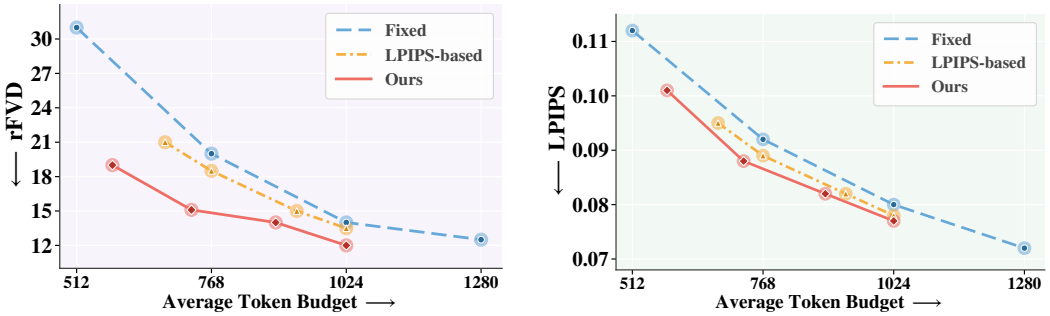


Figure 6: Comparison of token allocation strategies for reconstruction across varying token budgets.

Semantic Alignment. Tab. 5 validates our Representation Alignment module. Relying solely on pixel-level reconstruction (*w/o Rep. Align.*) yields poor semantic retention, severely crippling downstream generation. Incorporating V-JEPA semantic distillation dramatically improves generation quality, slashing gFVD from 45 to 34 (w/o router) and 57 to 44 (w/ router), alongside slight gains in reconstruction fidelity. This proves our strategy effectively injects structured, high-level semantics into discrete tokens, which is vital for autoregressive generation.

Table 5: Ablation study on the effectiveness of the router and representation alignment module.

Router	Rep. Align.	rTokens↓	PSNR↑	rFVD↓	gFVD↓
No	No	1024	29.35	15	45
No	Yes (JEPA)	1024	29.20	14	**34**
Yes	No	743	28.97	17	57
Yes	Yes (JEPA)	743	28.85	16	**44**

Router Comparison with InfoTok and EVATok. To further evaluate our allocation strategy, we integrate the routers of InfoTok (Ye et al., 2025) and EVATok (Xiong et al., 2026) into the same tokenizer, keeping all other model components unchanged, for a controlled comparison at an average budget of 743 tokens. As shown in Tab. 6, the three strategies achieve comparable reconstruction quality, but CodecTok requires only 0.09 ms for allocation, versus 63 ms and 15.5 ms, respectively. With a shared encoding and decoding cost of approximately 77 ms, CodecTok reduces the estimated total reconstruction latency by approximately 45% and 17%, respectively, demonstrating its substantial efficiency advantage.

Table 6: Controlled comparison with InfoTok and EVATok.

Router	ms↓	PSNR↑	LPIPS↓	MSE↓
InfoTok	63.0	**28.95**	0.075	0.002242
EVATok	15.5	28.79	0.074	0.002167
CodecTok	**0.09**	28.85	**0.073**	**0.002141**

5 Conclusion

We introduced **CodecTok**, rethinking video tokenization as a Variable Bitrate compression problem. We dynamically allocate tokens based on genuine spatiotemporal entropy via a Codec-informed Token Allocator to achieve highly efficient video tokenization, breaking the rigid Constant Bitrate paradigm. Featured with flexible encoding and a Semantic Alignment module, CodecTok natively processes arbitrary resolutions while ensuring that highly compressed discrete latents retain robust, high-level semantics. Extensive experiments demonstrate that CodecTok establishes new state-of-the-art performance in both video reconstruction and autoregressive generation on UCF-101. Crucially, it achieves this superior fidelity while reducing token consumption by over 40%, thereby significantly alleviating the computational bottleneck for downstream AR generators.

AI USE STATEMENT

We used a large language model to assist with translation, drafting and revising prose, condensing text, and LaTeX formatting. The authors reviewed the AI-assisted content for accuracy and consistency with the underlying experiments and sources and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Section 3 describes the codec-informed token allocator, flexible tokenizer architecture, semantic alignment mechanism, and training objectives. Section 4.1 summarizes the experimental setup. Appendices B–D provide implementation details, dataset descriptions, and training hyperparameters for the tokenizer and autoregressive models, with the main configurations summarized in Table 7. We will release the code to facilitate reproducibility.

REFERENCES

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Roman Bachmann, Jesse Allardice, David Mizrahi, Enrico Fini, Oğuzhan Fatih Kar, Elmira Amirloo, Alaaeldin El-Nouby, Amir Zamir, and Afshin Dehghan. Flextok: Resampling images into 1d token sequences of flexible length. In *Forty-second International Conference on Machine Learning*, 2025.

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

Zeyuan Chen, Kai Zhang, Zhuowen Tu, and Yuanjun Xiong. Soft tail-dropping for adaptive visual tokenization. *arXiv preprint arXiv:2601.14246*, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Shivam Duggal, Phillip Isola, Antonio Torralba, and William T Freeman. Adaptive length image tokenization via recurrent allocation. In *First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models*, 2025.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. *arXiv preprint arXiv:2204.03638*, 2022.

Pengbo Guo, Junke Wang, Zhen Xing, Chengxu Liu, Daoguo Dong, Xueming Qian, and Zuxuan Wu. Dera: Decoupled representation alignment for video tokenization. *arXiv preprint arXiv:2512.04483*, 2025.

Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *European Conference on Computer Vision*, pp. 393–411. Springer, 2024.

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

Mingkai Jia, Wei Yin, Xiaotao Hu, Jiaxin Guo, Xiaoyang Guo, Qian Zhang, Xiao-Xiao Long, and Ping Tan. Mgvq: Could vq-vae beat vae? a generalizable tokenizer with multi-group quantization. *arXiv preprint arXiv:2507.07997*, 2025.

Yan Li, Changyao Tian, Renqiu Xia, Ning Liao, Weiwei Guo, Junchi Yan, Hongsheng Li, Jifeng Dai, Hao Li, and Xue Yang. Learning adaptive and temporally causal video tokenization in a 1d latent space. *arXiv preprint arXiv:2505.17011*, 2025.

Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321*, 2025.

Lingjun Mao, Rodolfo Corona, Xin Liang, Wenhao Yan, and Zineng Tang. Images are worth variable length of representations. *arXiv preprint arXiv:2506.03643*, 2025.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023.

Keita Miwa, Kento Sasaki, Hidehisa Arai, Tsubasa Takahashi, and Yu Yamaguchi. One-d-piece: Image tokenizer meets quality-controllable compression. *arXiv preprint arXiv:2501.10064*, 2025.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mahmoud Assran, Koustuv Sinha, Michael Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831, 2021.

Junhong Shen, Kushal Tirumala, Michihiro Yasunaga, Ishan Misra, Luke Zettlemoyer, Lili Yu, and Chunting Zhou. Cat: Content-adaptive image tokenization. *arXiv preprint arXiv:2501.03120*, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, and Sergey Tulyakov. Hierarchical patch diffusion models for high-resolution video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7569–7579, 2024.

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.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

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

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Hanyu Wang, Saksham Suri, Yixuan Ren, Hao Chen, and Abhinav Shrivastava. Larp: Tokenizing videos with a learned autoregressive generative prior. *arXiv preprint arXiv:2410.21264*, 2024a.

Junke Wang, Yi Jiang, Zehuan Yuan, Binyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *Advances in Neural Information Processing Systems*, 37:28281–28295, 2024b.

Lingfeng Wang, Hualing Lin, Senda Chen, Tao Wang, Changxu Cheng, Yangyang Zhong, Dong Zheng, and Wuyue Zhao. Alto: Adaptive-length tokenizer for autoregressive mask generation. *arXiv preprint arXiv:2505.16495*, 2025.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024c.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable multi-modal generators. *arXiv preprint arXiv:2412.04332*, 2024.

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

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18770–18780, October 2025.

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

Wilson Yan, Volodymyr Mnih, Aleksandra Faust, Matei Zaharia, Pieter Abbeel, and Hao Liu. Elastictok: Adaptive tokenization for image and video. *arXiv preprint arXiv:2410.08368*, 2024.

Haotian Ye, Qiyuan He, Jiaqi Han, Puheng Li, Jiaojiao Fan, Zekun Hao, Fitsum Reda, Yogesh Balaji, Huayu Chen, Sheng Liu, et al. Infotok: Adaptive discrete video tokenizer via information-theoretic compression. *arXiv preprint arXiv:2512.16975*, 2025.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10459–10469. IEEE, 2023a.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023b.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37:128940–128966, 2024.

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

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

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025.

Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. *arXiv preprint arXiv:2406.07548*, 2024.

Yue Zhao, Hanwen Jiang, Zhenlin Xu, Chutong Yang, Ehsan Adeli, and Philipp Krähenbühl. Spherical leech quantization for visual tokenization and generation. *arXiv preprint arXiv:2512.14697*, 2025.

A LIMITATIONS

While CodecTok demonstrates significant improvements in token efficiency and perceptual quality for video generation, we acknowledge certain limitations that outline directions for future research. First, to maximize computational efficiency, our Codec-informed Token Allocator relies on signal processing metrics. Consequently, it may misjudge the actual semantic complexity in video scenes afflicted by severe sensor noise or compression artifacts, inadvertently treating them as high-frequency details. Second, variable-length generation remains a substantial challenge for autoregressive video models. Generating token sequences of arbitrary lengths tends to cause instability in the discrete latent space, which can subsequently degrade the overall generation quality. Exploring robust AR architectures that stably support variable-length tokens and developing noise-resistant allocation metrics are critical directions for our future work.

B EXPERIMENTAL SETTINGS

Visual Tokenizer. Formally, CodecTok comprises an encoder $\mathcal{E}$, a latent quantizer $\mathcal{Q}$, and a decoder $\mathcal{D}$. To handle varying video reconstruction scales ranging from $12 \times 128 \times 128$ to $24 \times 256 \times 256$, we integrate 3D Rotary Positional Embeddings Su et al. (2024) into both $\mathcal{E}$ and $\mathcal{D}$. Input videos are normalized to the $[-1, 1]$ range. For the codec-informed token allocator, we use $\alpha = 0.7$, $\beta = 0.3$, and $\gamma = 0.5$, with token-budget bounds of $N_{\min} = 384$ and $N_{\max} = 1536$ at the reference video shape of $16 \times 128 \times 128$. For high-fidelity reconstruction, we incorporate an adversarial objective using a ViT-based discriminator structurally analogous to $\mathcal{E}$. During training, the discriminator and tokenizer are updated asynchronously: the discriminator receives one update per five tokenizer steps, with its learning rate strictly scaled to 30% of the tokenizer's. The framework is optimized via Adam ($\beta_1 = 0.9, \beta_2 = 0.95$) with a base learning rate of 1×10^{-4}, modulated by a cosine annealing schedule and a linear warm-up phase. For ablation studies, CodecTok is trained for 75 epochs with a batch size of 64, while for the main results, we extend training to 150 epochs with a batch size of 128.

AR Generative Model. We adopt LLaMA-like Transformers Sun et al. (2024) as our autoregressive (AR) generative models, developing two variants with 343M and 632M parameters. To enable variable-length video generation, an End-of-Sequence (EOS) token is appended to the flattened sequence of quantized latent tokens during training. During inference, the generation process halts dynamically upon sampling this EOS token. Both models are trained on the UCF-101 Soomro et al. (2012) dataset for 3,000 epochs using a global batch size of 64. Optimization is performed via AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$, weight decay 0.05). The base learning rate is set to 6×10^{-4}, governed by a cosine annealing schedule preceded by a linear warm-up phase.

C TRAINING DATASETS

UCF-101. Comprising 13,320 video clips distributed across 101 action categories, UCF-101 Soomro et al. (2012) serves as a widely adopted benchmark for action recognition. The dataset is partitioned into 9,537 training samples and 3,783 testing samples. It remains a highly challenging benchmark due to its vast diversity in real-world conditions, including complex camera dynamics, viewpoint shifts, scale variations, changing illumination, and heavily cluttered backgrounds.

Kinetics-600. Kinetics-600 Carreira et al. (2018) is a comprehensive video benchmark featuring approximately 480K videos spanning 600 distinct action classes. The dataset is structured into subsets of 390K, 30K, and 60K videos for training, validation, and testing, respectively. Sourced directly from YouTube, each sample consists of a 10-second clip specifically trimmed to capture the most salient moments of the corresponding action.

D TRAINING SETTINGS

The detailed training hyper-parameter settings for CodecTok are reported in Table. 7. Both the fixed-length version(CodecTok-fix) and the Codec-informed Token Allocator version(CodecTok) share identical parameters during training.

Table 7: Hyperparameters for the main CodecTok experiments. The AR column reports the 632M generator. Video shapes are given as $T \times H \times W$ and, for the AR model, refer to the corresponding video clips. The tokenizer shape range refers to base training, excluding the additional 480p fine-tuning setting.

	CodecTok	CodecTok-fix	CodecTok-AR
Model parameters			
Parameters	157M	157M	632M
Min. video shape	$12 \times 128 \times 128$	$12 \times 128 \times 128$	$16 \times 128 \times 128$
Max. video shape	$24 \times 256 \times 256$	$24 \times 256 \times 256$	$16 \times 128 \times 128$
Patch size	$4 \times 8 \times 8$	$4 \times 8 \times 8$	–
Hidden size	768	768	1280
Transformer layers	12	12	30
Training			
Optimizer	Adam	Adam	AdamW
Learning rate	1×10^{-4}	1×10^{-4}	6×10^{-4}
β_1	0.9	0.9	0.9
β_2	0.95	0.95	0.95
Weight decay	0	0	0.05
Learning-rate schedule	Cosine	Cosine	Cosine
Warm-up epochs	8	8	4
Global batch size	128	128	64
Training epochs	150	150	3000

For tokenizer ablations, we use 75 training epochs and a global batch size of 64. The AR model operates on discrete tokens and therefore does not use a video patchification layer.

Training Stability. Despite incorporating a complex multi-objective optimization and operating under an arbitrary-resolution training paradigm, CodecTok exhibits remarkable training stability. In line with the evaluation protocol of LARP Wang et al. (2024a), Figure 7 illustrates the training loss trajectories across four model scales: CodecTok-Small, CodecTok-Base, CodecTok-Large, and CodecTok-Large-Long. Specifically, the first three variants are trained for 75 epochs with a global batch size of 64, while the Large-Long variant undergoes extended training for 150 epochs with a batch size of 128. Across all configurations and extended training schedules, the loss curves demonstrate a consistently smooth and stable convergence without any signs of divergence or fluctuation spikes.

E Baselines

We compare CodecTok with representative video tokenization and generation methods, including VideoFusion Luo et al. (2023), HPDM Skorokhodov et al. (2024), W.A.L.T-L Gupta et al. (2024), CogVideo Hong et al. (2022), TATS Ge et al. (2022), MAGVIT Yu et al. (2023a), MAGVIT-v2 Yu et al. (2023b), OmniTokenizer Wang et al. (2024b), ElasticTok Yan et al. (2024), AdapTok Li et al. (2025), STAT-Video Chen et al. (2026), LARP Wang et al. (2024a), and EVATok Xiong et al. (2026). For allocation ablations, we additionally compare fixed-budget and LPIPS-based allocation, and integrate the routers of InfoTok Ye et al. (2025) and EVATok Xiong et al. (2026) into the same tokenizer for controlled comparisons.

F Additional Results

More evaluation metrics. As illustrated in Table 8, to provide a comprehensive evaluation, we compare CodecTok against recent state-of-the-art models, such as LARP Wang et al. (2024a) and AdapTok Li et al. (2025), across both reconstruction metrics (PSNR, SSIM, LPIPS, rFVD) and generation metrics (gFVD). Specifically, our fixed-length variant, CodecTok-fix, achieves state-of-the-art performance across all metrics, establishing a new benchmark for both video reconstruction and autoregressive generation, thereby demonstrating the fundamental superiority of our model

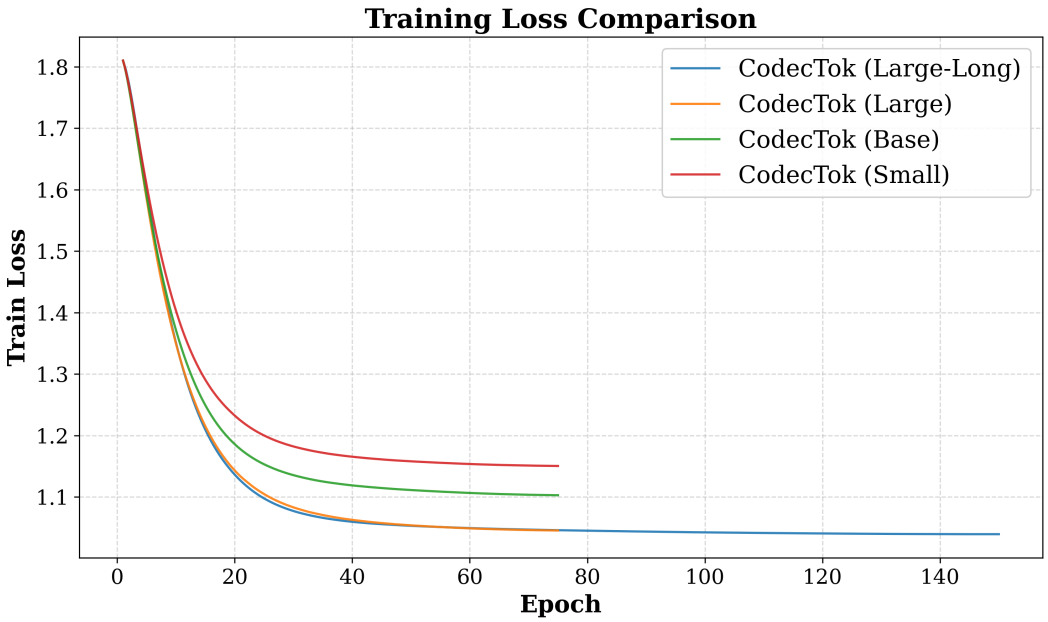


Figure 7: Training loss.

architecture. Furthermore, remarkably, the dynamic CodecTok utilizing merely 743 reconstruction tokens and 730 generation tokens still outperforms existing baselines (which use 1024 tokens) on both rFVD and gFVD, while remaining highly competitive across other pixel-level metrics. This firmly proves the extreme efficiency and effectiveness of our Codec-informed allocation mechanism in both compression and generative tasks.

Table 8: Comprehensive performance evaluation of CodecTok and existing baselines across video reconstruction and generation metrics.

Methods	PSNR↑	SSIM↑	LPIPS↓	rFVD↓	#rTokens↓	gFVD↓	#gTokens↓
TATS	24.22	-	0.206	162	1024	332	1024
OmniTokenizer	27.76	0.88	0.112	42	1280	191	1280
AdapTok	25.72	0.84	0.114	36	1024	67	1024
LARP-L-Long	28.70	0.89	0.075	20	1024	57	1024
CodecTok	27.95	0.88	0.087	19	**585**	53	**560**
CodecTok	<u>28.85</u>	<u>0.90</u>	<u>0.073</u>	<u>16</u>	<u>743</u>	<u>44</u>	<u>730</u>
CodecTok-fix	**29.20**	**0.91**	**0.071**	**14**	1024	**34**	1024

Detailed Comparison of Different Token Allocation Strategies. As a quantitative supplement to the reconstruction comparison in the main text, Table 9 further illustrates the impact of token allocation on video reconstruction. Our method achieves the lowest FVD at all three average token budgets. At budgets of 768, 890, and 1024 tokens, our codec-informed strategy reduces FVD from 20.0, 17.3, and 15.0 with fixed allocation to 15.0, 13.0, and 11.5, respectively. Compared with LPIPS-based allocation, the corresponding FVD reductions are 3.5, 3.0, and 2.0. These results support the effectiveness of codec-informed allocation for reconstruction under constrained token budgets.

Robustness against Disturbances. To evaluate the robustness of our method against real-world disturbances, we test its reconstruction performance under three conditions: original clean videos, Gaussian noise, and texture jittering. As shown in Tab. 10, introducing these disturbances only slightly

Table 9: rFVD at different token budgets. Lower is better.

Method	768	890	1024
Fixed	20.0	17.3	15.0
LPIPS-based	18.5	16.0	13.5
Ours	**15.0**	**13.0**	**11.5**

Table 10: Router robustness under Gaussian noise and texture jitter.

Condition	Tokens↓	PSNR↑	LPIPS↓
Original	743	28.85	0.07
+ Gaussian Noise	802	28.59	0.08
+ Texture Jitter	815	28.20	0.08

increases the allocated token budget (from 743 to around 810). This slight increase is expected, as noise inherently adds spatial entropy. Meanwhile, all reconstruction metrics (e.g., PSNR and LPIPS) remain highly stable. This confirms that our codec-informed allocator maintains strong robustness in most disturbed scenarios.

Ablation Study about Components of Token Allocator. To validate the necessity of integrating both spatial and temporal entropy in our Token Allocator, we conduct an ablation study isolating these two components. As shown in Fig. 8, relying solely on either spatial or temporal complexity proves insufficient to independently characterize video complexity, resulting in degraded rFVD performance. By synergistically fusing both spatial and temporal estimations, CodecTok achieves the optimal rate-distortion balance, dynamically routing representational capacity to where it is most needed and yielding the best overall performance.

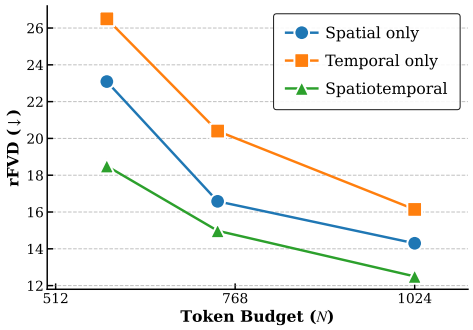


Figure 8: Ablation study on the measurements in allocator.

More Visualization Results

Reconstruction Visualization. As illustrated in Fig. 9 and Fig. 10, we present extensive qualitative reconstruction results of CodecTok on the UCF-101 test set. These visual comparisons display the original video sequences, our reconstructed outputs, and the corresponding error heatmaps. It is evident that CodecTok achieves near-lossless, high-fidelity reconstruction with almost zero perceivable error in the heatmaps, visually demonstrating the overwhelming superiority of our proposed method.

Generation Visualization. As depicted in Fig. 11, CodecTok demonstrates exceptional capability in synthesizing high-fidelity videos. The generated samples exhibit sharp spatial textures, highly coherent temporal dynamics, and robust structural integrity with minimal visual artifacts.

G Failure Cases

We qualitatively inspect the reconstruction results and token allocations of CodecTok. Under standard conditions, the model generally maintains high reconstruction quality. However, extreme texture jitter, strong noise, or abrupt color changes can distort the complexity estimates used by the token allocator. Such perturbations may cause the model to allocate excessive tokens to visual disturbances, reducing token efficiency and potentially degrading reconstruction quality.

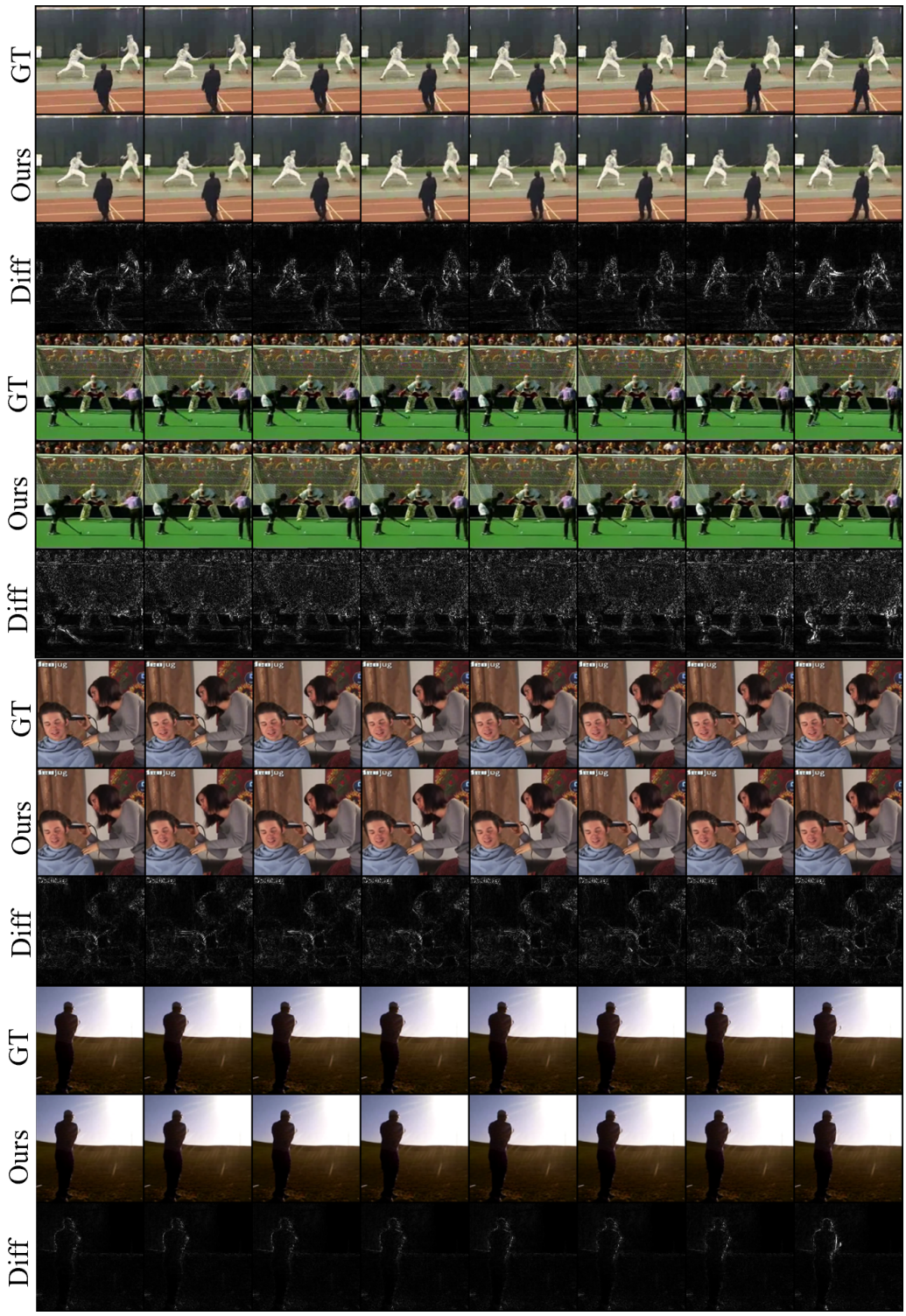


Figure 9: More reconstruction results on the UCF-101 dataset.

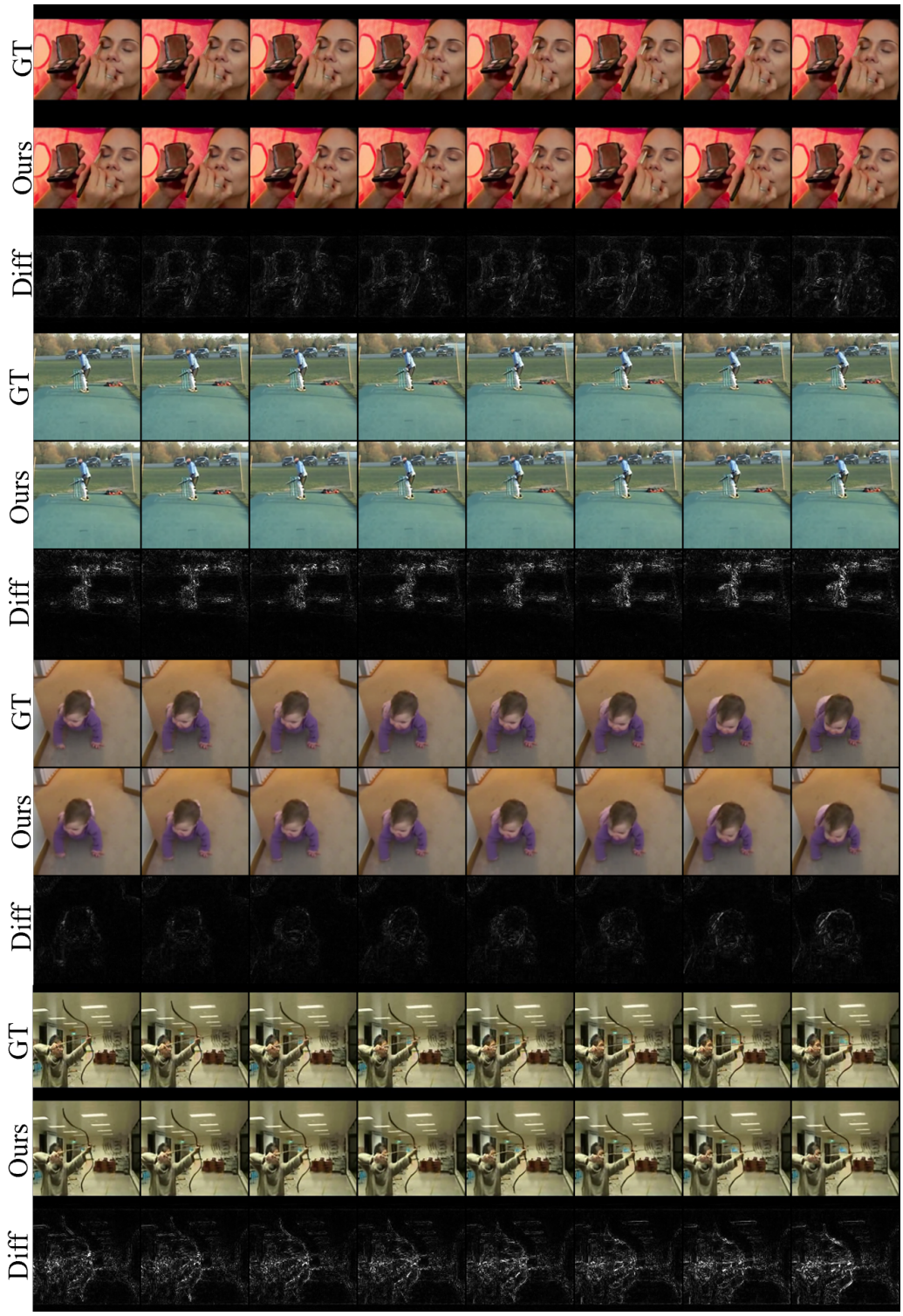


Figure 10: More reconstruction results on the UCF-101 dataset.



Figure 11: More generation results on the UCF-101 dataset.