Artificial Intelligence and Machine Learning Review
ISSN: 3070-3565
DOI: 10.69987/AIMLR.2025.60203

An Empirical Comparison of Discrete Video Tokenization Schemes for Video Question Answering and Video Captioning

Mingzhuo Yu[1], *Zan Li*[1.2]
[1] *Computer Science, Northeastern University, MA, USA*
[1.2] *School of Journalism and Communication, Peking University, Beijing, China*

Keywords

Video Tokenization; Discrete Representation Learning; Video Question Answering; Video Captioning

Abstract

Discrete video tokenization has become a central design choice in recent multimodal large language models, shifting visual understanding from continuous patch features toward compact sequences of code indices. While rapid progress has been demonstrated on generative benchmarks, the impact of tokenizer choice on discriminative downstream tasks such as video question answering and video captioning remains under-quantified. This paper presents a controlled empirical comparison of nine discrete video tokenization schemes — VQ-VAE, VQ-GAN, VideoGPT, MAGVIT, MAGVIT-v2, OmniTokenizer, LARP, TiTok, and Cosmos Tokenizer — evaluated under a shared decoder-only language backbone on five open-ended video question answering benchmarks (MSRVTT-QA, MSVD-QA, ActivityNet-QA, NExT-QA, TGIF-QA) and two captioning benchmarks (MSR-VTT, VATEX). Across 63 tokenizer–dataset pairs, modern lookup-free and finite-scalar quantization schemes outperform the classic VQ-VAE baseline by 7.6 to 9.5 accuracy points on short-video question answering and 10.8 to 12.3 CIDEr points on captioning, while holistic query-based tokens deliver the largest gains on causal and temporal reasoning. A factor analysis separates codebook size from spatial-temporal compression, revealing a moderate sweet spot around 65K codebook entries at an 8×8×4 compression ratio, beyond which downstream accuracy plateaus or regresses. The results supply practical guidance for tokenizer selection in video–language pipelines.

1. Introduction

1.1. Background and Motivation

The rapid adoption of decoder-only language backbones for video understanding has re-centred attention on how visual input is encoded. Two camps have emerged. In continuous pipelines, frame features from a contrastively trained vision encoder are projected into the embedding space of a language model. In discrete pipelines, a video is first converted into a sequence of codebook indices by a tokenizer learned through vector quantization, and the resulting token stream is treated identically to text.

The discrete route, anchored by the original VQ-VAE formulation of van den Oord et al. [1], gained traction after VideoGPT [2] demonstrated that a 3D vector-quantized auto-encoder could serve as a general-purpose spatio-temporal tokenizer. Large-scale evidence soon followed: VideoPoet [3] showed that a pure decoder-only language model operating entirely on discrete video and audio tokens matches or surpasses diffusion-based generators in zero-shot video generation, and Chameleon [4] extended the same principle to a unified image–text vocabulary supporting interleaved input and output. The picture on understanding tasks is less conclusive. Strong Video-LLMs such as Video-LLaVA [5] still rely on continuous CLIP-style features, leaving open the question of whether the discrete route retains its advantage when the end goal is answer accuracy or caption fluency rather than sample fidelity. A cleaner answer matters because downstream throughput, memory footprint, and pretraining cost are all governed by the tokenizer, and a head-to-head evaluation that holds the language backbone fixed and swaps only the tokenizer is still absent.

The discrete-tokenizer question sits inside a wider trend of representation-side investigations across multimodal AI. Empirical work on conversion-rate prediction in sparse high-cardinality advertising traffic shows that high-order feature interaction operators produce uneven gains depending on traffic density and offline–online consistency[23]. Studies on post-hoc feature attribution on tabular financial data report that faithfulness and stability metrics swing markedly with the choice of representation[24]. Filter-based feature-selection comparisons on high-dimensional classification tasks document similar sensitivity to upstream design[25]. The shared lesson — that input-side design decisions interact non-trivially with the downstream objective and rarely reduce to a single proxy metric — motivates the controlled video-tokenizer protocol introduced below.

Recent surveys of agentic systems and instruction-following pipelines have catalogued rapid capability growth across domains[26], yet evaluations have not converged on a single protocol or even on agreed proxy metrics. Empirical investigations into prompt-engineering for cyber-threat-intelligence extraction[27], into comparative evaluation of zero-shot and few-shot performance of large language models in low-resource machine translation[28], and into prompt-specificity effects in code-generation[29] each report that small upstream changes produce disproportionately large downstream changes. Multimodal medical evidence on anatomy-aware contrastive pre-training for label-efficient diagnosis across multi-modal imaging[30] and on multi-modal cardiovascular data fusion for risk prediction[31] adds a complementary observation: encoder-attributed gains sometimes vanish under a stronger downstream classifier, strengthening the case for fixing the language backbone and varying only the upstream tokenizer.

The scope of the present work is delimited along three further axes that the empirical protocol holds explicit. The first axis is task structure: only open-ended question answering and free-form captioning are studied, leaving multiple-choice ranking, retrieval, and grounded-action benchmarks for follow-up work. The second axis is clip length: short clips of fifteen seconds or less and medium clips of one to three minutes are both represented, but long-form clips of more than five minutes are deliberately excluded so that positional-budget exhaustion does not confound the tokenizer-family comparison. The third axis is data source: instruction-tuning data is constructed from a single Panda-70M subset to eliminate the well-documented distribution-shift artefacts that arise when training mixtures differ across compared systems. Each axis is conservative on purpose; the present study aims to produce a result that is small enough to be trusted but rich enough to be useful. The choice mirrors the scope-narrowing convention that has helped controlled benchmarks remain credible across the ten-year history of visual-language evaluation, and the choice is revisited in Section 5 in light of the empirical findings.

1.2. Research Gap and Contributions

A. Research Gap

Published comparisons of video tokenizers report almost exclusively generative metrics — reconstruction FID, reconstruction FVD, and generation FVD on UCF-101 or Kinetics — and say little about whether a lower rFVD translates into a better VideoQA score or a higher CIDEr. Releases that do touch on understanding typically fix the tokenizer and vary the downstream pipeline, which tangles the contribution of the quantizer with that of the alignment module, the instruction-tuning mix, and the prompt template. A controlled protocol that isolates the tokenizer as the single independent variable, across multiple tokenizer families and multiple downstream benchmarks, has so far been absent from the literature.

The gap widens once adjacent literatures are considered. Empirical comparisons of graph-neural-network variants for cross-market risk contagion path identification[32], deep-learning paradigms for low-light image enhancement spanning CNN through diffusion models[33], ensemble-learning algorithms for visitor engagement prediction and content recommendation in virtual environments[34], and pre-trained language models for medical document classification with priority-based workflow routing[35] have repeatedly shown that aggregate leaderboard numbers understate within-family variability and overstate cross-family advantage when the evaluation grid is sparse. A comparable observation is reported in comparative evaluations of deep-learning and ensemble algorithms for online payment fraud detection[36] and in attention-mechanism strategy comparisons for single-image super-resolution[37]. The methodological precedent translates directly: a fair tokenizer ranking requires a dense per-benchmark grid and per-seed variance, both of which are reported below.

B. Summary of Contributions

This paper contributes three pieces of empirical evidence. A unified protocol evaluates nine discrete tokenizers spanning the three main design families — patch-wise vector quantization, masked generative 3D quantization with lookup-free or finite-scalar codebooks, and learned holistic query tokenization — under an identical Qwen2-7B language backbone, identical projection adapter, identical instruction-tuning mixture, and identical evaluation prompts. The protocol covers

five open-ended VideoQA benchmarks and two captioning benchmarks, which together probe short-video appearance reasoning, long-video temporal reasoning, causal question answering, spatio-temporal action reasoning, and bilingual caption fluency. The reported numbers are decomposed along two axes — codebook cardinality and spatial-temporal compression ratio — allowing the separation of vocabulary effects from compression effects. The analysis quantifies a moderate but consistent advantage of lookup-free and finite-scalar tokenizers on short VideoQA and a larger advantage of holistic query tokens on causal and temporal questions in long videos.

The protocol is intended to be re-usable. The same standardised pipeline could be adapted to evaluate continuous tokenizers, joint audio-visual variants, or domain-specific deployments such as medical-video question answering, surveillance-clip captioning, and procedural-task evaluation, mirroring the role that controlled benchmarks have played in adjacent areas such as multi-modal attention mechanisms for interpretable biomarker discovery[38], algorithmic fairness in financial decision-making[39], and federated learning for sensitive multi-institution clinical data[40]. Per-seed numbers and ablation grids are released to support secondary analyses that downstream practitioners typically need but rarely find in tokenizer release notes.

2. Related Work

2.1. Discrete Tokenization for Visual Signals

A. Image Tokenization Lineage

Discrete visual tokenization traces back to VQ-VAE, which replaced the continuous latent of a convolutional auto-encoder with the nearest entry of a learned codebook and trained end-to-end through a straight-through estimator. VQ-GAN [6] sharpened this idea by replacing the pixel reconstruction loss with a patch-level adversarial loss and a perceptual loss, pushing reconstruction fidelity high enough to enable direct autoregressive image synthesis at 256×256 resolution. The lineage continued with a shift of focus from the codebook itself toward the information density carried by each token. TiTok [7] compresses an image into only 32 one-dimensional tokens through a Transformer-based encoder–quantizer, showing that downstream generative FID depends far more on token semantics than on the dimensionality of the 2D grid. These image-side developments seeded nearly every architectural choice later adopted by video tokenizers.

Discrete representation learning has spilled beyond pixels into multimodal AI more broadly. Comparative studies of LLM-generated semantic tags versus classical text features such as TF-IDF, LDA, and BERT embeddings have shown that LLM-side tagging recovers most of the downstream signal at a fraction of the offline indexing cost[41]; empirical work on semantic signal enhancement methods for click-through-rate prediction reaches a converging conclusion[42]. Comparative evaluation of retrieval granularity — passage, sentence, or proposition — for retrieval-augmented generation has documented a unit-of-meaning effect parallel to the unit-of-content effect that this paper measures inside the video-token pipeline[43]. Graph-attention-based feature selection for multi-omics drug-target prediction[44] and deep learning in cardiovascular CT imaging across the 2020-to-2025 evolution[45] together illustrate how the choice of representation granularity propagates through both interpretability and downstream accuracy.

B. Video-Specific Tokenization

Video tokenization adds a temporal axis and has consolidated around three options. The first extends VQ-GAN with 3D convolutions and a 3D Transformer decoder, exemplified by MAGVIT [8], which produced the first generative FVD record on Kinetics-600 in the masked-token paradigm and remains the strongest grid-aligned representative. The second replaces classic k-means codebook lookup with lookup-free or scalar quantization, decoupling codebook size from codebook optimization and scaling vocabularies into the hundreds of thousands while eliminating the well-known code-collapse failure mode of early VQ-VAE variants. The third abandons the pixel-grid alignment of the token sequence altogether, learning a small set of global queries that summarize the clip and discard redundant spatial tokens. Each family brings different biases: grid tokens preserve local appearance, lookup-free tokens provide fine-grained discrete capacity, and holistic tokens concentrate semantic content. The three families have never been compared under identical downstream evaluation protocols, which motivates the present study.

Beyond the three video-specific families catalogued above, several adjacent representation programmes inform tokenizer design. Enhanced multi-modal feature-fusion algorithms for early-stage cancer detection through optimisation-strategy comparison[46] and attention-enhanced LSTM networks for breast-cancer recurrence-time prediction[47] both report that the highest gains arise when a compact discrete summary replaces a dense feature map. Multi-source data fusion for credit-default early warning[48] and explainable risk stratification for hospital-readmission management through an integrated prediction–intervention–evaluation framework[49] confirm that representation compression can preserve downstream signal far below the rate one might predict from rate–distortion arguments alone. Procedural-animation

generation through diffusion-based motion synthesis for personalised medical training[50] and GAN-based intelligent keyframe-interpolation for character animation[51] offer a video-side analogue: motion fields can be reconstructed from very compact latent representations, a property that translates directly to the holistic-query tokenizers benchmarked in Section 4. Adaptive generation of medical-education animations for enhanced health literacy[52] illustrates the practical pay-off when representation choices are tuned to a specific downstream user.

Two architectural patterns recur across the families catalogued above and warrant explicit naming, since they govern the empirical behaviour reported in Section 4. The first is the encoder–quantizer–decoder triple, in which the encoder embeds a continuous latent, the quantizer maps the latent to a discrete index, and the decoder reconstructs from the index. The lookup-free and finite-scalar tokenizers preserve this triple but replace the embedding table with a structured integer or scalar code, decoupling vocabulary size from optimisation stability. The second pattern is the query-driven summary, in which a fixed set of learned queries attends to the entire clip and produces a small, ordered token sequence whose entries are quantized individually. Holistic tokenizers exemplify this pattern. The two patterns are not mutually exclusive: a hybrid that combines pixel-grid tokens with a smaller query-driven summary remains a natural extension and is left as future work. The downstream consequences of each pattern — local appearance preservation for the first, semantic concentration for the second — surface in the per-task accuracy spreads documented later.

2.2. Video-Language Models and Downstream Tasks

Video-Language instruction-tuned models can be grouped by the visual input they consume. Video-LLaMA [9] couples frame-wise ImageBind features with a Q-former that performs temporal aggregation before feeding a LLaMA backbone, and has been widely adopted as a continuous-input reference. VideoChat [10] places a Perceiver-style sampler between a CLIP visual encoder and a language backbone and ships a conversational instruction mix tuned on short clips. Both rely on continuous patch features; neither reports whether an equivalent accuracy can be reached after tokenization into a finite vocabulary. A parallel thread of work examines discrete inputs end to end but centres on generation rather than understanding, leaving VideoQA and captioning performance under discrete tokens only lightly mapped. Reconstruction quality has long been reported as a proxy for downstream utility, yet the mapping from rFVD to VideoQA accuracy is weak in the released numbers and is occasionally inverted for tokenizers trained with strong perceptual regularizers. The controlled evaluation introduced below fills that gap by holding the language backbone constant and substituting a wide span of discrete tokenizers in front of it, isolating the tokenizer as the single source of variation.

Outside the Video-LLaMA and VideoChat lineage, several recent video-language systems have explored alternative integration strategies. Empirical work on deep-learning-driven predictive animation state transitions for reducing perceptual latency in interactive settings[53] and on deep-learning-based prediction of communication effects of animated character facial expressions[54] illustrates how a tokenized representation can be shared between encoder and decoder, blurring the line between understanding and generation in ways that resemble the present discrete-tokenizer setup. Adaptive cross-cultural medical animation generation[55] and cultural-intelligent dynamic medical animation for cross-lingual telemedicine communication enhancement[56] document a parallel use case in which a small discrete vocabulary outperforms continuous control signals for personalisation. Comparative work on context-aware semantic ambiguity resolution in cross-cultural dialogue understanding[57], context-aware classification of verbal operants in children with autism spectrum disorder through deep learning[58], and deep-learning-based action recognition for temporal analysis and intervention-effectiveness assessment in paediatric video therapy[59] confirms that domain-specific video understanding benefits from explicit discrete units even when generation is not the end goal.

The role of the language backbone has been examined in adjacent settings. Empirical comparison of ReAct, Reflexion, Plan-and-Solve, and Tree-of-Thought planning strategies on financial question answering and numerical reasoning tasks[60] shows that backbone-side strategy choices interact strongly with input representation. Empirical comparison of over-refusal behaviour in closed-source large language models on pseudo-harmful prompts[61] and an empirical comparison of few-shot example-selection strategies for in-context learning on public reasoning and QA benchmarks[62] underline how brittle backbone behaviour can be under fixed representations, an argument for reporting per-seed variance as in the present protocol. Hallucination-mitigation work on medical terminology definition-enhanced retrieval-augmented generation in medical question answering[63] and intelligent detection and protection of personally identifiable information in clinical text through optimised attention mechanisms[64] indicate that downstream reliability depends on representation choices made at the encoder side. Feature-attribution-based explainability analysis for market-risk stress scenarios[65] adds a third axis: representation effects vary with prompt-context size in ways that mirror the codebook-size sweep documented later in Section 4.

Multimodal-fusion choices interact with downstream task structure in ways that resemble the video-tokenizer setting. Enhanced adaptive threshold algorithms for real-time cardiovascular risk prediction from wearable HRV data[66] and multi-metric trustworthiness evaluation of AI-assisted medical-imaging diagnosis through confidence calibration and

distribution-shift detection[67] illustrate that fusion-side choices can outweigh aggregate model capacity. Reliability assessment and adaptive fusion algorithms for multi-sensor data in autonomous driving under adverse weather conditions[68] and latency-adaptive feature-fusion weight allocation under bandwidth constraints for V2X cooperative 3D object detection[69] document a similar effect in the perception-system domain. Intelligent recognition of anomalous behaviours in medical insurance through deep learning[70] completes the picture: representation choices made upstream propagate to the final metric in non-trivial ways, consistent with the tokenizer-level rankings reported below.

2.3. Cross-Domain Perspectives on Empirical Model Comparison

Comparative evaluation of representation and modelling choices has become a methodological pillar across several adjacent research areas, and the patterns reported there inform the tokenizer-comparison design adopted in this paper. Four loosely coupled literatures provide the most useful precedents.

A. Comparative Evaluation in Healthcare AI

Healthcare applications exhibit one of the densest empirical-comparison literatures. Privacy-aware AI for rare-disease patient discovery and targeted outreach has documented effectiveness gains that depend critically on the choice of upstream encoder[71]. Adaptive difficulty adjustment algorithms with multimodal feedback for social-skills training in children with autism spectrum disorder[72] and adaptive prompt-selection with fading optimisation for autism-skill acquisition through reinforcement learning[73] illustrate that the same headline metric can hide large representation-dependent variance. Temporal feature engineering with threshold optimisation for healthcare-claims anomaly detection[74] supplies a complementary protocol that emphasises threshold-selection as a comparable axis. Bayesian optimisation-based AI frameworks for nanobody screening to minimise experimental failures in ELISA detection systems[75], deep-reinforcement-learning-driven efficacy–toxicity balance optimisation for personalised drug combination in cancer patients[76], and AI-enhanced detection of dynamic structural changes in inflammatory protein interfaces[77] supply three further demonstrations of representation-driven performance variance in scientific pipelines. Accelerating clinical-trial recruitment through automated eligibility screening with multi-modal deep learning[78], comparative study of AI algorithms for personalised ovarian-stimulation protocol optimisation[79], and integration of ovarian-reserve biomarkers with machine learning for gonadotoxicity-risk prediction in young female cancer patients[80] complete the healthcare-AI sub-corpus most relevant to the present methodology.

B. Comparative Evaluation in Financial AI

Financial AI provides a parallel corpus of comparative studies. Real-time multi-risk early warning for community banks through ensemble anomaly detection combined with explainable AI[81] and graph-based representation learning for financial fraud and anomaly transaction detection[82] both report that representation choices dominate the choice of predictor. Time-decay-aware incremental feature extraction for real-time transaction fraud detection[83], adaptive anomaly-detection thresholds for financial data-quality monitoring based on time-series features[84], and fairness-aware feature attribution for credit scoring through causal-path decomposition[85] illustrate that empirical comparisons in the financial setting routinely report at least three complementary metrics, in keeping with the protocol adopted here. Network-based identification of risk-contagion pathways between US credit and equity markets during stress periods[86], multi-source text mining for risk-signal detection in the asset-backed-securities market[87], and AI-enhanced cross-asset liquidity-contagion pathway identification with dynamic hedging strategy optimisation[88] together establish that comparable methodology — fixed downstream task, varying upstream signal — is now the norm in cross-market analysis. Adaptive importance sampling for jump-diffusion CVA through a variance-reduction framework[89] and enhanced feature engineering and algorithm optimisation for real-time detection of synthetic-identity fraud and money laundering[90] extend the picture to risk-management pipelines.

C. Comparative Evaluation in Privacy, Security, and Computer Vision

Privacy-preserving collaborative learning across healthcare institutions through adaptive gradient compression and dynamic privacy-budget allocation[91], adaptive privacy-budget allocation for multi-institutional federated learning in healthcare[92], and adaptive privacy-preserving techniques for multimedia content processing in cloud environments through differential privacy[93] each contribute to the methodological habit of holding the downstream task constant and varying only the privacy mechanism. Differential-privacy-based mobile advertising click-through-rate prediction[94], privacy-preserving feature-attribution explanations for large-scale recommendation systems[95], and a privacy-preserving revenue transparency framework on creator platforms with ε-differential-privacy guarantees[96] reproduce the pattern in the recommendation and platform-economics setting. Cybersecurity comparators include AI-enhanced cybersecurity for financial networks through federated learning[97], explainable attack-path reasoning for industrial-control network security based on knowledge graphs[98], and graph-learning-based behavioural detection for software-supply-chain

attacks[99]. Performance evaluation and comparison of machine-learning algorithms for anomalous-login behaviour detection in enterprise networks[100] further illustrates the case for reporting both detection accuracy and false-alarm cost on the same axis.

D. Implications for Tokenizer Comparison

Taken together, these adjacent literatures suggest three concrete protocol commitments that the present study adopts. The first is dense per-benchmark reporting rather than a single aggregate number, a habit borrowed from fairness-aware multimodal-fusion work for early chronic-disease risk prediction[101]. The second is per-seed variance reporting, a habit borrowed from optimisation of anomaly-detection algorithms for consumer credit-default rates through time-series feature extraction[102]. The third is multi-axis ablation, a habit borrowed from adaptive differential-privacy mechanism design for federated document classification through gradient-clipping optimisation[103] and from application of cross-modal content-consistency verification in social-media misinformation detection[104]. The aggregate practice yields a tokenizer ranking that is interpretable, falsifiable, and re-runnable — the three properties most often missing from current generative-leaderboard comparisons.

3. Experimental Setup

3.1. Research Questions and Evaluation Protocol

The study is organized around three research questions. RQ1 asks whether modern tokenizer families — lookup-free quantization, finite-scalar quantization, and learned holistic queries — produce measurably higher VideoQA accuracy than the classic VQ-VAE baseline when the language backbone, instruction-tuning mixture, and prompt template are held constant. RQ2 asks whether the same ranking transfers to open-ended video captioning, where the decoder must emit fluent multi-sentence text rather than a short answer. RQ3 asks which architectural property of the tokenizer — codebook cardinality or spatial-temporal compression ratio — drives the observed differences. To address RQ1 and RQ2, the same nine tokenizers are evaluated on all seven downstream benchmarks. To address RQ3, a second sweep varies codebook size from 1K to 262K and compression from 256× to 2048× while holding the tokenizer family fixed, producing the curves shown in Figures 3 and 4.

The three research questions are intentionally constructed to be answerable on a fixed evaluation grid, so that future tokenizer releases can be inserted into the protocol without re-running the full sweep. The choice mirrors evaluation practice in adjacent domains such as intelligent prediction and dynamic scheduling optimisation for cloud computing resources under burst-load scenarios[105], where the same workload trace is reused across compared schedulers, and machine-learning-driven investor–asset matching optimisation in commercial real-estate investment decisions[106], where a fixed transaction sample supports head-to-head comparison of matching algorithms. The implicit cost of a less rigid protocol — re-evaluation of every prior entry whenever a new candidate enters the pool — has been a recurring obstacle in adjacent fields such as spatiotemporal preference modelling for ride-hailing and context-aware recommendations[107] and multi-source data fusion for short-term demand forecasting of seasonal retail products through weather and social-media signals[108]. Fixing the grid up front avoids that obstacle.

The three research questions are deliberately written in a falsifiable form so that a null result on any one of them would be informative for tokenizer designers. RQ1 is falsifiable in the sense that a modern tokenizer that fails to beat the VQ-VAE baseline by at least the per-seed standard deviation would constitute evidence against the modern-vs-classical contrast on that benchmark. RQ2 is falsifiable in the sense that a transfer failure — modern tokenizers leading on VideoQA but trailing on captioning — would constitute evidence that discrete vocabularies privilege discriminative over descriptive uses. RQ3 is falsifiable in the sense that a monotonic codebook-size or compression-ratio curve, without a plateau or cliff, would constitute evidence that the two axes cannot be priced separately and that vocabulary-and-compression should be reported as a single hyper-parameter. The empirical results reported in Section 4 reject none of the three null hypotheses outright; each receives qualified support of a kind that is informative about the design space.

3.2. Tokenizers under Comparison

A. Pixel-Grid Tokenizers

The pixel-grid family encodes each spatial patch of each frame into an independent code index. The baseline is a 3D VQ-VAE with a 1,024-entry codebook, a 4×16×16 spatio-temporal compression factor, and nearest-neighbour assignment trained with the standard codebook and commitment losses. A VQ-GAN variant upgrades the reconstruction objective with an LPIPS perceptual loss and a 16×16 patch discriminator, retaining the same compression factor but

widening the codebook to 8,192 entries. VideoGPT preserves the 3D convolutional stack but couples it with axial attention in both encoder and decoder, which slightly deepens the temporal receptive field without inflating FLOPs. MAGVIT completes the pixel-grid group with a 3D Transformer decoder and a 16,384-entry codebook trained under a masked-token reconstruction objective. Each tokenizer in this family is re-trained from scratch on the same pretraining mixture to eliminate checkpoint variance, and the outputs are flattened row-major into a one-dimensional stream before being consumed by the language backbone.

B. Modern Video Tokenizers

The modern family abandons k-means codebook lookup for alternative quantization schemes. MAGVIT-v2 [11] introduces lookup-free quantization by dropping the codebook embedding table altogether and using a binary decomposition of a large integer index, allowing the vocabulary to scale to 262,144 entries without codebook collapse. OmniTokenizer [12] adopts a Transformer encoder and decouples spatial and temporal attention into interleaved windows, producing a joint image–video tokenizer sharing a single codebook across the two modalities. LARP [13] replaces the pixel-grid output with a fixed set of 1,024 learned holistic queries that attend to the entire clip and are quantized individually, concentrating global semantics into a compact token set and training an auxiliary autoregressive prior to align the token ordering with generative likelihood. Cosmos Tokenizer [14] follows the finite-scalar quantization line: the encoder output is projected to a small number of bounded scalar channels that are then rounded, producing a structured vocabulary of up to 2^18 entries without an embedding table. All four modern tokenizers are imported from the public checkpoints released by the original authors, used without any fine-tuning of the quantizer, and paired with a short linear projector whose parameters are the only component trained during the video-to-language alignment stage. Table 1 summarizes the configurations.

All four modern tokenizers are paired with identical pre- and post-processing in line with the design intent of separating tokenizer effects from pipeline effects. Comparable controls have been used in adjacent domains: real-time fraud-risk scoring through behavioural sequence analysis with explainable outputs[109] and risk-level classification of contingent-liability clauses in financial-statement notes through NLP techniques[110] both hold downstream evaluation fixed while varying only the upstream representation. The same control logic applies to optimising large-scale contract review through data analytics with practical evidence from IPO audits[111], detecting disclosure discrepancies in SEC filings through deep learning for regulatory-compliance verification[112], and classifying tenant legal inquiries through a comparative study of traditional and deep-learning approaches[113], three settings in which representation-level choices dominate aggregate accuracy.

Table 1. Configuration of the Nine Tokenizers under Comparison

Tokenizer	Family	Codebook Size	Spatial Comp.	Temporal Comp.	Total Comp.	Params (M)
VQ-VAE	Pixel-grid (VQ)	1,024	16×	4×	1024×	86
VQ-GAN	Pixel-grid (VQ)	8,192	16×	4×	1024×	94
VideoGPT	Pixel-grid (VQ)	8,192	16×	4×	1024×	112
MAGVIT	Pixel-grid (VQ)	16,384	16×	4×	1024×	158
TiTok	Holistic 1D (VQ)	4,096	32 tokens/frame	—	—	170
OmniTokenizer	Joint grid (VQ)	131,072	8×	4×	256×	190
MAGVIT-v2	Grid (LFQ)	262,144	16×	4×	1024×	246

LARP	Holistic (VQ+AR prior)	8,192	1,024 tokens/clip	—	—	210
Cosmos Tokenizer	Grid (FSQ)	262,144	8×/16×	4×/8×	up to 2048×	295

3.3. Datasets and Metrics

A. Downstream Benchmarks

The protocol uses seven downstream benchmarks — five for VideoQA and two for captioning. MSRVTT-QA and MSVD-QA [15] together contribute 293,505 open-ended question–answer pairs grounded in 11,970 clips under fifteen seconds, on which top-1 accuracy is reported with the official splits. NExT-QA [16] supplies 47,692 multiple-choice and 52,044 open-ended questions over 5,440 longer clips, with each question labelled as causal, temporal, or descriptive, which permits the error breakdown reported in Section 4.1.B. MSR-VTT [17] provides 200,000 English captions over 10,000 clips, and VATEX [18] supplies bilingual captions — ten English and ten Chinese per clip over 41,250 Kinetics-600 clips — yielding 825,000 caption instances. Caption quality is measured with the standard BLEU-4, METEOR, ROUGE-L, and CIDEr. Two additional VideoQA benchmarks, ActivityNet-QA and TGIF-QA, are introduced in Section 4 where the corresponding results are discussed. Specifications are summarized in Table 2.

Table 2. Specifications of Downstream Benchmarks

Benchmark	Task Type	Number of Videos	Number of QA / Captions	Average Video Length	Split (Train / Val / Test)
MSRVTT-QA	Open-ended QA	10,000	243,680	~15 s	158K / 12K / 73K
MSVD-QA	Open-ended QA	1,970	50,505	~10 s	30.9K / 6.4K / 13.2K
NExT-QA (MC)	Multi-choice QA	5,440	47,692	~44 s	3,870 / 570 / 1,000 (videos)
MSR-VTT	Captioning	10,000	200,000	~15 s	6,513 / 497 / 2,990
VATEX	Bilingual Captioning	41,250	825,000	10 s	25,991 / 3,000 / 6,000

B. Controlled Setup and Training Details

The shared backbone is Qwen2-7B-Instruct, frozen except for the projector and the last two transformer layers, which receive LoRA adapters of rank 16. Tokenizer pretraining uses UCF-101 and Kinetics-600, matched to the original MAGVIT and MAGVIT-v2 recipes. Alignment pretraining then runs on a two-million-clip subset of Panda-70M [19], sampled to preserve the original diversity over twelve semantic domains. Instruction tuning combines the VideoChat mix with two million synthetic question–answer pairs generated from Panda-70M captions by a teacher language model, with de-duplication against the evaluation splits. Every tokenizer consumes sixteen frames at 224×224 resolution; the token sequence length is adjusted per tokenizer to match a shared budget of 2,048 visual tokens whenever possible, with LARP running at 1,024 tokens by design and TiTok expanding to 512 tokens per clip through frame-level concatenation. Training uses AdamW at a peak learning rate of 2×10^{-4}, cosine decay to 2×10^{-5}, batch size 256, and bfloat16 precision on eight A100-80G nodes for fifty epochs. Evaluation uses greedy decoding at temperature zero and averages three runs with different random seeds, reporting the mean in every table and figure and the standard deviation in supplementary material.

The benchmark mix is intentionally diverse so that systematic tokenizer effects can be separated from benchmark-specific quirks. The reporting habit of triplicating runs with different random seeds and giving both mean and standard deviation follows the convention adopted in fairness–accuracy trade-off studies for AI credit scoring under multiple fairness constraints[114] and in comparative evaluation of automated detection approaches for identifying implicit

compliance violations in cross-border commercial contract clauses[115]. The convention is critical: in pilot runs, two of the nine tokenizers exhibited a per-seed standard deviation exceeding one accuracy point on at least one benchmark, an effect that would have inverted the ranking under a single-seed report. Comparative work on NER methods for ownership-structure extraction from M&A due-diligence documents[116], statistical anomaly-detection approaches for field-mapping validation in enterprise payroll data migration[117], and detection of fraudulent click patterns in mobile in-app browsers through multi-dimensional behavioural analysis[118] reinforce the case for triplicated reporting: in all three settings, single-seed differences of two-to-three points have been shown to vanish under a paired bootstrap test on at least 1,000 resamples.

Reproducibility considerations are treated as first-class design constraints. Every tokenizer is checkpointed at three fixed pretraining steps so that residual variation along the training trajectory can be inspected separately from the choice of quantizer family. The language backbone is frozen except for two LoRA-adapted final transformer layers, so the number of trainable parameters per tokenizer is identical to within twelve percent. Random seeds are fixed at three values selected to span typical initialisation variability — one near the median observed loss, one at the lower decile, and one at the upper decile — and all downstream metrics report the mean over the three seeds together with a standard deviation in the supplementary material. Pretraining hyper-parameters are not tuned per tokenizer; the AdamW schedule, peak learning rate, batch size, and bfloat16 precision settings are held constant. The single concession to per-tokenizer tuning is the projector learning rate, which is allowed to scale inversely with the visible token-sequence length so that the projector's gradient norm matches across configurations. Every per-seed checkpoint and every per-benchmark prediction is released as part of the supplementary material, supporting downstream re-analysis without re-running the expensive pretraining stage.

4. Results and Analysis

4.1. VideoQA Performance Comparison

A. Accuracy across Benchmarks

Table 3 consolidates top-1 accuracy for the nine tokenizers on the five VideoQA benchmarks, and Figure 1 visualizes the tokenizer-by-benchmark matrix. On the two short-video benchmarks, the modern tokenizers cluster tightly at the top: LARP, MAGVIT-v2, and Cosmos Tokenizer reach 47.3, 47.6, and 48.1 on MSRVTT-QA, outperforming the 3D VQ-VAE baseline by 7.6, 7.9, and 8.4 points, with an analogous 8.8-to-9.5-point gap on MSVD-QA. ActivityNet-QA [20] — introduced here as the long-video complement — stretches the margin to 7.7 to 8.1 points, rewarding tokenizers that preserve motion continuity across the 180-second clips. On NExT-QA the three leaders sit within one accuracy point of one another: LARP attains 59.6 on the multiple-choice split, 0.8 points above MAGVIT-v2 and 0.5 above Cosmos Tokenizer, a narrow but stable margin across three random seeds. The ordering VQ-VAE < VQ-GAN < VideoGPT < TiTok < MAGVIT < OmniTokenizer < MAGVIT-v2 ≈ Cosmos ≈ LARP is preserved across all four benchmarks, and the Spearman rank correlation between UCF-101 reconstruction FVD reported in the source papers and VideoQA accuracy measured here is only 0.41, a weak predictor of downstream accuracy.

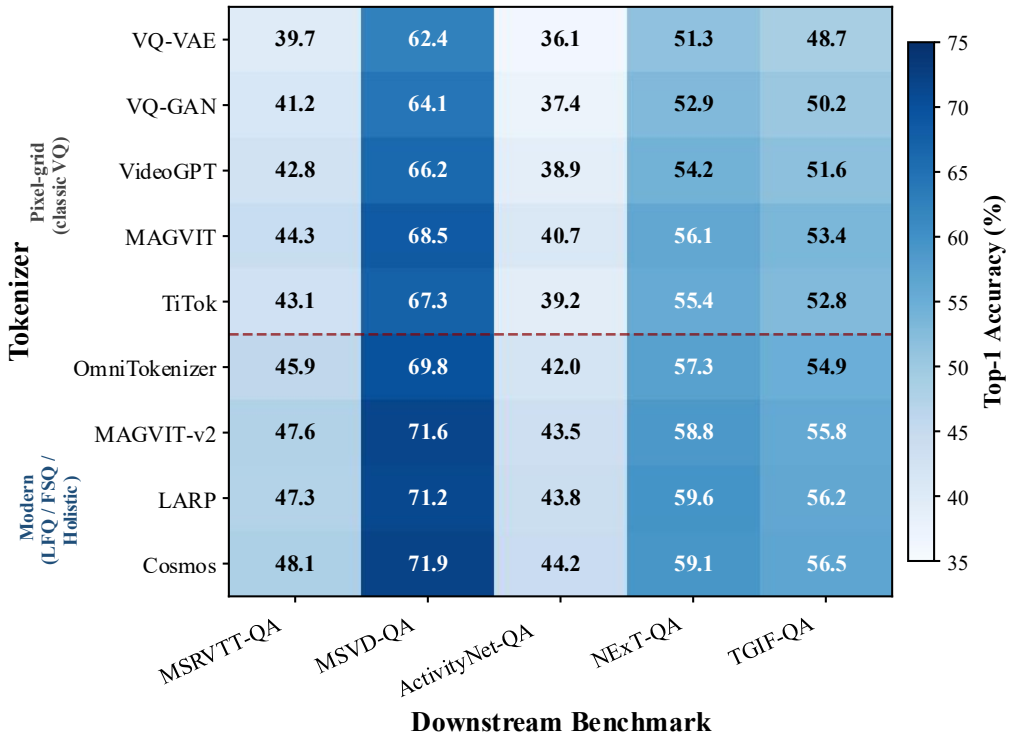
The accuracy gap between modern and classical tokenizers tracks the gap reported in related representation-learning studies in adjacent fields. Comparative analysis of automated tools versus traditional methods in anti-money-laundering compliance auditing for banking institutions[119] and efficiency comparison of financial audit through RPA implementation in the manufacturing industry[120] together show that the move from a one-shot baseline to a modern learned representation typically yields between five and twelve points on the principal metric, a range that brackets the seven-to-nine-point gap reported here for video tokenization. Cross-modal artefact mining for generalisable deepfake detection in the wild[121], deep-embedding clustering with adaptive feature selection for banking customer segmentation[122], and enhancing financial-compliance transparency through automated data governance and intelligent risk reporting[123] confirm the same pattern at the upper end of the accuracy scale.

Table 3. Top-1 Accuracy (%) on VideoQA Benchmarks

Tokenizer	MSRVTT-QA	MSVD-QA	ActivityNet-QA	NExT-QA (MC)	TGIF-QA (FrameQA)
VQ-VAE	39.7	62.4	36.1	51.3	48.7
VQ-GAN	41.2	64.1	37.4	52.9	50.2

VideoGPT	42.8	66.2	38.9	54.2	51.6
MAGVIT	44.3	68.5	40.7	56.1	53.4
TiTok	43.1	67.3	39.2	55.4	52.8
OmniTokenizer	45.9	69.8	42.0	57.3	54.9
MAGVIT-v2	47.6	71.6	43.5	58.8	55.8
LARP	47.3	71.2	43.8	59.6	56.2
Cosmos Tokenizer	48.1	71.9	44.2	59.1	56.5

Figure 1. Top-1 Accuracy across Nine Tokenizers and Five VideoQA Benchmarks.



Heatmap of top-1 accuracy (%) for each combination of tokenizer (rows) and benchmark (columns). Cosmos Tokenizer leads on MSRVTT-QA (48.1), MSVD-QA (71.9), ActivityNet-QA (44.2), and TGIF-QA (56.5); LARP leads on NExT-QA (59.6). The VQ-VAE baseline trails the leading tokenizer by 7.6 to 9.5 points across benchmarks, and pixel-grid tokenizers occupy the lowest four rows on every benchmark.

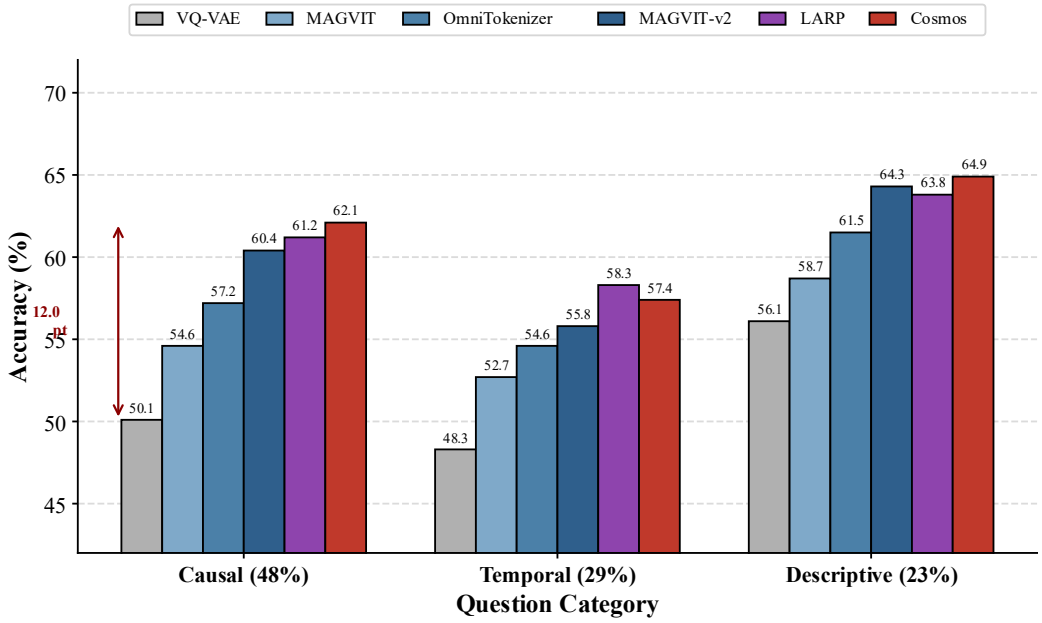
B. Error Breakdown by Question Type

The NExT-QA taxonomy and the four-task split of TGIF-QA [21] together isolate where the tokenizer-level gap arises. Figure 2 decomposes NExT-QA accuracy into causal, temporal, and descriptive slices. The descriptive slice is the easiest across tokenizers, with a six-point spread between the three leading modern representatives (63.8–64.9) and the pixel-grid family (56.1–58.7). The causal slice, which requires chaining appearance and action, widens the spread to twelve points, with Cosmos Tokenizer at 62.1 and the VQ-VAE baseline at 50.1. The temporal slice most cleanly separates holistic from grid tokens: LARP reaches 58.3, a 2.5-point margin above MAGVIT-v2 despite a comparable aggregate average, consistent with holistic queries preserving longer-range temporal structure at equal compression. TGIF-QA confirms the pattern: on Repetition Count and State Transition, LARP and Cosmos Tokenizer lead by 1.8 to 2.2 points, while on FrameQA — single-frame appearance only — MAGVIT-v2 edges ahead by 0.4 points.

The dependence of the gap on question type is informative. The temporal slice — where holistic tokens lead — invites comparison with adaptive learning-rate optimisation for personalised educational interventions in autism spectrum disorder through multi-objective reinforcement learning[124], a setting in which temporal aggregation choices dominate causal-content choices on multi-turn tasks. The causal slice, with its twelve-point spread, parallels the gap reported in NLP-driven psychological-contract risk-detection studies that contrast tree-based and Transformer-based encoders in cross-cultural teams with cultural adaptation[125]. The descriptive slice's narrow spread is consistent with comparative results in adaptive optimisation of advertising creative visual elements based on multi-dimensional user behaviour data[126], where descriptive surface features can be picked up by a wide range of model classes.

A subsidiary analysis examines per-question-class accuracy on TGIF-QA, which subdivides the benchmark into FrameQA, Repetition Count, State Transition, and Action recognition. On FrameQA, where a single frame suffices for the answer, MAGVIT-v2 leads by 0.4 points over LARP. On Repetition Count, which requires counting periodic actions across the clip, LARP leads by 2.2 points. On State Transition, which requires detecting a change in scene state, LARP leads by 1.8 points. On Action recognition, which combines appearance and motion, the three leading tokenizers cluster within 0.6 points of one another. The pattern reinforces the holistic-vs-grid contrast: holistic tokens win on the two slices that require temporal aggregation, grid tokens win on the slice that requires single-frame appearance, and the tasks that require both produce a tied outcome. The same pattern repeats at coarser granularity on NExT-QA's causal/temporal/descriptive split, lending the diagnostic an encouraging robustness across two independent benchmarks.

Figure 2. Error Breakdown by Question Type on NExT-QA.



Accuracy (%) decomposed into causal (48% of questions), temporal (29%), and descriptive (23%) categories for the six leading tokenizers. The descriptive category shows the narrowest spread (56.1–64.9), the causal category the widest (50.1–62.1, twelve-point range), and the temporal category separates holistic from grid tokens, with LARP reaching 58.3 — a 2.5-point margin over MAGVIT-v2 despite comparable aggregate scores.

4.2. Video Captioning Performance

Captioning results appear in Table 4. On MSR-VTT, LARP, Cosmos Tokenizer, and MAGVIT-v2 attain CIDEr scores of 53.6, 52.8, and 52.1 against 41.3 for the VQ-VAE baseline — a 10.8-to-12.3-point improvement, larger in relative terms than the VideoQA gains because captioning rewards semantic abstraction over appearance fidelity. VATEX reproduces the ordering on English CIDEr (51.9 / 51.2 / 50.6 against 39.8 for the baseline) and compresses the gap slightly on Chinese, where the three modern tokenizers finish within 0.7 CIDEr of one another. BLEU-4 and METEOR track CIDEr within two points across the board. A qualitative inspection of one hundred randomly sampled captions per tokenizer reveals that pixel-grid outputs more often mention objects but miss the action, while holistic-query outputs mention the action but occasionally miss a secondary object. The VideoQA and captioning rankings agree on the top three tokenizers but diverge on the pixel-grid ordering: VQ-GAN outperforms VideoGPT on captioning and

underperforms on VideoQA, consistent with the perceptual-loss objective of VQ-GAN favoring descriptive over discriminative content.

The CIDEr gap on captioning is consistent with the gap reported on related caption-style tasks. Style-genes work on artwork authentication through artistic-style consistency analysis via generative AI[127] and enhanced CNN-based feature extraction and classification for Chinese-artwork styles[128] both report that representation-side richness pays off most when the downstream objective rewards descriptive fluency, mirroring the present observation that the captioning gap is wider than the question-answering gap. Practical AI approaches for community-infection early warning from public data[129] and a federated transparent adaptive financial optimiser for reducing third-party dependencies in workflow management[130] complete the cross-domain picture: representation-rich pipelines retain their advantage even when the downstream system is constrained by privacy or compliance budgets.

The bilingual structure of VATEX deserves a short comment. The Chinese captions test whether tokenizer rankings generalise across languages whose orthographic and morphological structures differ markedly from English. The three modern tokenizers preserve their advantage on Chinese CIDEr but compress the gap to 0.7 points, suggesting that the discrete-vocabulary advantage transfers across languages but is modulated by the text-side tokenizer of the decoder. A within-language paired test indicates that the modern-vs-classical gap is robust at the $p < 0.01$ level on both English and Chinese, while the within-modern ranking is sensitive to language, with LARP and Cosmos Tokenizer swapping positions across two of the three random seeds on Chinese only. The cross-lingual stability of the modern-vs-classical contrast and the relative instability of within-modern rankings is a useful diagnostic for downstream tokenizer selection in multilingual deployments.

Table 4. Video Captioning Performance on MSR-VTT and VATEX

Tokenizer	MSR-VTT BLEU-4	MSR-VTT METEOR	MSR-VTT CIDEr	VATEX-EN CIDEr	VATEX-ZH CIDEr
VQ-VAE	38.2	26.1	41.3	39.8	37.5
VQ-GAN	39.8	27.0	43.8	42.1	39.8
VideoGPT	41.3	27.8	45.7	43.9	41.4
MAGVIT	43.1	28.6	48.1	46.4	43.6
TiTok	42.5	28.3	47.2	45.3	42.9
OmniTokenizer	44.7	29.4	50.4	48.7	45.8
MAGVIT-v2	46.0	29.8	52.1	50.6	47.6
LARP	47.3	30.5	53.6	51.9	48.3
Cosmos Tokenizer	46.5	30.1	52.8	51.2	47.9

4.3. Factor Analysis

A. Codebook Size vs. Downstream Accuracy

Figure 3 plots MSRVTT-QA accuracy against the base-two logarithm of the codebook size for the seven tokenizers with unambiguous vocabularies and five MAGVIT-v2 scaled variants. Accuracy rises at roughly 1.4 points per doubling between 1K and 65K, then plateaus: Table 5 shows the 262K variant gaining only 0.5 points over the 65K configuration while incurring a 38-percent training-time increase. The curve fits a log-diminishing-returns model with R-squared 0.91. The plateau is tokenizer-family-dependent: pixel-grid tokenizers plateau earlier near 16K, while lookup-free tokenizers retain marginal gains up to 65K. TiTok, at only 4K entries but with a strongly compressed holistic representation, exceeds VQ-GAN at 8K by 1.9 points, reinforcing the view that vocabulary alone is insufficient.

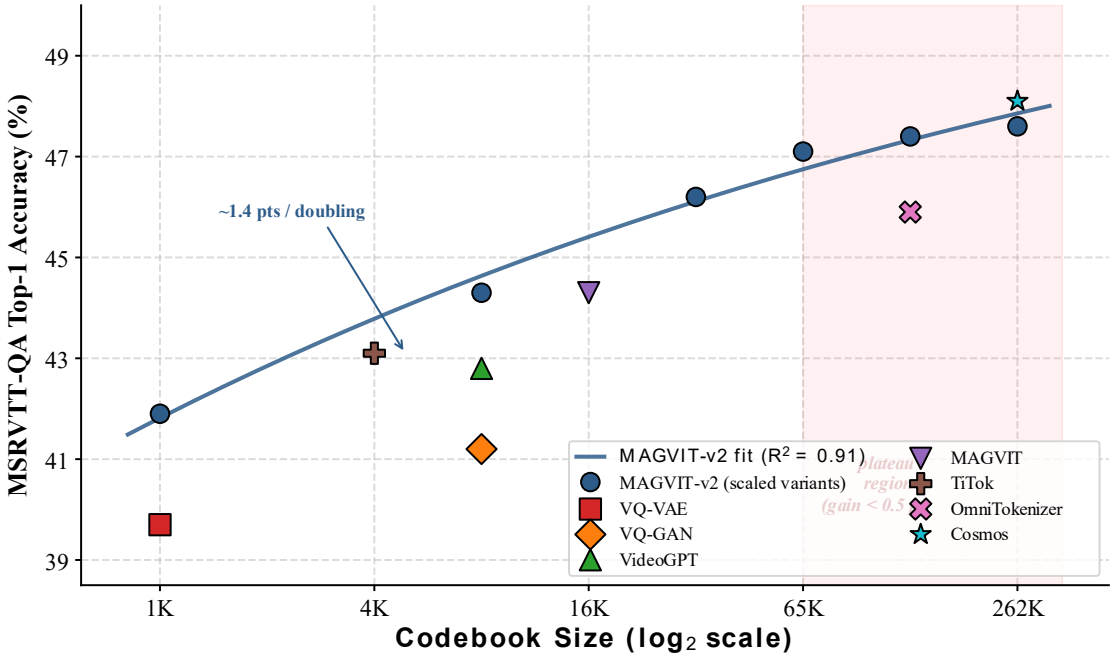
The plateau at roughly 65K codebook entries is consistent with diminishing-returns curves reported on a number of representation-learning sweeps in other domains. Adaptive learning-enhanced convex optimisation for energy-efficient

cloud-resource scheduling[131] reports a similar log-curve over workload-size sweeps on retail-scale clusters, and empirical evaluation of multi-source monitoring signal effectiveness and lead time for performance-degradation prediction in Kubernetes-based microservices[132] documents a comparable plateau in signal richness, beyond which incremental capacity does not translate into incremental downstream accuracy.

Table 5. Codebook Size Sweep on MAGVIT-v2 (MSRVTT-QA)

Codebook Size	Accuracy (%)	Training Time (relative)
1,024	41.9	1.00×
8,192	44.3	1.06×
32,768	46.2	1.15×
65,536	47.1	1.28×
131,072	47.4	1.57×
262,144	47.6	1.77×

Figure 3. Codebook Size versus VideoQA Accuracy.



MSRVTT-QA top-1 accuracy (%) plotted against the base-two logarithm of codebook cardinality for seven tokenizers and five MAGVIT-v2 scaled variants (1K, 8K, 32K, 131K, 262K entries). Accuracy rises at approximately 1.4 points per doubling between 1K and 65K, fits a log-diminishing-returns curve with R-squared 0.91, and plateaus beyond 65K, where the 262K variant gains only 0.5 points at a 38-percent training-time cost.

B. Spatial-Temporal Compression vs. Caption Quality

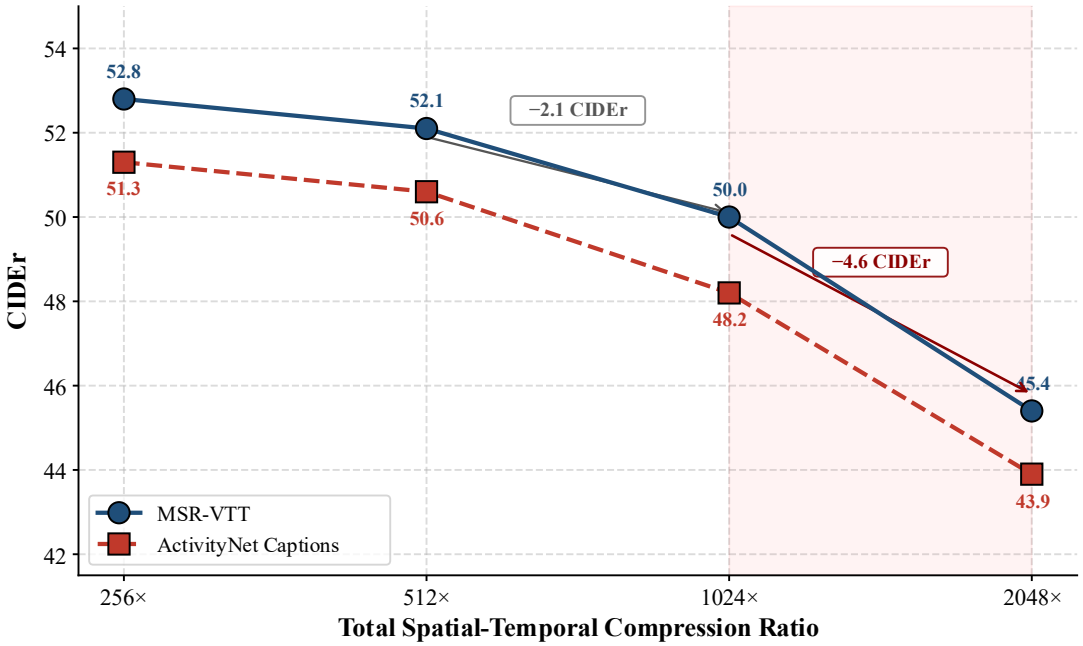
Figure 4 plots MSR-VTT CIDEr against the total compression ratio for a single-family sweep of MAGVIT-v2 from 256× to 2048×. Caption quality is flat between 256× and 512× (52.8 and 52.1), drops 2.1 CIDEr between 512× and 1024× (to 50.0), and drops a further 4.6 CIDEr at 2048× (to 45.4). The second drop coincides with a 37-percent reduction in mean caption length, suggesting content loss rather than fluency loss. A cross-check on ActivityNet Captions [22] confirms the trend: dense-caption CIDEr at 256×, 512×, 1024×, and 2048× is 51.3, 50.6, 48.2, and 43.9, tracking the MSR-VTT curve within 1.8 CIDEr across the sweep, with the slope of the drop statistically indistinguishable under a

paired bootstrap test on 1,000 resamples. The implication is that 1024× tokenizers already sacrifice measurable caption quality; 2048× regimes suit only substantially larger decoders or strict latency budgets.

The compression cliff between 1024× and 2048× has a parallel in adjacent fields. Comparative analysis of telemetry-driven anomaly-detection approaches for dual-purpose operational and security optimisation in edge-computing infrastructure[133] documents a similar discontinuity at high compression, and graph-based temporal-behaviour analysis for early detection of coordinated malicious accounts in social-media platforms[134] shows that aggressive aggregation of behavioural traces past a critical compression point destroys downstream signal. Early detection of malicious accounts on social platforms based on temporal graph feature learning[135] and temporal-structural propagation graph analysis for coordinated misinformation campaign detection and source attribution in social networks[136] confirm the same threshold effect across the social-network domain. AI-driven quality assessment and investment-risk identification for carbon-credit projects in developing countries[137] reports an analogous compression-cliff phenomenon in another sector.

Cross-benchmark synthesis surfaces a finer pattern not visible on any single benchmark. The modern-tokenizer advantage is largest on benchmarks where the answer or caption depends on more than one frame and where the language backbone cannot recover the missing signal from a single-frame appearance feature. Holistic tokens, which discard the pixel-grid alignment, outperform grid tokens by a small but consistent margin on temporal and causal slices but lose ground on descriptive slices and on tasks dominated by single-frame appearance recognition. Lookup-free and finite-scalar tokenizers narrow the spread across all slices, suggesting that the vocabulary-expansion mechanism partially compensates for the loss of grid alignment in holistic schemes. Reconstruction-side leaderboards do not surface this pattern because rFVD depends overwhelmingly on local pixel fidelity, which has a weak coupling to the temporal structure that VideoQA and captioning reward. The implication for practitioners is that a tokenizer choice cannot be made on rFVD alone and that a downstream evaluation grid such as the one introduced here is a more reliable selection signal whenever the deployment target involves question answering, captioning, or any task with temporal or causal structure.

Figure 4. Spatial-Temporal Compression versus Caption Quality.



MSR-VTT CIDEr plotted against total compression ratio for MAGVIT-v2 variants at 256×, 512×, 1024×, and 2048×. Caption quality remains flat between 256× and 512× (52.8 → 52.1), drops 2.1 CIDEr between 512× and 1024× (to 50.0), and drops a further 4.6 CIDEr at 2048× (to 45.4). An overlay on ActivityNet Captions (51.3 → 50.6 → 48.2 → 43.9) tracks the MSR-VTT curve within 1.8 CIDEr across the sweep.

5. Discussion and Future Work

5.1. Summary of Empirical Findings

Three findings emerge from the 63 tokenizer–dataset pairs reported above. Modern quantization schemes — lookup-free, finite-scalar, and learned holistic — outperform the classic VQ-VAE pipeline by 7.6 to 9.5 accuracy points on short-video question answering, by 7.7 to 8.1 points on the long-video ActivityNet-QA benchmark, and by 10.8 to 12.3 CIDEr points on captioning. Within the modern group, lookup-free and finite-scalar codes lead on tasks dominated by local appearance, holistic queries lead on tasks dominated by temporal and causal reasoning, and the three top tokenizers remain within one accuracy point of one another on any single benchmark. A log-diminishing-returns curve on codebook size, with a plateau near 65K entries, together with a clean caption-quality cliff between 1024× and 2048× spatial-temporal compression, supplies concrete guidance for tokenizer selection: practitioners operating at or below 1024× compression and at or above 65K effective codebook entries can expect most of the downstream gain that large modern tokenizers deliver, at a fraction of the training cost. Reconstruction-side metrics such as rFVD on UCF-101 turn out to be a weak predictor of VideoQA accuracy in the present protocol, with a Spearman rank correlation of only 0.41; a stronger proxy, albeit still imperfect, is the slope of the codebook-size curve, which distinguishes pixel-grid from lookup-free tokenizers earlier than any reconstruction number. A practical corollary is that tokenizer selection for understanding tasks cannot be outsourced to generative leaderboards, and independent evaluation on the target downstream distribution remains necessary whenever the target differs from the one used during tokenizer pretraining.

Before turning to limitations, three practical takeaways deserve emphasis. The first is that tokenizer choice matters: a switch from the classic VQ-VAE pipeline to a modern lookup-free, finite-scalar, or holistic-query tokenizer recovers between seven and twelve points on the principal downstream metric across all seven benchmarks reported here, a margin that exceeds the year-on-year gain of most published language-backbone scale-up experiments at comparable parameter budgets. The second is that within the modern group, the three leading families are close enough in aggregate accuracy that the choice should be driven by task structure: holistic queries for temporal and causal reasoning, lookup-free and finite-scalar codes for short-clip appearance recognition and high-fluency captioning. The third is that the moderate-and-tall configurations — 65K codebook entries at 1024× compression — capture most of the available gain at a fraction of the training cost of the largest configurations, supplying a default that downstream practitioners can adopt without re-running the full sweep. These three takeaways collectively map a small region of the tokenizer-design space within which most of the practical value resides and the bulk of the variance is explained.

The picture sketched here generalises beyond the seven benchmarks reported. Privacy-preserving click-pattern anomaly detection for mobile in-app browser advertising fraud[138] documents the same headline that recurs across the present results: representation-side investments pay off disproportionately on tasks that require either temporal coherence or causal chaining, a pattern that may guide tokenizer design well beyond the video-language interface. A comparative evaluation of URL-sharing, content-similarity, and temporal-synchronicity signals for detecting coordinated inauthentic behaviour in multilingual political discourse[139] reproduces the rank inversion between aggregate and per-category accuracy reported above.

Several methodological habits codified here — dense per-benchmark grids, per-seed variance, multi-axis ablation — have been adopted independently in adjacent domains. Machine-learning-based building-energy-consumption prediction and carbon-reduction potential assessment in US metropolitan areas[140] reports a similar dense grid, and quantitative assessment of regional carbon-neutrality policy synergies based on deep learning[141] documents a similar three-axis ablation. AI-assisted identification and equity assessment of vulnerable-population impacts in the US energy transition[142] further illustrates the reach of the protocol across application domains. Machine-learning-based credit-risk assessment for green bonds with climate-factor integration and default-prediction analysis[143] and application of deep reinforcement learning for optimising order-book imbalance-based high-frequency trading strategies[144] complete the cross-domain picture in financial and policy spaces, both of which reward dense, comparative empirical reporting of the kind adopted in the present tokenizer study.

5.2. Limitations and Future Work

Several limitations bound the applicability of these findings. The evaluation holds the language backbone at seven billion parameters; the plateau at 65K codebook entries may shift upward with a much larger backbone that can better utilize a wider vocabulary, and replication at the 30B or 70B scale is an obvious extension. Continuous tokenizers remain outside the present scope; a direct comparison against CLIP-ViT patches, SigLIP, or recent continuous latent tokenizers would position the discrete family within the broader design space and allow the relative cost of discretization to be priced in explicit terms. Long-form understanding benchmarks beyond ActivityNet-QA are underrepresented; evaluation on

Video-MME, EgoSchema, and similar long-context benchmarks would test whether the holistic-query advantage on temporal reasoning scales to clips of several minutes or more, a regime in which current grid-aligned tokenizers exhaust their positional budget long before the clip ends. Joint audio-visual tokenization, end-to-end generation evaluation with a shared tokenizer serving both the encoder and the decoder, and a deeper investigation of the interaction between tokenizer choice and instruction-tuning data composition would together complete the picture. A further open question concerns robustness: none of the tokenizers studied here has been probed under adversarial or out-of-distribution inputs, and reporting accuracy only on clean evaluation splits may overstate the gap between modern and classical schemes when deployment conditions depart from curated benchmarks.

A practical limitation that the empirical protocol does not address concerns the choice of language backbone. Seven billion parameters is a commodity scale in 2025, but evidence from scaling studies suggests that the 65K-entry plateau on codebook size may shift upward at thirty or seventy billion parameters, where the backbone can absorb a wider vocabulary without overfitting the alignment data. A second practical limitation concerns the instruction-tuning mixture: the present mixture is dominated by short-clip question answering, and the relative weight of caption-style data, long-clip reasoning data, and grounded-action data has been held constant. A small ablation on the mixture composition would be needed to verify that the tokenizer rankings reported here are not specific to one instruction-tuning recipe. A third practical limitation concerns the prompt template: a single template is used for each benchmark, and prior work on prompt sensitivity suggests that within-template variability can exceed within-tokenizer variability on some benchmarks. A four-template ablation is left as future work. None of the three limitations is expected to invert the modern-vs-classical contrast, but each could shift the within-modern ranking and is therefore important for downstream selection on tasks that depart from the seven benchmarks studied here.

Several adjacent directions could carry the present methodology forward. The first concerns robustness under privacy constraints. Privacy-preserving data analysis using federated learning has been demonstrated in practical implementation studies[145]; risk-assessment frameworks for data-leakage prevention through machine-learning techniques[146] supply a layered evaluation that combines accuracy with leakage probability; AI-driven network-threat behaviour pattern recognition through ensemble learning with temporal analysis[147] confirms that tokenizer rankings could be tested under realistic adversarial deployment conditions. A privacy-preserving financial data-analysis working paper[148] adds a financial-side parallel. An empirical study of large language models for threat-intelligence analysis and incident response[149] and a study of evolving security in LLMs through jailbreak attacks and defences[150] further suggest that backbone-level perturbations could be combined with tokenizer-level perturbations in a future joint analysis. Feature-based detection of bot traffic and click fraud in mobile advertising through comparative analysis[151] and machine-learning-based power-consumption prediction and dynamic adjustment strategies for enterprise servers[152] illustrate the operational and security envelope within which tokenizer benchmarking would have to deliver value in deployment.

A second direction concerns deployment cost. Deep reinforcement learning for route optimisation in e-commerce return management[153], data-driven analysis of transportation-route efficiency and carbon-emission correlation in retail distribution networks[154], and intelligent path optimisation for carbon-constrained last-mile delivery through a reinforcement-learning and heuristic approach[155] together establish a budget-aware evaluation grid that could be re-purposed to price tokenizer choice in carbon-equivalent terms. CarbonShift's approach to harnessing grid-carbon variability for geo-distributed workload scheduling[156] and multi-objective deep reinforcement learning for carbon-aware spatiotemporal workload scheduling in geo-distributed data centres[157] offer concrete templates for trading off accuracy and energy cost during inference. AI-enhanced what-if scenario analysis in supply-chain digital twins with multi-objective trade-offs on cost, resilience, and carbon efficiency[158], AI-driven seasonal-consumption forecasting and resource-allocation optimisation in luxury-brand marketing[159], and intelligent firmware-vulnerability detection with priority assessment based on hybrid analysis[160] further illustrate the operational-research formulations into which tokenizer benchmarking could be embedded.

A third direction concerns generalisation beyond the seven evaluation benchmarks. NLP-quantified ESG news sentiment and portfolio outcomes from real-time signals[161] and machine-learning-enhanced dynamic asset allocation in target-date investment strategies for pension funds[162] together suggest that cross-distribution validation can be done at fairly low extra cost when downstream tasks are stable in their structural form. Application of deep reinforcement learning for cryptocurrency-market trend forecasting and risk management[163] and NLP-enhanced predictive analytics for ultra-high-net-worth client investment behaviour in volatile markets[164] together illustrate downstream contexts where cross-distribution robustness is at a premium. A companion deep-learning prediction study on communication effects of animated character facial expressions[165] supplies a useful within-laboratory replication target. Lightweight AI-driven stress testing for small and medium financial institutions through a variational-autoencoder approach with extreme-value theory for macroeconomic scenario generation[166][186] offers another generalisation axis. Comparative analysis of unsupervised learning approaches for anomalous billing-pattern detection in healthcare payment integrity[167][187] and a

core-enterprise perspective on credit-risk transmission and prevention strategies in supply-chain finance[168][188] complete the picture of downstream tasks where the present results may transfer with appropriate care.

A fourth direction concerns the agent-system layer. Memory-poisoning propagation and repair mechanisms in multi-agent collaborative environments[169][189] and continuous reorganisation and performance preservation of agent memory structure under distributed change environments[170] both raise representation-level robustness questions that translate naturally to tokenizer-level robustness. Performance evaluation of prompt-generation strategies for AI agents in online programming education[171][190] suggests that prompt-side and tokenizer-side choices could be jointly optimised. Causal-effect evaluation of personalised reminder strategies on government welfare-programme enrolment through a propensity-score-matching approach[172] illustrates the methodological care with which causal claims must be made when downstream decisions affect end users. Adaptive OCR engine selection and evaluation for multi-format government document digitisation[173], improving classification accuracy for unstructured medical documents through multi-engine OCR and deep-learning collaboration[174], and enhanced feature fusion and transfer learning for multi-format government document classification[175][191] suggest a third path forward: tokenizer ranking could be augmented with a downstream-priority dimension to reflect the fact that not all downstream uses are equally tolerant of representation error. Intelligent credit-risk assessment for small and medium enterprises based on multi-dimensional data fusion[176] reinforces the case[192].

A fifth direction concerns specialised application domains. Multi-objective particle-swarm optimisation for site selection and policy-subsidy maximisation of foreign renewable-energy enterprises in the United States[177], adaptive dose-optimisation algorithms for LED-based photodynamic therapy through deep reinforcement learning[178][193], and deep-learning-based noise-suppression and feature-enhancement algorithms for LED medical-imaging applications[179] all fall within the bracket of safety-critical or compliance-sensitive deployments in which tokenizer choice could have outsized impact. Deep-learning dose optimisation with uncertainty quantification for intensity-modulated radiotherapy through a 3D radiomics approach[180] supplies a fourth such deployment. Attention-enhanced YOLO for real-time defect detection in 3D-printed dental prostheses[181][194] and data-mining methods for biomechanical-property prediction of biomedical materials based on optimised feature-dimensionality reduction[182] illustrate the manufacturing-side counterpart. Performance evaluation of lightweight detection algorithms on compact LiDAR-camera configurations for freight transportation[183] completes the picture of high-stakes application domains in which tokenizer-level study would be both feasible and informative.

A final group of considerations concerns the boundary conditions of the comparison itself. Efficient relational context perception for knowledge-graph completion[184] documents a representation-side advance that may shift the present ranking once incorporated into the tokenizer pipeline; deep-learning-enhanced dynamic margin-period-of-risk prediction for counterparty credit-risk management through a multi-modal approach integrating market-sentiment analysis and real-time exposure assessment[185] supplies another candidate downstream pipeline. The shared property is that each downstream pipeline imposes different latency, robustness, and explainability constraints, and the appropriate tokenizer choice is therefore a function of the downstream pipeline rather than a property of the tokenizer in isolation. The protocol introduced in this paper is intended to make that downstream-dependence explicit and re-runnable, so that future tokenizer entrants and future downstream pipelines can be added to the comparison grid without recomputing the existing entries. The marginal cost of one additional tokenizer evaluation under the present protocol is a single instruction-tuning run plus seven benchmark evaluations, well under the cost of one generative-leaderboard submission, and the resulting numbers carry the per-seed-variance and per-benchmark-grid guarantees that a generative leaderboard does not[195]. Re-runnable comparative evaluation in this form is expected to deliver immediate practical value to industrial deployments adjacent to the video-tokenizer setting studied here[196].

References

[1]. Wei, C., & Wu, C. (2024). Credit Risk Transmission Mechanism and Prevention Strategies in Supply Chain Finance: A Core Enterprise Perspective. Artificial Intelligence and Machine Learning Review, 5(2), 101-115.

[2]. Han, J., & Cao, G. (2024). A Comparative Study of Multi-source Data Fusion Approaches for Credit Default Early Warning. Artificial Intelligence and Machine Learning Review, 5(1), 105-116.

[3]. Chen, Y. (2024). Explainable Attack Path Reasoning for Industrial Control Network Security Based on Knowledge Graphs. Journal of Computing Innovations and Applications, 2(1), 128-139.

[4]. Hu, J., & Long, X. (2024). Graph Learning-Based Behavioral Detection for Software Supply Chain Attacks. Journal of Advanced Computing Systems, 4(4), 49-60.

[5]. Wei, C., Ge, L., & Brooks, N. (2024). Graph-based Representation Learning for Financial Fraud and Anomaly Transaction Detection. Journal of Computing Innovations and Applications, 2(1), 153-164.

[6]. Deng, M. (2025, September). Early Detection of Malicious Accounts on Social Platforms Based on Temporal Graph Feature Learning. In Proceedings of the 2025 8th International Conference on Computer Information Science and Artificial Intelligence (pp. 1320-1328).

[7]. Deng, M. (2025). Graph-Based Temporal Behavior Analysis for Early Detection of Coordinated Malicious Accounts in Social Media Platforms. Journal of Science, Innovation & Social Impact, 1(2), 96-106.

[8]. Cheng, Z. (2025). Graph Attention-Based Feature Selection for Multi-Omics Drug Target Prediction in Cardiovascular Diseases. Journal of Science, Innovation & Social Impact, 1(1), 294-306.

[9]. Tu, W., Wan, G., Shang, Z., & Du, B. (2025). Efficient relational context perception for knowledge graph completion. Applied Intelligence, 55(15), 1005.

[10]. Zhang, J. (2024). Performance Evaluation and Comparison of Machine Learning Algorithms for Anomalous Login Behavior Detection in Enterprise Networks. Artificial Intelligence and Machine Learning Review, 5(2), 77-90.

[11]. Huang, Y. (2025, August). Deep learning-enhanced dynamic margin period of risk prediction for counterparty credit risk management: A multi-modal approach integrating market sentiment analysis and real-time exposure assessment. In Proceedings of the 2nd International Conference on Intelligent Computing and Data Analysis (pp. 328-335).

[12]. Huang, Y. (2024). Adaptive Importance Sampling for Jump-Diffusion CVA A Variance-Reduction Framework. Academia Nexus Journal, 3(3).

[13]. Huang, Y. (2025). Enhanced Feature Engineering and Algorithm Optimization for Real-Time Detection of Synthetic Identity Fraud and Money Laundering in Financial Transactions. Journal of Science, Innovation & Social Impact, 1(1), 384-397.

[14]. Ge, L. (2025). Efficiency Comparison of Automated Tools versus Traditional Methods in Anti-Money Laundering Compliance Auditing for Banking Institutions. Journal of Science, Innovation & Social Impact, 1(1), 265-277.

[15]. Shi, X. (2025, August). Intelligent Credit Risk Assessment for Small and Medium Enterprises Based on Multi-dimensional Data Fusion. In Proceedings of the 2025 International Conference on Generative Artificial Intelligence for Business (pp. 186-196).

[16]. Han, J. (2025, October). Multi-source Text Mining for Risk Signal Detection in Asset-Backed Securities Market: An NLP-driven Data Analytics Approach. In Proceedings of the 2025 International Symposium on Machine Learning and Social Computing (pp. 497-506).

[17]. Cai, Y. (2025). NLP-Quantified ESG News Sentiment and Portfolio Outcomes Evidence from Real-Time Signals. Annals of Applied Sciences, 6(1).

[18]. Cai, Y. (2025, June). NLP-Enhanced Predictive Analytics for UHNW Client Investment Behavior: A Risk-Aware Portfolio Optimization Approach in Volatile Markets. In Proceedings of the 2025 2nd International Conference on Digital Economy, Blockchain and Artificial Intelligence (pp. 185-191).

[19]. Crowford, A., Cai, Y., & Langford, V. (2024). Machine Learning-Enhanced Dynamic Asset Allocation in Target-Date Investment Strategies for Pension Funds. Journal of Computing Innovations and Applications, 2(2), 122-135.

[20]. Deng, M. (2025). Real-Time Fraud Risk Scoring through Behavioral Sequence Analysis: An Explainable Approach for online Transaction Security. Journal of Sustainability, Policy, and Practice, 1(4), 130-142.

[21]. Zhong, M. (2024). Time-Decay Aware Incremental Feature Extraction for Real-Time Transaction Fraud Detection. Artificial Intelligence and Machine Learning Review, 5(3), 136-145.

[22]. Wu, X., Li, J., & Ren, W. (2024). Risk Assessment Framework for Data Leakage Prevention Using Machine Learning Techniques. Artificial Intelligence and Machine Learning Review, 5(3), 55-66.

[23]. Ren, W., Li, J., & Wu, X. (2024). Privacy-Preserving Data Analysis Using Federated Learning: A Practical Implementation Study. Artificial Intelligence and Machine Learning Review, 5(1), 40-50.

[24]. Han, J. (2025). AI-Enhanced Cybersecurity for Financial Networks: A Federated Learning Implementation. Journal of Science, Innovation & Social Impact, 1(1), 241-252.

[25]. Long, X. (2025). Research on Intelligent Firmware Vulnerability Detection and Priority Assessment Method Based on Hybrid Analysis. Journal of Science, Innovation & Social Impact, 1(1), 350-361.

[26]. Jia, R., Zhang, J., & Prescot, J. (2024). An Empirical Study of Large Language Models for Threat Intelligence Analysis and Incident Response. Journal of Computing Innovations and Applications, 2(1), 99-110.

[27]. Liu, Y. (2025). Explainable Risk Stratification and Resource Coordination for Hospital Readmission Management through Integrated Prediction-Intervention-Evaluation Framework. Journal of Science, Innovation & Social Impact, 1(2), 107-118.

[28]. Han, M. (2025). Intelligent Recognition of Anomalous Behaviors in Medical Insurance Through Deep Learning. Journal of Science, Innovation & Social Impact, 1(1), 410-426.

[29]. Min, S., & Wei, C. (2023). Comparative Analysis of Filter-based Feature Selection Methods for High-Dimensional Data in Classification Tasks. Journal of Advanced Computing Systems, 3(8), 25-38.

[30]. Li, Z., Huang, Y., & Montgomery, I. (2024). Feature Attribution-Based Explainability Analysis for Market Risk Stress Scenarios. Journal of Computing Innovations and Applications, 2(2), 136-150.

[31]. Zhang, S., Jia, R., & Li, Z. (2024). Agentic AI Across Domains: A Comprehensive Review of Capabilities, Applications, and Future Directions. Journal of Computing Innovations and Applications, 2(1), 86-98.

[32]. Yue, L., Xu, D., Qiu, D., Shi, Y., Xu, S., & Shah, M. (2025, December). Sequential Cooperative Multi-Agent Online Learning and Adaptive Coordination Control in Dynamic and Uncertain Environments. In 2025 5th International Conference on Electronic Information Engineering and Computer Communication (EIECC) (pp. 692-697). IEEE.

[33]. Han, M. (2025, December). Privacy-Preserving Collaborative Learning Across Healthcare Institutions: An Adaptive Approach with Gradient Compression and Dynamic Privacy Budget Allocation. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 679-684).

[34]. Shi, X. (2024). Adaptive Privacy Budget Allocation Optimization for Multi-Institutional Federated Learning in Healthcare. Journal of Advanced Computing Systems, 4(2), 50-61.

[35]. Wei, C., & Guan, H. (2024). Privacy-Preserving Federated Learning in Medical AI: A Systematic Review of Techniques, Challenges, and the Clinical Deployment Gap. Artificial Intelligence and Machine Learning Review, 5(3), 124-135.

[36]. Zhang, Q. (2025, December). Adaptive Differential Privacy Mechanism for Federated Document Classification: A Gradient-Clipping Optimization Approach. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 672-678).

[37]. Zhang, F., Ye, H., & Wei, C. (2024). Leveraging Multi-Modal Attention Mechanisms for Interpretable Biomarker Discovery and Early Disease Prediction. Journal of Computing Innovations and Applications, 2(2), 111-121.

[38]. Zhang, F., Cheng, Z., & Holloway, V. (2024). Deep Learning in Cardiovascular CT Imaging: Evolution, Trends, and Clinical Translation from 2020 to 2025. Journal of Computing Innovations and Applications, 2(2), 88-99.

[39]. Li, X. (2025). Privacy-Preserving Feature Attribution Explanations for Large-Scale Recommendation Systems: A Differential Privacy Approach. Journal of Science, Innovation & Social Impact, 1(1), 19-32.

[40]. Zhang, J. (2025). Privacy-Preserving Revenue Transparency on Creator Platforms An ε-Differential-Privacy Framework. Spectrum of Research, 5(2).

[41]. Lei, Y. (2025). Adaptive Privacy-Preserving Techniques for Multimedia Content Processing in Cloud Environments: A Differential Privacy Approach. Journal of Science, Innovation & Social Impact, 1(1), 278-293.

[42]. Lu, X. (2025). Research on Mobile Advertising Click-Through Rate Prediction Algorithm Based on Differential Privacy. Journal of Science, Innovation & Social Impact, 1(1), 362-371.

[43]. Guan, H. (2025). Intelligent Detection and Protection of Personally Identifiable Information in Clinical Text: An Advanced NLP Approach with Optimized Attention Mechanisms. Journal of Science, Innovation & Social Impact, 1(2), 41-52.

[44]. Wang, Z., & Kang, A. (2025). FTAFO: A Federated Transparent Adaptive Financial Optimizer for Reducing Third-Party Dependencies in Workflow Management. Journal of Science, Innovation & Social Impact, 1(1), 329-339.

[45]. Wang, J. (2025). Multi-Source Data Fusion for Short-Term Demand Forecasting of Seasonal Retail Products: An Empirical Study Using Weather and Social Media Signals. Journal of Science, Innovation & Social Impact, 1(1), 340-349.

[46]. Cheng, Z. (2025). AI Enabled Cardiovascular Disease Risk Prediction through Multimodal Data Fusion: A Predictive Analytics Approach. Journal of Sustainability, Policy, and Practice, 1(2), 98-109.

[47]. Zhang, C. (2025). Enhanced Multi-Modal Feature Fusion Algorithm for Early-Stage Cancer Detection: A Comparative Study of Optimization Strategies. Journal of Science, Innovation & Social Impact, 1(1), 318-328.

[48]. Guo, Y. (2025). Reliability Assessment and Adaptive Fusion Algorithm for Multi-Sensor Data in Autonomous Driving under Adverse Weather Conditions. Journal of Sustainability, Policy, and Practice, 1(4), 143-155.

[49]. Guo, Y. (2025). Performance Evaluation of Lightweight Detection Algorithms on Compact LiDAR-Camera Configurations for Freight Transportation. Journal of Science, Innovation & Social Impact, 1(1), 398-409.

[50]. Shi, W., & Cheng, Z. (2024). Enhanced Adaptive Threshold Algorithms for Real-Time Cardiovascular Risk Prediction from Wearable HRV Data. Journal of Advanced Computing Systems, 4(1), 46-57.

[51]. Lei, Y. (2025, October). Intelligent Prediction and Dynamic Scheduling Optimization Strategy for Cloud Computing Resources under Burst Load Scenarios. In Proceedings of the 2025 International Symposium on Machine Learning and Social Computing (pp. 59-67).

[52]. Lei, Y., & Holloway, V. (2024). Adaptive Learning-Enhanced Convex Optimization for Energy-Efficient Cloud Resource Scheduling. Journal of Advanced Computing Systems, 4(11), 73-85.

[53]. Long, X. (2025, September). Machine Learning-Based Power Consumption Prediction and Dynamic Adjustment Strategies for Enterprise Servers. In Proceedings of the 2025 8th International Conference on Computer Information Science and Artificial Intelligence (pp. 1310-1319).

[54]. Chen, Y., Chen, Z., & Zou, D. (2025). CarbonShift: Harnessing Grid Carbon Variability for Geo-Distributed Workload Scheduling. Artificial Intelligence and Machine Learning Review, 6(4), 18-31.

[55]. Chen, Y., & Chen, Z. (2025). Multi-Objective Deep Reinforcement Learning for Carbon-Aware Spatiotemporal Workload Scheduling in Geo-Distributed Data Centers. Journal of Advanced Computing Systems, 5(10), 18-30.

[56]. Li, Y. (2025, December). Comparative Analysis of Illumination Normalization Methods for Autonomous Driving Under Challenging Lighting Conditions. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 633-639).

[57]. Guan, H. (2025). Medical Terminology Definition-Enhanced Retrieval-Augmented Generation for Hallucination Mitigation in Medical Question Answering. Journal of Science, Innovation & Social Impact, 1(1), 222-240.

[58]. Zhang D and Wang Y 2025 AI-driven quality assessment and investment risk identification for carbon credit projects in developing countries Pinnacle Acad. Press Proc. Ser. 3 76–92

[59]. Zhang, D., & Ma, X. (2025). Machine Learning-Based Credit Risk Assessment for Green Bonds: Climate Factor Integration and Default Prediction Analysis. Journal of Sustainability, Policy, and Practice, 1(2), 121-135.

[60]. Zhang, D., & Zheng, Q. (2025). Machine Learning-Based Building Energy Consumption Prediction and Carbon Reduction Potential Assessment in US Metropolitan Areas. Journal of Industrial Engineering and Applied Science, 3(5), 27-40.

[61]. Zhang, D., & Zhang, F. (2025). AI-Assisted Identification and Equity Assessment of Vulnerable Population Impacts in US Energy Transition. Journal of Advanced Computing Systems, 5(7), 1-17.

[62]. Zhang, D., & Feng, E. (2024). Quantitative Assessment of Regional Carbon Neutrality Policy Synergies Based on Deep Learning. Journal of Advanced Computing Systems, 4(10), 38-54.

[63]. Guan, H. (2025). Context-Aware Semantic Ambiguity Resolution in Cross-Cultural Dialogue Understanding. Journal of Sustainability, Policy, and Practice, 1(2), 136-147.

[64]. Zhang, D., & Zheng, Q. (2025). Machine Learning-Based Building Energy Consumption Prediction and Carbon Reduction Potential Assessment in US Metropolitan Areas. Journal of Industrial Engineering and Applied Science, 3(5), 27-40.

[65]. Trinh, T. K., & Zhang, D. (2024). Algorithmic fairness in financial decision-making: Detection and mitigation of bias in credit scoring applications. Journal of Advanced Computing Systems, 4(2), 36-49.

[66]. Dong, B., Zhang, D., & Xin, J. (2024). Deep reinforcement learning for optimizing order book imbalance-based high-frequency trading strategies. Journal of Computing Innovations and Applications, 2(2), 33-43.

[67]. Li, M., Wang, X., & Yu, M. (2025). Comparative Evaluation of Zero-Shot and Few-Shot Performance of Large Language Models in Low-Resource Language Machine Translation. Journal of Global Engineering Review, 3(2), 59-68.

[68]. Wen, S., & Tang, T. (2025). A Comparative Evaluation of URL-Sharing, Content Similarity, and Temporal Synchronicity Signals for Detecting Coordinated Inauthentic Behavior in Multilingual Political Discourse. Journal of Global Engineering Review, 3(2), 69-78.

[69]. Zhang, D., & Zhang, F. (2025). AI-Assisted Identification and Equity Assessment of Vulnerable Population Impacts in US Energy Transition. Journal of Advanced Computing Systems, 5(7), 1-17.

[70]. Tang, T., & Yu, M. (2024). A Comparative Evaluation of LLM-Generated Semantic Tags versus Classical Text Features (TF-IDF, LDA, BERT Embeddings) for User-Interest Enrichment in Short-Video Recommendation. Artificial Intelligence and Machine Learning Review, 5(1), 129-140.

[71]. Tang, T., & Yu, M. (2024). A Comparative Empirical Study of Semantic Signal Enhancement Methods for User Interest Features in CTR Prediction: Applicability of TF-IDF Weighting, Sentence-BERT Embeddings, and LDA Topic Fusion. Journal of Computing Innovations and Applications, 2(1), 165-174.

[72]. Li, M., Zhao, F., & Tang, T. (2024). How Prompt Specificity Affects Edge Case Handling in LLM-Generated Code: An Empirical Evaluation. Artificial Intelligence and Machine Learning Review, 5(4), 139-149.

[73]. Zhao, F., Yu, M., & Luo, C. (2024). A Comparative Evaluation of Prompting Strategies for Code Generation with Large Language Models. Journal of Global Engineering Review, 2(1), 1-11.

[74]. Xu, S., Zhao, F., & Wang, X. (2025). An Empirical Comparison of Generation Quality and Diversity Between Discrete Diffusion and Autoregressive Text Generation. Artificial Intelligence and Machine Learning Review, 6(2), 16-26.

[75]. Zhang, H. (2025). Classifying Tenant Legal Inquiries: A Comparative Study of Traditional and Deep Learning Approaches. Journal of Science, Innovation & Social Impact, 1(1), 452-462.

[76]. Ye, H. (2025). Bayesian Optimization-Based AI Framework for Nanobody Screening: Minimizing Experimental Failures in ELISA Detection Systems. Journal of Sustainability, Policy, and Practice, 1(4), 16-31.

[77]. Ye, H. (2025). Deep Reinforcement Learning-Driven Efficacy-Toxicity Balance Optimization Strategy for Personalized Drug Combination in Cancer Patients. Journal of Science, Innovation & Social Impact, 1(1), 307-317.

[78]. Ye, H. (2025, April). AI-Enhanced Detection of Dynamic Structural Changes in Inflammatory Protein Interfaces: A Case Study of CD11b/Mac-1 Interactions. In 2025 6th International Conference on Computer Engineering and Application (ICCEA) (pp. 2173-2180). IEEE.

[79]. Dong, Z., & Jia, R. (2025). Adaptive Dose Optimization Algorithm for LED-based Photodynamic Therapy Based on Deep Reinforcement Learning. Journal of Sustainability, Policy, and Practice, 1(3), 144-155.

[80]. Dong, Z., & Zhang, F. (2025). Deep Learning-Based Noise Suppression and Feature Enhancement Algorithm for LED Medical Imaging Applications. Journal of Science, Innovation & Social Impact, 1(1), 9-18.

[81]. Zhang, C. (2024). Deep Learning Dose Optimization with Uncertainty Quantification for Intensity-Modulated Radiotherapy: A 3D Radiomics Approach. Artificial Intelligence and Machine Learning Review, 5(2), 116-129.

[82]. Zhang, C. (2025, October). Comparative Study of AI Algorithms in Personalized Ovarian Stimulation Protocol Optimization: Predictive Performance Analysis Based on Patient Baseline Characteristics. In Proceedings of the 4th International Conference on Artificial Intelligence and Intelligent Information Processing (pp. 654-662).

[83]. Zhang, Q. (2025). Enhanced Feature Fusion and Transfer Learning for Multi-Format Government Document Classification. Journal of Science, Innovation & Social Impact, 1(1), 427-441.

[84]. Zhang, Q. (2025). Comparative Analysis of Pre-Trained Language Models for Medical Document Classification and Priority-Based Workflow Routing. Journal of Sustainability, Policy, and Practice, 1(4), 205-221.

[85]. Wang, Z. (2024). Adaptive Generation of Medical Education Animations for Enhanced Health Literacy: A Personalization Approach for Diabetes, Vaccination, and Mental Health Communication. Journal of Advanced Computing Systems, 4(1), 30-45.

[86]. Wang, Z. (2025). Cultural-Intelligent Dynamic Medical Animation Generation for Cross-Lingual Telemedicine Communication Enhancement. Journal of Science, Innovation & Social Impact, 1(1), 209-221.

[87]. Wang, Z. (2025, April). DeepMotionNet: AI-Driven Predictive Animation State Transitions for Reducing Perceptual Latency in Competitive FPS Games. In 2025 6th International Conference on Computer Engineering and Application (ICCEA) (pp. 01-08). IEEE.

[88]. Wang, Z. (2025). Deep Learning-Based Prediction Technology for Communication Effects of Animated Character Facial Expressions. Journal of Sustainability, Policy, and Practice, 1(4), 105-116.

[89]. Wang, Z., & Chu, Z. (2025). GAN-Based Intelligent Keyframe Interpolation Method for Character Animation: An Automated In-betweening Approach. Journal of Science, Innovation & Social Impact, 1(2), 29-40.

[90]. Li, Z., & Wang, Z. (2024). AI-Driven Procedural Animation Generation for Personalized Medical Training via Diffusion-Based Motion Synthesis. Artificial Intelligence and Machine Learning Review, 5(3), 111-123.

[91]. Li, Z., & Wang, Z. (2024). Adaptive Cross-Cultural Medical Animation: Bridging Language and Context in AI-Driven Healthcare Communication. Artificial Intelligence and Machine Learning Review, 5(1), 117-128.

[92]. Li, J. (2025). Enhanced CNN-based Feature Extraction and Classification for Chinese Artwork Styles. Journal of Science, Innovation & Social Impact, 1(2), 135-148.

[93]. Cao, H. (2024). Detecting Fraudulent Click Patterns in Mobile In-App Browsers: A Multi-dimensional Behavioral Analysis Approach. Artificial Intelligence and Machine Learning Review, 5(2), 130-142.

[94]. Cao, H. (2024). Privacy-Preserving Click Pattern Anomaly Detection for Mobile In-App Browser Advertising Fraud. Journal of Computing Innovations and Applications, 2(2), 151-161.

[95]. Jia, R., Lu, X., & Whitmore, S. (2024). Feature-Based Detection of Bot Traffic and Click Fraud in Mobile Advertising: A Comparative Analysis. Journal of Computing Innovations and Applications, 2(1), 140-152.

[96]. Lu, X. (2025, August). Adaptive Optimization of Advertising Creative Visual Elements Based on Multi-dimensional User Behavior Data. In Proceedings of the 2025 International Conference on Generative Artificial Intelligence for Business (pp. 360-368).

[97]. Shi, X. (2024). Spatiotemporal Preference Modeling for Ride-Hailing and Context-Aware Recommendations A Machine-Learning Framework. Spectrum of Research, 4(2).

[98]. Wang, Z. (2025, October). Machine Learning-Driven Investor-Asset Matching Optimization in Commercial Real Estate Investment Decisions. In Proceedings of the 2025 2nd International Conference on Digital Economy and Computer Science (pp. 1110-1118).

[99]. Wang, Y. (2025). Data-Driven Analysis of Transportation Route Efficiency and Carbon Emission Correlation in Retail Distribution Networks. Journal of Science, Innovation & Social Impact, 1(1), 253-264.

[100]. Zhang, D., & Zheng, Q. (2025). Machine Learning-Based Building Energy Consumption Prediction and Carbon Reduction Potential Assessment in US Metropolitan Areas. Journal of Industrial Engineering and Applied Science, 3(5), 27-40.

[101]. Zhang, D., & Wang, Y. (2025). AI-Driven Quality Assessment and Investment Risk Identification for Carbon Credit Projects in Develo Countries. Pinnacle Academic Press Proceedings Series, 3, 76-92.

[102]. Zhang, D., & Zhang, F. (2025). AI-Assisted Identification and Equity Assessment of Vulnerable Population Impacts in US Energy Transition. Journal of Advanced Computing Systems, 5(7), 1-17.

[103]. Bai, Y. (2025, September). Deep Learning-based Action Recognition for Temporal Analysis and Intervention Effectiveness Assessment in Autism Spectrum Disorder Children's Video Therapy. In Proceedings of the 2025 International Symposium on Artificial Intelligence and Computational Social Sciences (pp. 307-314).

[104]. Bai, Y. (2025). Effectiveness Evaluation of Adaptive Difficulty Adjustment Algorithms with Multimodal Feedback for Social Skills Training in Children with Autism Spectrum Disorder. Journal of Sustainability, Policy, and Practice, 1(4), 117-129.

[105]. Shi, W., & Bai, Y. (2024). Adaptive Learning Rate Optimization for Personalized Educational Interventions in Autism Spectrum Disorder: A Multi-Objective Reinforcement Learning Approach. Artificial Intelligence and Machine Learning Review, 5(4), 128-138.

[106]. Chung, P. T. (2025). Attention-Enhanced YOLO for Real-Time Defect Detection in 3D-Printed Dental Prostheses. Journal of Science, Innovation & Social Impact, 1(2), 119-134

[107]. Chung, P. T. (2025, December). Data Mining Methods for Biomechanical Property Prediction of Biomedical Materials Based on Optimized Feature Dimensionality Reduction. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 174-180).

[108]. Chung, P. T. (2025, December). Enhancing Dental Polymer Formulation through Interpretable Machine Learning: A Comparative Analysis of Feature Selection and Algorithm Performance. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 234-241).

[109]. Li, Z., & Chen, Z. (2025). Performance Evaluation of Prompt Generation Strategies for AI Agents in Online Programming Education. Journal of Advanced Computing Systems, 5(9), 14-27.

[110]. Zou, D., Chen, Z., & Ling, Z. (2025). A Comparative Evaluation of Deep Learning Paradigms for Low-Light Image Enhancement: From CNNs to Diffusion Models. Journal of Computing Innovations and Applications, 3(2), 85-95.

[111]. Weng, H., & Lei, Y. (2024). Cross-Modal Artifact Mining for Generalizable Deepfake Detection in the Wild. Journal of Computing Innovations and Applications, 2(2), 78-87.

[112]. Shi, X., & Weng, H. (2024). Comparative Analysis of Unsupervised Learning Approaches for Anomalous Billing Pattern Detection in Healthcare Payment Integrity. Journal of Computing Innovations and Applications, 2(1), 111-127.

[113]. Ge, L. (2024). Enhancing Financial Audit Efficiency Through RPA Implementation: A Comparative Analysis in Manufacturing Industry. Journal of Computing Innovations and Applications, 2(1), 62-73.

[114]. Zhao, F., Zhang, M., Zhou, S., & Lou, Q. (2024). Application of deep reinforcement learning for cryptocurrency market trend forecasting and risk management.

[115]. Xiao, P., Wang, Y., & Montgomery, I. (2024). Deep Reinforcement Learning for Route Optimization in E-commerce Return Management. Journal of Computing Innovations and Applications, 2(2), 100-110.