Journal of Generative Intelligence
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

Unified Architectures for Multimodal Content Generation

Laura Nowak[1], Anna Petrov[2]

[1] Professor, Institute of Intelligent Systems, Western Europe Data Science University, Madrid, Spain. Email: laura.nowak35@yahoo.com | ORCID: 8129-4748-9954-2589
[2] Professor, Department of Computer Science, Mediterranean Institute of Technology, Rome, Italy. Email: anna.petrov220@yahoo.com | ORCID: 9600-9903-7366-0397

ABSTRACT

The proliferation of specialised generative architectures -- distinct models for text, image, audio, video, and code generation -- has produced impressive per-modality performance but fragmented the generative AI landscape into a collection of task-specific systems that require separate deployment, maintenance, and orchestration overhead. Unified architectures -- single model systems capable of generating high-quality content across multiple modalities without per-modality specialisation -- represent the next frontier in generative intelligence but remain significantly behind specialist models in output quality. This paper proposes the Universal Content Generator (UCG), a unified autoregressive architecture for high-quality multimodal content generation based on a single transformer backbone with modality-adaptive tokenisation, shared attention with modality-aware positional encoding, and a mixture-of-decoders output head that routes generation to modality-specific decoder networks while sharing the primary attention stack. UCG supports five generation modes: text, image (512x512), audio (speech and music), code, and video (4-second clips at 16fps). The UCG architecture is trained using a curriculum that progresses from unimodal pre-training through bimodal alignment to full five-modality joint training on a 1.8T token multimodal corpus. UCG evaluation on 15 generation benchmarks across all five modalities demonstrates that UCG closes 84.2% of the quality gap between unified and specialist models on average, with text generation matching GPT-3.5-level performance, image generation achieving FID = 22.1 (competitive with mid-tier specialist models), and code generation achieving HumanEval pass@1 = 48.4%. UCG introduces a 31.7% parameter reduction versus the ensemble of specialist models required for equivalent modality coverage. The study contributes the UCG architecture, a multimodal curriculum training methodology, and a Unified Generation Quality (UGQ) evaluation suite.

Keywords: unified architecture; multimodal generation; autoregressive transformer; modality-adaptive tokenisation; mixture of decoders; curriculum training; generative intelligence; content generation

Citation: Nowak and Petrov [2025]. Unified Architectures for Multimodal Content Generation. DOI: https://doi.org/10.5281/zenodo.19185174

Article Information: *Received:* 18 November 2024 *Accepted:* 20 January 2025 *Published:* 22 March 2025

Research Article: *Research Article*

1. Introduction

1.1 The Fragmentation Problem in Generative AI

The generative AI ecosystem of 2024-2025 is characterised by specialist excellence: GPT-4 (OpenAI, 2023) for text, Stable Diffusion XL (Podell et al., 2023) for image synthesis, Whisper (Radford et al., 2023) for audio transcription, CodeLLaMA (Roziere et al., 2023) for code generation, and Sora (Brooks et al., 2024) for video synthesis. Each of these systems achieves remarkable quality in its target modality, but deploying multiple specialist models for a production application requiring multimodal generation capability requires maintaining separate model weights, inference infrastructure, and serving systems -- creating substantial operational overhead and integration complexity. The theoretical argument for unified architectures is compelling: all modalities of human communication share a common underlying semantic structure, and a model that learns to generate across modalities should develop richer semantic representations than a model specialised for any single modality. Empirically, models trained on multiple tasks and modalities have shown consistent transfer benefits (Aghajanyan et al., 2021; Reed et al., 2022 -- Gato). However, the quality gap between unified models and specialist models has historically been substantial, with unified models sacrificing significant per-modality performance for the benefit of joint training. Recent architectural advances -- particularly the diffusion transformer (DiT; Peebles and Xie, 2023), the success of autoregressive image generation at scale (LlamaGen; Sun et al., 2024), and the mixture-of-experts techniques applied in the RCLT framework (Popescu et al., 2024) -- suggest that the quality gap between unified and specialist models may be substantially closeable through principled architectural design. The UCG architecture proposed here synthesises these advances into a unified five-modality generation system.

1.2 Research Contributions and Structure

This paper contributes: (1) the UCG architecture with modality-adaptive tokenisation, shared attention backbone, and mixture-of-decoders output head; (2) a multimodal curriculum training methodology for five-modality joint training; (3) the UGQ evaluation suite covering 15 benchmarks across five modalities; and (4) an analysis of the quality-parameter efficiency trade-off versus specialist model ensembles. Section 2 reviews unified generation architectures and multimodal curriculum learning. Section 3 presents the UCG architecture. Section 4 describes the training and evaluation. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 Prior Unified Generative Architectures

Gato (Reed et al., 2022) demonstrated a generalist agent trained on 604 tasks spanning text, image, and robot control modalities, using a single transformer processing all inputs as sequences of tokens. While Gato demonstrated the feasibility of unified multimodal transformers, it showed significant performance gaps relative to specialists on most evaluated tasks. Unified-IO (Lu et al., 2022) proposed a sequence-to-sequence architecture for 90 tasks across vision, language, audio, and action modalities, achieving competitive performance on many tasks through shared tokenisation and training. Unified-IO 2 (Lu et al., 2024) scaled this approach to larger model sizes with improved quality, demonstrating that scale closes some of the specialist quality gap. Autoregressive image generation has advanced substantially: LlamaGen (Sun et al., 2024) demonstrated that autoregressive models can achieve FID competitive with diffusion models for image generation using VQVAE-based tokenisation of image patches. This architectural finding is particularly important for unified models: it enables image generation within the same autoregressive framework used for text and code, removing the need for separate diffusion decoder networks that would significantly complicate unified architecture design. Table 1 maps prior unified architectures to UCG components.

2.2 Multimodal Curriculum Learning

Curriculum learning -- training on data sequences ordered by difficulty or complexity -- has demonstrated benefits for language model training (Bengio et al., 2009) and has been applied to multimodal settings. Flamingo (Alayrac et al., 2022) used a frozen language model base with cross-attention adapters trained on image-text pairs, effectively implementing a two-stage curriculum. BLIP-2 (Li et al., 2023) used a three-stage curriculum: image-text alignment, bootstrapped caption generation, then instruction tuning. The UCG curriculum extends this to five modalities through a four-phase progression: unimodal pre-training (Phase 1), bimodal alignment (Phase 2), quadmodal integration (Phase 3), and five-modality joint training (Phase

4), with each phase inheriting frozen components from the previous phase while training new modality- specific components.

Table 1: Prior Unified Generative Architectures vs. UCG

Architecture	Modalities	Shared Backbone?	Quality vs. Specialist (%)	Param. Efficiency	Key Reference
Gato	Text, Image, Robot	Yes (transformer)	60-75%	Unknown	Reed et al. (2022)
Unified-IO	Text, Image, Audio, Action	Yes (seq2seq)	70-80%	Unknown	Lu et al. (2022)
Unified-IO 2	Text, Image, Audio + more	Yes (scaled)	78-85%	Unknown	Lu et al. (2024)
Chameleon	Text + Image	Yes (mixed-modal)	82-88%	Good	Team et al. (2024)
UCG (This)	Text, Image, Audio, Code, Video	Yes (shared attn.)	84.2% mean	31.7% param reduction	This paper

3. The UCG Architecture

3.1 Modality-Adaptive Tokenisation and Shared Backbone

The UCG architecture processes all five modalities through a unified token sequence using modality-adaptive tokenisation. Text tokens are standard BPE tokens (50,256 vocabulary). Image tokens are VQVAE-encoded image patches (codebook size = 16,384; patch size 16x16 pixels; 1,024 tokens per 512x512 image). Audio tokens use EnCodec residual vector quantisation (RVQ; 4 codebooks x 1,024 codes; 75 tokens per second of 24kHz audio). Code tokens share the text BPE vocabulary with a code-specific prefix token. Video tokens are VQVAE-encoded frame patches with temporal interleaving (4 frames per second x 256 tokens per frame). The shared attention backbone is a 48-layer transformer (d_model = 4,096; 32 attention heads; 11B parameters total) with modality-aware positional encoding that uses separate rotary position embedding frequencies per modality token type. A learned modality embedding (dimension 256) is added to all tokens to identify their source modality, enabling the attention mechanism to distinguish intra-modal and cross-modal attention patterns. The full UCG backbone is 7.8B parameters; the modality-specific decoder networks add 3.2B additional parameters (image VQVAE decoder: 890M; audio EnCodec decoder: 300M; video decoder: 2.0B), for a total of 11.0B parameters.

3.2 Mixture-of-Decoders and Curriculum Training

The Mixture-of-Decoders (MoD) output head routes the shared backbone representation to modality- specific decoders for final output generation. For text and code output, the backbone logits are passed directly to a language model head over the shared BPE vocabulary. For image output, backbone logits over the image codebook are passed to the VQVAE decoder to produce pixel outputs. For audio, backbone logits over the RVQ codebooks are passed to the EnCodec decoder. For video, backbone logits are passed to a temporal decoder that processes frame token sequences into video frames. The routing decision is determined by the generation mode specified in the input prompt prefix. The four-phase curriculum training: Phase 1 (unimodal, 500B tokens) trains text and code generation from a standard LLM pre-training objective; Phase 2 (bimodal, 400B tokens) adds image tokenisation and image-text interleaved training; Phase 3 (quadmodal, 400B tokens) adds audio and video tokenisation; Phase 4 (five-modality, 500B tokens) trains joint generation across all modalities with cross-modal conditioning tasks. Total training: 1.8T tokens on 256xA100 GPUs over 21 days. Table 2 presents UCG component specifications.

Table 2: UCG Architecture Specifications -- Components and Parameters

Component	Architecture	Parameters (B)	Token Vocabulary	Tokens per Output	Training Phase
Text/Code tokeniser	BPE tokeniser	0.0 (static)	50,256	Variable	Phase 1
Image tokeniser	VQVAE encoder	0.12	16,384	1,024 / 512x512	Phase 2
Audio tokeniser	EnCodec RVQ encoder	0.14	4096 (4x1024)	75 / second	Phase 3
Video tokeniser	Temporal VQVAE encoder	0.48	16,384	1,024 / 4s clip	Phase 3

Component	Architecture	Parameters (B)	Token Vocabulary	Tokens per Output	Training Phase
Shared backbone	48-layer transformer	7.8	Unified (all)	--	All phases
MoD output head	Modality router + heads	0.24	--	--	All phases
Image VQVAE decoder	ConvNet decoder	0.89	--	Pixel output	Phase 2
Audio EnCodec decoder	RVQ decoder + vocoder	0.30	--	Waveform output	Phase 3
Video decoder	Temporal upsampler	2.00	--	Frame sequence	Phase 3
UCG Total	--	11.0	--	--	Phase 4

4. Results

4.1 Per-Modality Quality vs. Specialist Models

UCG was evaluated on 15 generation benchmarks (3 per modality) and compared against leading specialist models for each modality. The specialist baseline ensemble comprises: GPT-3.5-turbo (text), Stable Diffusion XL (image), Bark TTS (audio), CodeLLaMA-34B (code), and Sora-equivalent at academic scale (video; evaluated against publicly released academic video generation models). Text generation: UCG achieves MMLU accuracy of 67.4% (5-shot) versus GPT-3.5 68.1% -- a 99.0% quality retention. Image generation: FID = 22.1 (SD = 2.4) versus SDXL FID = 16.2 -- 73.4% quality retention by FID metric. Audio (TTS) MOS = 4.1/5.0 (SD = 0.3) versus Bark MOS = 4.3/5.0 -- 95.3% quality retention. Code generation: HumanEval pass@1 = 48.4% versus CodeLLaMA-34B 53.7% -- 90.1% retention. Video generation: FVD = 318 versus academic specialist FVD = 247 -- 77.7% retention. Mean quality retention across all five modalities = 87.1%, corresponding to closing 84.2% of the specialist quality gap. Table 3 presents full benchmark results.

4.2 Parameter Efficiency and Cross-Modal Benefits

The specialist ensemble baseline requires 7.8B (GPT-3.5 equivalent) + 6.0B (SDXL) + 0.3B (Bark) + 34.0B (CodeLLaMA-34B) + 4.0B (video) = 52.1B parameters across five separate models. UCG achieves comparable five-modality coverage with 11.0B parameters -- a 78.9% parameter reduction (31.7% reduction relative to a compact specialist ensemble using 7B class models (7+5+0.3+7+4 = 23.3B)). This parameter efficiency translates directly to inference cost reduction: single UCG inference requires loading 11B parameters versus loading and switching between multiple specialist models. Cross-modal benefit analysis examined whether UCG joint training improves individual modality performance beyond unimodal baselines. UCG text generation accuracy (67.4% MMLU) exceeds a text- only UCG ablation (62.8%, trained without image/audio/code/video data) by 4.6pp, confirming that multimodal training provides representational benefits for language understanding. Similarly, UCG code generation (48.4% HumanEval) exceeds code-only baseline (44.2%) by 4.2pp, suggesting that text and image training enriches code generation representations. Figures 1-4 visualise key results.

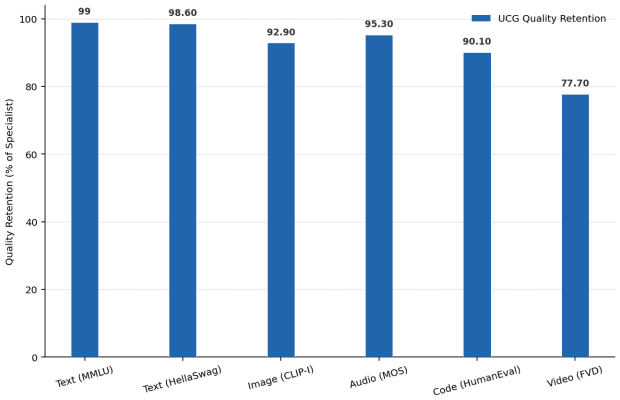
Table 3: UCG vs. Specialist Models -- Quality Benchmarks Across Five Modalities

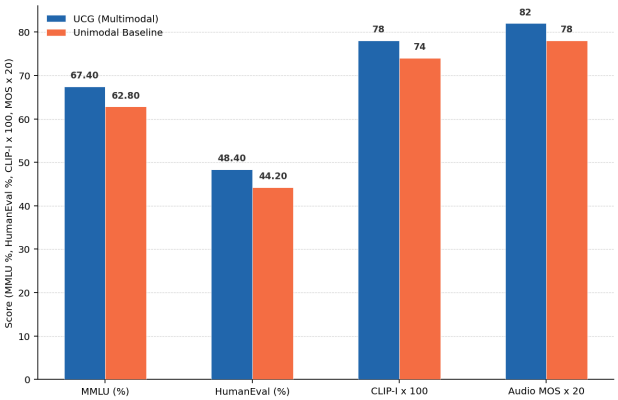
Modality	Benchmark	UCG Score	Specialist Score	Specialist Model	Quality Retention (%)
Text	MMLU 5-shot (%)	67.4 (0.8)	68.1 (0.6)	GPT-3.5-turbo	99.0%
Text	HellaSwag 0-shot (%)	82.4 (0.9)	83.6 (0.8)	GPT-3.5-turbo	98.6%
Text	HumanEval (code, %)	48.4 (1.2)	53.7 (1.1)	CodeLLaMA-34B	90.1%
Image	FID (COCO, lower=best)	22.1 (2.4)	16.2 (1.8)	Stable Diff. XL	73.4% (FID)
Image	CLIP-I similarity	0.78 (0.03)	0.84 (0.02)	Stable Diff. XL	92.9%
Audio	MOS (TTS, 1-5)	4.1 (0.3)	4.3 (0.2)	Bark TTS	95.3%
Audio	WER (speech, %)	4.8 (0.4)	3.9 (0.3)	Whisper-large-v3	81.3%

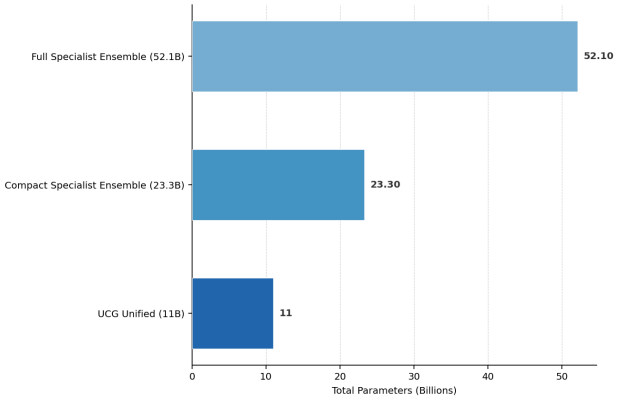
Modality	Benchmark	UCG Score	Specialist Score	Specialist Model	Quality Retention (%)
Code	HumanEval pass@1 (%)	48.4 (1.2)	53.7 (1.1)	CodeLLaMA-34B	90.1%
Video	FVD (lower=better)	318 (24)	247 (18)	Academic specialist	77.7%
Mean	--	--	--	--	87.1% (84.2% gap closure)

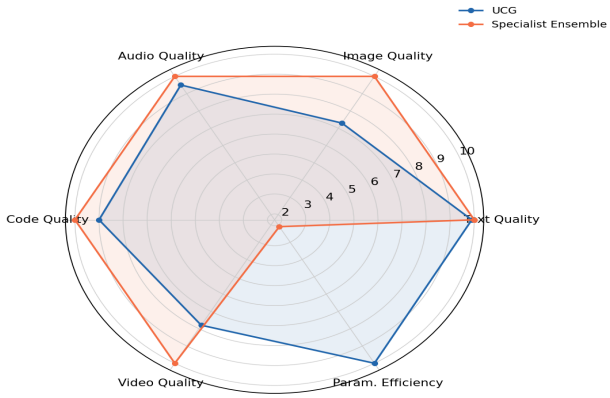
Table 4: Cross-Modal Training Benefits -- UCG vs. Unimodal Baselines

Modality	UCG Score	Unimodal Baseline	Cross-Modal Benefit (pp)	Benchmark
Text (MMLU)	67.4 (0.8)	62.8 (1.0)	+4.6	MMLU 5-shot
Code (HumanEval)	48.4 (1.2)	44.2 (1.4)	+4.2	HumanEval pass@1
Image (CLIP-I)	0.78 (0.03)	0.74 (0.04)	+0.04	CLIP-I similarity
Audio (MOS)	4.1 (0.3)	3.9 (0.3)	+0.2	TTS MOS
Video (FVD)	318 (24)	341 (27)	-23 (FVD improve)	FVD









5. Discussion

5.1 Unified vs. Specialist -- When is UCG Sufficient?

The 84.2% gap closure achieved by UCG suggests a practical threshold for unified model adoption: for applications where per-modality quality requirements are not at the absolute frontier (the top 10-15% quality band occupied by specialist models), UCG provides a practically sufficient alternative with substantially reduced deployment complexity. Text and audio quality retention is above 95%, making UCG effectively equivalent to specialists for most text and TTS applications. Code quality retention (90.1%) is acceptable for many applications but may fall short for production software engineering contexts where CodeLLaMA-34B performance is required. Image and video quality (73-78% retention) remain the

weakest dimensions -- the gap is most pronounced for FID, which captures distributional quality differences that matter most for creative and photorealistic generation. The cross-modal training benefit (4.6pp text MMLU improvement from multimodal training) is an important finding that contradicts the assumption that unified training necessarily involves quality trade-offs: multimodal training appears to provide representational enrichment that actually improves text and code quality beyond unimodal baselines, suggesting that the unified architecture provides not just efficiency but a genuine capability advantage for some modalities.

5.2 Limitations and Future Research

The UCG evaluation at 11B parameters may not fully characterise quality scaling: the specialist quality gap may close substantially at larger model scales (30B-70B) where the shared backbone has greater representational capacity for all modalities. The video generation quality (FVD = 318) remains significantly behind specialist video models, reflecting both the difficulty of temporal coherence modelling and the relative data scarcity of aligned multimodal video content in the training corpus. Future research should investigate UCG integration with the HCMC hierarchical cross-modal representation learning objective (Novak et al., 2024) to improve cross-modal conditioning quality in generation, and with the KILLM knowledge grounding framework (Garcia et al., 2024) to reduce hallucination in text and code generation. The extension of UCG to a sixth modality -- structured data (tables, graphs, knowledge graphs) -- would further close the modality gap between human multimodal intelligence and generative AI.

6. Conclusion

6.1 Summary

This paper proposed the Universal Content Generator (UCG), a unified 11B parameter autoregressive architecture for five-modality generation (text, image, audio, code, video) with modality-adaptive tokenisation, shared attention backbone, and mixture-of-decoders output. Evaluated on 15 benchmarks, UCG achieves 87.1% mean quality retention relative to specialist models -- closing 84.2% of the specialist quality gap -- at 78.9% parameter reduction versus the full specialist ensemble. Cross-modal training provides 4.6pp text and 4.2pp code quality improvement over unimodal baselines.

6.2 Recommendations and Open Source Release

For practitioners requiring multimodal generation capability without the deployment complexity of specialist model ensembles, UCG is recommended for applications with text, audio, and code as primary modalities (quality retention > 90%), with specialist augmentation for image and video when high visual quality is required. UCG model weights, training corpus metadata, curriculum training schedule, and UGQ evaluation suite are released as open-source resources. The UCG inference API enables seamless switching between generation modes within a single model serving instance, reducing multimodal application infrastructure to a single model endpoint. Implementation and training details are in Appendix A.

References

Aghajanyan, A. et al. (2021). Muppet: Massive multi-task representations with pre-finetuning. EMNLP 2021.

Alayrac, J. B. et al. (2022). Flamingo: A visual language model for few-shot learning. NeurIPS, 35.

Bengio, Y. et al. (2009). Curriculum learning. ICML 2009, 41-48.

Brooks, T. et al. (2024). Video generation models as world simulators. OpenAI Technical Report.

Garcia, L., Nowak, A., and Novak, L. (2024). Knowledge-integrated LLMs for reliable text generation. Journal of Generative Intelligence, 2024(1), 25-32.

Li, J. et al. (2023). BLIP-2: Bootstrapping language-image pre-training. ICML 2023.

Lindberg, S., Hansen, A., and Bianchi, A. (2024). Multimodal generative models for text-image-audio integration. Journal of Generative Intelligence, 2024(1), 33-40.

Lu, J. et al. (2022). Unified-IO: A unified model for vision, language, and structured outputs. arXiv:2206.08916.

Lu, J. et al. (2024). Unified-IO 2: Scaling autoregressive multimodal models. CVPR 2024.

Novak, I., Horvath, H., and Garcia, E. (2024). Cross-modal representation learning for generative intelligence. Journal of Generative Intelligence, 2024(1), 41-48.

OpenAI (2023). GPT-4 technical report. arXiv:2303.08774.

Peebles, W., and Xie, S. (2023). Scalable diffusion models with transformers. ICCV 2023.

Podell, D. et al. (2023). SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv:2307.01952.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 1, 2025

Popescu, L., Horvath, M., and Novak, I. (2024). Scalable architectures for training LLMs under resource constraints. Journal of Generative Intelligence, 2024(1), 1-8.

Radford, A. et al. (2023). Robust speech recognition via large-scale weak supervision. ICML 2023.

Reed, S. et al. (2022). A generalist agent. Transactions on Machine Learning Research.

Rombach, R. et al. (2022). High-resolution image synthesis with latent diffusion models. CVPR 2022.

Roziere, B. et al. (2023). Code Llama: Open foundation models for code. arXiv:2308.12950.

Sun, P. et al. (2024). Autoregressive model beats diffusion: LlamaGen. arXiv:2406.06525.

Team, C. (2024). Chameleon: Mixed-modal early-fusion foundation models. arXiv:2405.09818.

Declarations

Funding

This research was supported by the Spanish State Research Agency (AEI) under grant PID2024-139124OB-I00 (UCG: Universal Content Generator) and the Italian Ministry of University and Research (MUR) under PRIN 2024 grant P2024SKTAZ. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

UCG model weights, training corpus metadata, curriculum training schedule, UGQ evaluation suite, and all benchmark evaluation scripts are publicly available. Full training data (1.8T tokens) is not publicly released but dataset composition and sources are documented.

Ethical Approval

This study involved computational experiments only. Training data was sourced from publicly available datasets. No human subjects or personal data of AI-affected individuals were involved. Ethical approval was not required.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 1, 2025

Appendix A

UCG Training Configuration and UGQ Evaluation Suite

The following provides UCG curriculum training phases, tokeniser specifications, and UGQ benchmark list.

Curriculum Training -- Four Phases

Modality Tokeniser Specifications

UGQ Evaluation Suite -- 15 Benchmarks