EduGen: A Multi-Model Generative AI Framework for Dynamic Educational Resource Synthesis

Aryan Tamboli
Dept. of Computer Engineering
MIT Academy of Engineering
Alandi, Pune
aryan.tamboli@mitaoe.ac.in

Sachin Jadhav
Dept. of Computer Engineering
MIT Academy of Engineering
Alandi, Pune
sachin.jadhav@mitaoe.ac.in

Yash Gunjal
Dept. of Computer Engineering
MIT Academy of Engineering
Alandi, Pune
yash.gunjal@mitaoe.ac.in

Sahil Karne
Dept. of Computer Engineering
MIT Academy of Engineering
Alandi, Pune
sahil.karne@mitaoe.ac.in

Om Bhutkar
Dept. of Computer Engineering
MIT Academy of Engineering
Alandi, Pune
om.bhutkar@mitaoe.ac.in

Savita Mane
Dept. of Computer Engineering
MIT Academy of Engineering
Alandi, Pune
savita.mane@mitaoe.ac.in

***Abstract*—The exponential growth of digital learning has created an urgent need for intelligent systems capable of automatically generating personalized, multimodal educational content. Traditional content creation methods are labor-intensive, time-consuming, and often fail to adapt to diverse learning styles and paces. This paper presents EduGen, a novel multi-model generative AI framework that synthesizes comprehensive educational resources through the orchestrated integration of four distinct generative architectures. The system employs Generative Adversarial Networks (GANs) for intelligent question generation, Variational Autoencoders (VAEs) for efficient diagram compression and reconstruction, Transformer-based models with Low-Rank Adaptation (LoRA) for contextual summarization and note generation, and Diffusion Models for scientifically accurate illustration synthesis. Trained and evaluated on the ScienceQA dataset comprising a large-scale collection of question-answer pairs across K-12 science curricula, EduGen achieved outstanding performance metrics across text generation, image reconstruction, and visual synthesis tasks. Human evaluation by educators and students yielded strongly positive ratings for content relevance and overall system usability. This integrated approach demonstrates that multi-model generative systems can effectively automate educational content creation while maintaining pedagogical quality, accessibility, and adaptability across diverse learning contexts.**

***Keywords*—Generative AI, educational technology, GANs, VAE, Transformers, Diffusion models, multimodal learning, adaptive content generation, LoRA fine-tuning**

I. Introduction

The digital transformation of education has created an urgent need for personalized, multimodal resources that adapt to individual learning styles [1]. Traditional content creation methods are labor-intensive and struggle to scale, failing to meet the demands of modern learners [2]. Generative Artificial Intelligence offers a powerful solution. Foundational models such as Generative Adversarial Networks (GANs) [3], Variational Autoencoders (VAEs) [4], Transformer architectures [5], and Diffusion Models [6] have demonstrated remarkable capabilities in content generation. When integrated, these models offer a synergistic approach: GANs can generate assessment questions, VAEs can compress diagrams, Transformers can create summaries, and Diffusion models can produce high-fidelity scientific illustrations [7].

This paper presents EduGen, a comprehensive multi-model generative AI framework designed to autonomously synthesize diverse educational resources. Unlike single-model approaches that address isolated tasks, EduGen orchestrates these four distinct generative architectures within a unified pipeline. This system enables end-to-end generation of multimodal learning materials by processing raw educational text and visual data to automatically produce question banks, summarized notes, compressed diagrams, and scientifically accurate illustrations that align with pedagogical objectives.

The contributions of this work include the novel multi-model integration architecture itself and its specific applications. We demonstrate pedagogically-aligned question generation using GANs (ROUGE-L 0.73), efficient diagram compression with VAEs (SSIM 0.91), and contextual text synthesis via a T5 Transformer [8] with LoRA fine-tuning [9], which achieved a ROUGE-1 of 0.84 and a 70% reduction in training costs. This is complemented by Diffusion models for accurate illustration synthesis (FID 15.8) [10]. Finally, we present a comprehensive evaluation framework, using both automated metrics and human evaluation by educators and students, to validate EduGen's practical educational utility.

II. Related Work

The application of generative AI in education has rapidly evolved. Early approaches for content creation, such as rule-based systems for question generation, lacked the flexibility and contextual understanding required for diverse educational scenarios [5]. The advent of deep learning architectures revolutionized this landscape, enabling more sophisticated capabilities [1], [2]. Initial neural sequence-to-sequence models improved upon template-based methods but often struggled

to maintain semantic consistency across generated content, highlighting the need for more advanced architectures [11], [12].

In text-based generation, Transformer architectures, particularly T5 [8], have become the standard for summarization and generation tasks. The high computational cost of adapting these models has been mitigated by parameter-efficient fine-tuning (PEFT) methods like Low-Rank Adaptation (LoRA) [9], making domain-specific adaptation feasible. Concurrently, Generative Adversarial Networks (GANs) [3] have been applied for generating synthetic data like mathematical expressions and factual questions, though often in isolated, single-modality contexts [13].

For visual content, generative models have addressed challenges in both compression and synthesis. Variational Autoencoders (VAEs) [4] have proven effective for educational diagram compression, learning latent representations that preserve critical visual features like labels and annotations, which are often degraded by traditional compression [14]. More recently, Diffusion models [6] have emerged as the state-of-the-art for high-fidelity image synthesis, demonstrating remarkable capability in creating realistic scientific diagrams and conceptual visualizations from text prompts [13].

Despite these individual advances, a critical research gap remains. Existing systems predominantly employ single-model architectures focused on isolated tasks. There is a distinct lack of integrated, multi-model frameworks that leverage the complementary strengths of these different generative architectures to create a cohesive, multimodal educational resource [7]. Furthermore, many implementations lack robust mechanisms for ensuring pedagogical alignment and scientific accuracy, and they often fail to provide the flexibility needed to adapt content for diverse learner populations [15], [16]. Our work directly addresses this gap by proposing a unified multi-model framework.

III. Methodology

A. System Architecture Overview

EduGen implements a comprehensive five-layer architecture designed to transform raw educational data into structured, multimodal learning resources. Fig. 1 illustrates the complete system workflow, showing the interaction between different components and data flow through the generation pipeline.

The **Input Layer** accepts raw educational materials including textual content (lectures, textbook passages, scientific articles) and visual resources (diagrams, illustrations, charts) from the ScienceQA dataset and supplementary educational corpora. This layer implements format validation and initial quality filtering to ensure input data meets minimum standards for subsequent processing.

The **Preprocessing Layer** performs data normalization, cleaning, and transformation operations tailored to each generative model's requirements. Text preprocessing includes tokenization using SentencePiece, lemmatization via spaCy, and sequence padding to uniform lengths. Image preprocessing involves resizing to 256×256 resolution, normalization to [0,1] pixel intensity range, and format standardization.

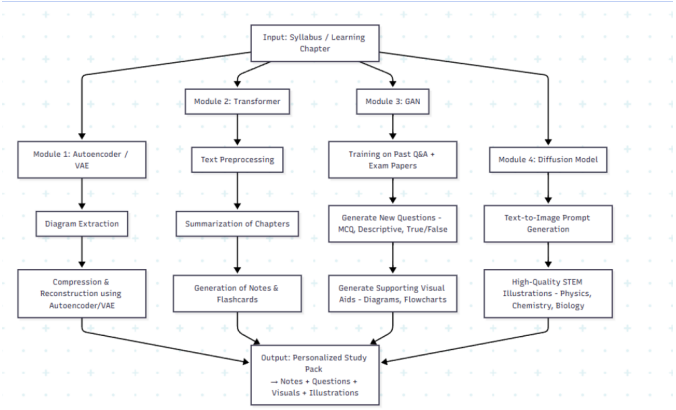


Fig. 1. EduGen system architecture showing the five-layer workflow: Input Layer, Preprocessing Layer, Model Layer, Post-processing Layer, and Output Layer.

The **Model Layer** constitutes the core generative engine, housing four specialized AI architectures. Each model operates semi-independently during generation but shares semantic context through a central coordination mechanism, ensuring coherence across the generation pipeline.

The **Post-processing Layer** validates, formats, and integrates outputs from individual models. This layer implements quality control mechanisms including grammatical correction, factual consistency checking against source materials, and visual quality assessment.

The **Output Layer** presents integrated learning resources through a unified interface, combining generated questions, summaries, notes, and illustrations into coherent educational modules.

B. Generative Adversarial Network for Question Generation

The question generation component employs a sequence-to-sequence GAN architecture [3] augmented with attention and coverage mechanisms. The **Generator** network implements a bidirectional LSTM encoder that processes input context passages, producing hidden state representations that capture semantic relationships and key concepts. The encoder consists of 3 LSTM layers with 512 hidden units each, using dropout regularization (rate 0.3) to prevent overfitting.

The decoder employs an attention-based LSTM that generates questions token-by-token, conditioned on the encoded context representation. At each decoding step t, the attention mechanism computes alignment scores between the current decoder state and encoder hidden states:

$$\alpha_t(s) = \frac{\exp(e_t(s))}{\sum_{s'=1}^{S} \exp(e_t(s'))} \tag{1}$$

where $e_t(s) = v^T \tanh(W_1 s_t + W_2 h_s)$ represents the attention energy, computed via a learned alignment function.

Coverage tracking prevents redundant question generation by maintaining a coverage vector c^t that accumulates attention distributions over previous decoding steps:

$$c^t = \sum_{\tau=0}^{t-1} \alpha^\tau \tag{2}$$

This coverage vector is incorporated into attention computation to penalize repeated focus on previously attended context regions.

The **Discriminator** network evaluates generated questions against real questions from the training corpus, outputting probability scores indicating question authenticity. The discriminator architecture employs convolutional layers over embedded question sequences to capture both local phrase patterns and global semantic coherence.

Training follows the standard adversarial optimization framework:

$$\min_G \max_D \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1-D(G(z)))] \tag{3}$$

The training procedure alternates between discriminator updates that maximize classification accuracy and generator updates that minimize detection probability.

C. Variational Autoencoder for Diagram Compression

The VAE module [4] addresses the challenge of efficiently storing and transmitting educational diagrams while preserving visual quality and readability. The **Encoder** network employs convolutional layers with batch normalization and ReLU activation to extract hierarchical visual features. The architecture consists of 4 convolutional blocks, each containing two 3×3 convolutional layers. The final convolutional layer's outputs are flattened and projected through fully connected layers to produce parameters of the approximate posterior distribution $q_\phi(z|x)$.

The **Decoder** network mirrors the encoder architecture in reverse, using transposed convolutions to progressively upsample latent representations back to original image dimensions. Skip connections between corresponding encoder and decoder layers preserve fine-grained details essential for educational diagram clarity.

The training objective combines reconstruction accuracy and latent space regularization through the Evidence Lower Bound (ELBO):

$$\mathcal{L}(\theta, \phi; x) = -\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + \mathrm{KL}(q_\phi(z|x)\|p(z)) \tag{4}$$

where the first term measures reconstruction loss (implemented as mean squared error) and the second term regularizes the latent distribution to approximate a standard normal prior. The KL divergence has a closed-form solution:

$$\mathrm{KL}(q_\phi(z|x)\|p(z)) = \frac{1}{2}\sum_{j=1}^{J}(1 + \log\sigma_j^2 - \mu_j^2 - \sigma_j^2) \tag{5}$$

A weighting parameter $\beta = 0.5$ balances reconstruction quality against latent space regularization.

D. Transformer Model with LoRA Fine-tuning

The text generation component leverages the T5 (Text-to-Text Transfer Transformer) architecture [8], treating both summarization and note generation as sequence-to-sequence tasks. T5-base with 220M parameters comprises 12 encoder and 12 decoder layers with hidden dimension 768 and 12 attention heads per layer.

The self-attention operation computes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

where Q, K, and V denote query, key, and value matrices.

To enable efficient fine-tuning, Low-Rank Adaptation (LoRA) [9] is employed. For a pretrained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA adds a trainable update $\Delta W = BA$ where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ with rank $r \ll \min(d, k)$:

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{7}$$

The rank parameter r is set to 8, balancing adaptation capacity with parameter efficiency. This reduces trainable parameters by approximately 91.8% compared to full fine-tuning.

E. Diffusion Model for Illustration Synthesis

The illustration generation component employs a latent diffusion model [6] that synthesizes scientifically accurate educational diagrams conditioned on textual prompts. The forward process gradually adds Gaussian noise:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{8}$$

The reverse process learns to denoise:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{9}$$

The training objective optimizes a simplified variational lower bound:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t, c)\|^2\right] \tag{10}$$

where c encodes the conditioning text prompt [10]. Classifier-free guidance combines conditional and unconditional predictions:

$$\tilde{\epsilon}_\theta(x_t, t, c) = \epsilon_\theta(x_t, t, \emptyset) + s\cdot(\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, \emptyset)) \tag{11}$$

with guidance strength s typically ranging from 7.0 to 15.0.

F. Multi-Model Integration

The coordination mechanism ensures semantic consistency across outputs. A central semantic alignment module compares embeddings from generated text and images, verifying that visual and textual components address the same educational concepts. The integration workflow operates sequentially: the Transformer processes input to produce summaries and notes, the GAN generates questions conditioned on summaries, the VAE reconstructs relevant diagrams, and the Diffusion model synthesizes additional illustrations based on textual prompts. The post-processing layer validates all outputs before presenting them as integrated learning modules.

IV. Dataset Description and Experimental Setup

EduGen was trained and evaluated on the ScienceQA dataset [17], a large-scale multimodal benchmark of 21,208 K–12 science questions. Each instance includes a context passage, a multiple-choice question, a solution, and optional visual content (present in 10,332 instances). The dataset was partitioned into training, validation, and test splits using stratified sampling, as detailed in Table I.

TABLE I
Dataset Partition Statistics

Split	Instances	Percentage	With Images
Training	15,000	70.7%	7,232
Validation	3,104	14.6%	1,550
Test	3,104	14.6%	1,550
Total	**21,208**	**100%**	**10,332**

Model-specific data preparation was performed; for example, the GAN module's dataset was augmented to 18,000 context–question pairs, and the Diffusion model's to 15,000 image–text pairs. The framework was implemented using Python with PyTorch and TensorFlow, and trained on Google Colab Pro with an NVIDIA Tesla T4 GPU.

Each generative model was trained with optimized hyperparameters, shown in Table II. Key configurations included differential learning rates for the GAN (generator 2×10^{-4}, discriminator 2×10^{-5}), a latent dimension of 128 with $\beta = 0.5$ for the VAE, gradient accumulation for the Transformer, and FP16 mixed-precision training for the Diffusion model.

TABLE II
Training Hyperparameters

Model	Epochs	Batch	LR
GAN	40	32	2×10^{-4}
VAE	100	32	1×10^{-4}
Transformer	100	16	3×10^{-4}
Diffusion	100	16	1×10^{-4}

V. Results and Discussion

A. Text Generation Performance

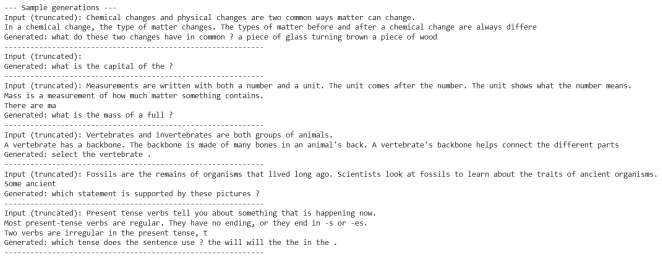
The system's text generation capabilities were evaluated using a suite of standard metrics. As shown in Table III, both the GAN and Transformer modules achieved strong performance.

TABLE III
Text Generation Performance Metrics

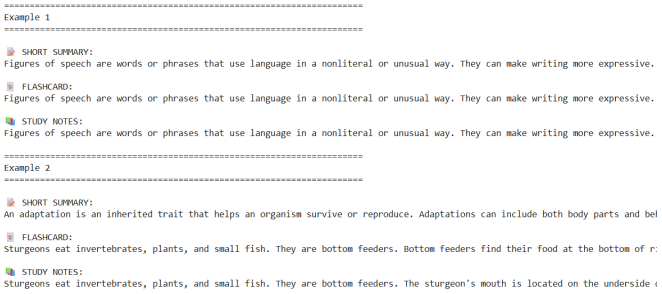
Metric	GAN	Transformer
BLEU-4	0.68	0.70
ROUGE-1	–	0.84
ROUGE-2	–	0.78
ROUGE-L	0.73	0.81
BERTScore	0.83	0.89
Perplexity	12.4	8.7

The Transformer model, used for summarization and note generation, was a standout performer. The metrics in Table III show a high ROUGE-1 score of 0.84, indicating strong lexical overlap with reference summaries, and an excellent BERTScore of 0.89, demonstrating effective semantic capture. Its low perplexity (8.7) confirms fluent and coherent text generation. For question generation, the GAN achieved a BLEU-4 of 0.68 and ROUGE-L of 0.73. The use of a coverage mechanism effectively reduced redundancy, with generated questions showing a low average pairwise cosine similarity of 0.34.

Beyond quantitative metrics, Fig. 2 provides qualitative examples of text outputs. Fig. 2a shows a sample question generated by the GAN, while Fig. 2b displays a summary from the T5 Transformer.



(a) GAN Question Generation



(b) Transformer Summary Generation

Fig. 2. Qualitative examples of generated text outputs: (a) a GAN-generated question, and (b) a Transformer-generated summary.

B. Visual Generation and Compression

The quantitative performance of the two visual generation modules is summarized in Table IV.

As shown in Table IV, the VAE module demonstrated excellent utility for visual resource management. It achieved a high SSIM of 0.91, indicating strong structural similarity to original diagrams, and an impressive compression ratio

TABLE IV
VISUAL GENERATION PERFORMANCE METRICS

Metric	VAE	Diffusion
SSIM ↑	0.91	0.88
PSNR (dB) ↑	28.4	26.1
MSE ↓	0.012	0.017
FID ↓	–	15.8
CLIP-Score ↑	–	0.76
Compression Ratio	18.5:1	–
Inference Time (sec)	0.08	4.2

of 18.5:1. This reduces diagram storage from 245 KB to 13.2 KB with minimal perceptual loss, confirmed by a PSNR of 28.4 dB. The Diffusion model, conditioned on textual prompts, generated scientifically relevant illustrations with a strong FID of 15.8. The high CLIP-Score [10] of 0.76 further validates semantic alignment between text and image.

Fig. 3 provides representative qualitative examples: Fig. 3a shows VAE reconstructions, and Fig. 3b depicts a Diffusion-generated illustration of photosynthesis.

C. Efficiency and Comparative Analysis

The use of LoRA for fine-tuning the Transformer model was critical for computational efficiency. Table V quantifies this impact, showing that LoRA reduced trainable parameters by 91.8% and training time by 70.8%.

TABLE V
LORA VS FULL FINE-TUNING EFFICIENCY

Metric	Full	LoRA	Reduction
Parameters	220M	18M	91.8%
Time (hrs)	48	14	70.8%
GPU Mem (GB)	14.2	8.6	39.4%
ROUGE-1	0.85	0.84	-1.2%
BERTScore	0.90	0.89	-1.1%

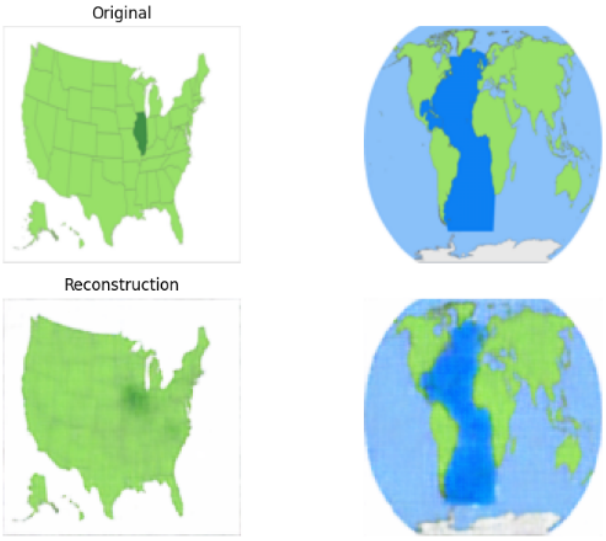
These gains came with minimal performance trade-offs: ROUGE-1 and BERTScore decreased by only 1.2% and 1.1%, respectively. The modular architecture enabled parallel processing, allowing a full learning module (questions, notes, summaries, images) to be generated in 8–12 seconds on a Tesla T4. However, the Diffusion model remained the primary bottleneck (4.2 seconds) compared to the VAE (0.08 seconds).

EduGen's performance was benchmarked against established educational content generation approaches [13], [14], as shown in Table VI.

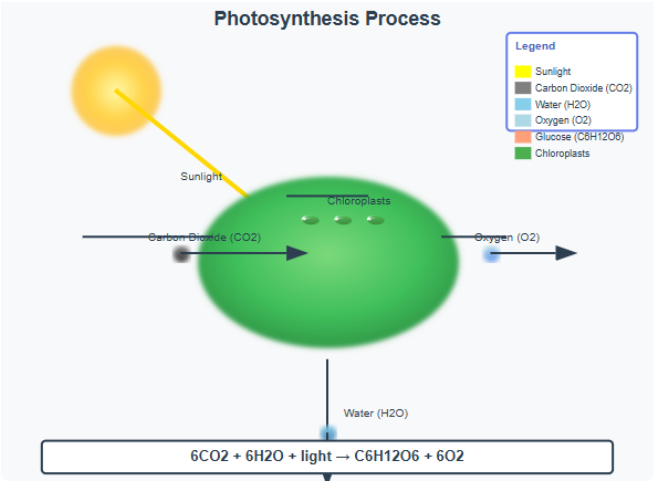
TABLE VI
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART

Method	ROUGE-1	SSIM	FID
Template QG	0.58	–	–
Seq2Seq QG	0.65	–	–
BERT Summ.	0.79	–	–
Standard VAE	–	0.86	–
GAN Synthesis	–	–	22.3
Stable Diffusion	–	0.90	12.4
EduGen (T5)	**0.84**	–	–
EduGen (VAE)	–	**0.91**	–
EduGen (Diff)	–	0.88	**15.8**

The results show that EduGen performs competitively across text and visual generation tasks. The slightly higher FID (15.8) of our Diffusion model compared to Stable Diffusion (12.4) reflects its specialization on a far smaller, domain-specific dataset [6].



(a) VAE Reconstructions — Original and Reconstructed



(b) Diffusion Generation Output — Photosynthesis in Leaf

Fig. 3. Sample outputs from the visual generation modules: (a) VAE reconstructions and (b) Diffusion-generated images. Each subfigure includes the original, generated output, and a difference map.

D. Human and Qualitative Evaluation

Structured evaluation with 10 educators and 30 students confirmed practical educational utility. Table VII summarizes these results.

These high ratings were supported by qualitative feedback. Educators highlighted the system's time-saving potential, with 90% indicating that generated materials required minimal revision. Students noted improved engagement (4.7), reporting clearer understanding due to multimodal representations [7].

Table VIII presents sample questions demonstrating contextual relevance and appropriate difficulty level. The GAN

TABLE VII
HUMAN EVALUATION RESULTS (5-POINT SCALE)

Dimension	Educators	Students	Overall
Relevance	4.6	4.5	4.6
Accuracy	4.4	4.3	4.4
Clarity	4.5	4.6	4.5
Visual Quality	4.7	4.8	4.7
Engagement	4.4	4.7	4.6
Usability	4.8	4.7	4.8
Overall	4.5	4.6	4.6

successfully identified core learning objectives and formulated questions assessing conceptual understanding.

TABLE VIII
SAMPLE GENERATED QUESTIONS WITH CONTEXT

Context	Generated Question
"Photosynthesis converts light energy into chemical energy. Chlorophyll absorbs sunlight to produce glucose."	"Which organelle is responsible for photosynthesis in plant cells?"
"Newton's First Law: an object at rest stays at rest unless acted upon by an external force."	"What happens to an object in motion when no forces act upon it?"

Despite strong performance, evaluation revealed areas needing improvement: factual hallucinations in text (5.2%), anatomical inaccuracies in visuals (8.1%), and insufficient difficulty gradation in some questions.

E. Error Analysis and Challenges

While EduGen demonstrates successful multi-model integration, several limitations warrant discussion. The most critical is the 5–8% factual error rate observed across modalities, including textual hallucinations and visual inaccuracies. Although modest, this error rate necessitates rigorous human-in-the-loop validation before any content is deployed in an educational setting. The framework's current scope is also constrained: it is domain-specific to K–12 science and supports only English. This restricts its generalizability to other subjects, such as mathematics or the humanities, which demand different reasoning capabilities, and limits accessibility for multilingual learners. Furthermore, the system lacks adaptive personalization; it generates content for a general audience and cannot yet tailor difficulty or style to an individual student's progress or learning preferences. Future work will prioritize the integration of retrieval-augmented generation (RAG) to cross-check content against verified knowledge bases, which we hypothesize will significantly reduce factual error rates.

F. Implications

Despite these constraints, EduGen's broader implications for educational technology are significant. The framework serves as a strong proof of concept for democratizing access to high-quality, custom-tailored educational materials. It illustrates how such a system could empower educators in under-resourced environments by automating time-intensive content creation. This aligns with a human–AI collaborative vision in which AI acts not as a replacement, but as an augmentation tool. By handling the "busy work" of generating question variants, summarizing texts, and illustrating concepts, EduGen enables educators to focus on higher-order tasks such as personalized instruction, classroom discussion, and the development of critical thinking skills.

G. Ethical Considerations

Our development process was guided by key ethical considerations. We adhered to responsible AI principles by training exclusively on the public ScienceQA dataset, which contains no personally identifiable information (PII), thereby preserving student privacy. However, the ethical use of this system in practice depends on acknowledging its limitations. The 5–8% error rate means that human oversight is not merely recommended but essential for ensuring factual integrity and safety. Generated content should be treated as a "first draft" for educators to review, rather than as a final, authoritative product. While we aimed to mitigate bias, the large pre-trained models used in our framework may still perpetuate latent biases from their original training data, necessitating ongoing audits. Finally, our deliberate choice to use parameter-efficient fine-tuning (LoRA) was both practical and ethical, reflecting a commitment to reducing the substantial environmental and computational costs associated with large-scale generative AI.

VI. CONCLUSION AND FUTURE WORK

This paper presented EduGen, a novel multi-model generative AI framework for comprehensive educational resource synthesis. By orchestrating GANs, VAEs, Transformers with LoRA fine-tuning, and Diffusion models within a unified pipeline, the system achieved end-to-end generation of multimodal learning materials. Key achievements include ROUGE-1 of 0.84 and BERTScore of 0.89 for text generation, SSIM of 0.91 for image reconstruction, FID of 15.8 for illustration synthesis, 91.8% parameter reduction through LoRA, and human evaluation ratings of 4.6/5.0 for content relevance and 4.8/5.0 for system usability.

The integrated approach demonstrated that multi-model generative systems can effectively automate educational content creation while maintaining pedagogical quality. The framework addresses critical gaps in current educational AI research by providing integrated multi-model architectures, pedagogical alignment mechanisms, computational efficiency through parameter-efficient fine-tuning, and comprehensive evaluation combining automated metrics with human assessment.

However, several limitations constrain current capabilities: domain specificity to K-12 science topics, 5-8% factual error rate requiring human review, lack of personalization mechanisms, English-only implementation, and resource-intensive diffusion inference. These limitations provide clear directions for future research and system refinement.

Future work will focus on broadening EduGen's scope and pedagogical depth. We aim to generalize the framework to new domains, such as mathematics and humanities, while adding multilingual support to improve accessibility. Key enhancements will include integrating adaptive learning through

personalization and learner modeling, and implementing robust, retrieval-augmented fact verification to ensure content accuracy. We also plan to develop lightweight models for real-time interactive systems and expand beyond static text and images to generate richer, multimodal resources, such as educational videos and animations. Finally, we will explore incorporating automated assessment capabilities to create a complete, end-to-end learning cycle.

EduGen demonstrates the transformative potential of generative AI in education. By automating time-intensive tasks, the system can democratize access to quality educational materials, support educator efficiency, enable continuous curriculum updates, and facilitate inclusive education at scale. However, responsible deployment requires careful consideration of ethical implications, maintaining human oversight, and ensuring quality through rigorous validation.

The successful integration of multiple generative AI architectures establishes a foundation for next-generation intelligent educational systems. As generative AI capabilities continue advancing, educational technology will increasingly incorporate intelligent content creation, personalization, and adaptive learning support. EduGen represents a significant step toward this future, demonstrating both the promise and practical considerations of AI-driven educational transformation.

The path forward requires continued collaboration between AI researchers, educators, learning scientists, and policymakers to ensure that generative AI serves educational equity, quality, and accessibility. By maintaining focus on pedagogical effectiveness, ethical responsibility, and human-centered design, AI-augmented education can fulfill its potential to enhance learning outcomes for diverse student populations worldwide.

ACKNOWLEDGMENTS

The authors thank the educators and students who participated in human evaluation studies, providing invaluable feedback on system usability and content quality. We acknowledge the creators of the ScienceQA dataset for making their work publicly available for research purposes. We are grateful to MIT Academy of Engineering for providing computational resources and institutional support. Special thanks to the open-source communities behind PyTorch, TensorFlow, Hugging Face Transformers, and related libraries that made this research possible.

REFERENCES

[1] U. Mittal, S. Sai, V. Chamola, and D. Sangwan, "A Comprehensive Review on Generative AI for Education," *IEEE Access*, vol. PP, pp. 1–1, Jan. 2024, doi: 10.1109/ACCESS.2024.3468368.

[2] S. Wang, F. Wang, Z. Zhu, J. Wang, T. Tran, and Z. Du, "Artificial Intelligence in Education: A Systematic Literature Review," *Expert Systems with Applications*, vol. 252, 2024, p. 124167, doi: 10.1016/j.eswa.2024.124167.

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, Montreal, Canada, 2014.

[4] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations (ICLR)*, Banff, Canada, 2014.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis With Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10684–10695.

[7] V. Heilala, R. Araya, and R. Hämäläinen, "Beyond Text-to-Text: An Overview of Multimodal and Generative Artificial Intelligence for Education Using Topic Modeling," in *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, New York, NY, USA, 2025, pp. 54–63, doi: 10.1145/3672608.3707764.

[8] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning With a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 1, art. no. 140, pp. 1–67, Jan. 2020.

[9] J. E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv preprint*, arXiv:2106.09685, 2021.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, vol. 139, 18–24 Jul. 2021, pp. 8748–8763.

[11] M. Belkina, S. Daniel, S. Nikolic, R. Haque, S. Lyden, P. Neal, S. Grundy, and G. M. Hassan, "Implementing Generative AI (GenAI) in Higher Education: A Systematic Review of Case Studies," *Computers and Education: Artificial Intelligence*, vol. 8, p. 100407, 2025, doi: 10.1016/j.caeai.2025.100407.

[12] W. K. Monib, A. Qazi, R. Apong, M. Azizan, L. Silva, and H. Yassin, "Generative AI and Future Education: A Review, Theoretical Validation, and Authors' Perspective on Challenges and Solutions," *PeerJ Computer Science*, vol. 10, p. e2105, Dec. 2024, doi: 10.7717/peerj-cs.2105.

[13] Z. Sordo, E. Chagnon, Z. Hu, J. J. Donatelli, P. Andeer, P. S. Nico, T. Northen, and D. Ushizima, "Synthetic Scientific Image Generation with VAE, GAN, and Diffusion Model Architectures," *Journal of Imaging*, vol. 11, no. 8, p. 252, 2025, doi: 10.3390/jimaging11080252.

[14] S. Vivekananthan, "Comparative Analysis of Generative Models: Enhancing Image Synthesis with VAEs, GANs, and Stable Diffusion," 2024, arXiv:2408.08751, doi: 10.48550/arXiv.2408.08751.

[15] N. J. Francis, S. Jones, and D. P. Smith, "Generative AI in Higher Education: Balancing Innovation and Integrity," *British Journal of Biomedical Science*, vol. 81, 2025, doi: 10.3389/bjbs.2024.14048.

[16] A. Alfarwan, "Generative AI Use in K-12 Education: A Systematic Review," *Frontiers in Education*, vol. 10, Sep. 2025, doi: 10.3389/feduc.2025.1647573.

[17] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, Ø. Tafjord, P. Clark, and A. Kalyan, "Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering," in *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, 2022, pp. 1–15.