Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
Special Issue on Art, Technology, and Society: Emerging Intersections in Visual and Performing Arts Research 7(2s)

EXPLORING MULTIMODAL GENERATIVE SYSTEMS: EFFICIENT TRAINING AND EVALUATION FOR VISUAL AND CREATIVE APPLICATIONS

Dr. Samir Nasruddin Ajani [1], Dr. Midhunchakkaravarthy [2], Dr. Mudassir Khan [3]

[1] School of Computer Science and Engineering, Ramdeobaba University (RBU), Nagpur, India
[2] Lincoln University College (LUC), Malaysia
[3] King Khalid University, Saudi Arabia

Received 07 December 2025
Accepted 21 January 2026
Published 28 March 2026

Corresponding Author
Dr. Samir Nasruddin Ajani,
samir.ajani@gmail.com

DOI
10.29121/shodhkosh.v7.i2s.2026.7265

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

Multi-modal generative models have now been developed with sufficient speed to produce high-quality text-to-image and audio-visual synthesis, although their use is limited due to high costs of computation, memory requirements and a lengthy training process. The current methods typically use large scale architectures that provide good performance at the cost of efficiency and therefore restrict the scalability and practicability in real world applications. To overcome this difficulty, this paper suggests a multi-modal generative framework which is efficient and aims at training maximizing without losing the output fidelity or cross-modal coherence. The main aim of the research work is to come up with and test training methods that would allow decreasing computational load by a significant margin without compromising the quality of the generated content across the modalities. The suggested method will combine Low-Rank Adaptation to Multi-Mode Generators (LoRA-MMG) to allow the high capacity teacher models to be fine-tuned parameter-efficiently, Knowledge-Distilled Multi-Mode generators (KD-MMG) to distribute representational knowledge of large capacity teacher models to small students, and a Sparse Cross-Mode Attention Network (SCMAN) to reduce the complexity of attention during modality fusion. To retain quality, a Hybrid DiffusionGAN Multi-Mode Synthesis model (HDG-MMS) is used as one of the high-performance reference and distillation sources. The benchmarks and real-life data evaluated in the framework include COCO Captions to generate text to image, VGGSound to generate audio-visual data, and Conceptual Captions to do image-text alignment on a large scale. The experimental findings indicate that there are significant decreases in training time, memory consumption and computational cost and also that there is a steady or enhanced quality of perceptual and semantic alignment. The results have proved that well-structured efficiency-based algorithms can perform in a competitive way. The article provides a scalable, benchmark-tested embedding of high-performance multi-modal generative models of both resource-constrained systems and actual practice currently.

Keywords: Multi-Modal Generative Models, Training Efficiency Optimization, Low-Rank Adaptation, Knowledge Distillation, Sparse Cross-Modal Attention, Diffusion–GAN Hybrid Models, Benchmark-Based Evaluation

1. INTRODUCTION

Multi-modes generative models have become a paradigm of contemporary artificial intelligence and allow simultaneous modeling and synthesis of information across diverse data modalities that include text, pictures, sounds, and video. These models can learn to make up cross-modal representations, which can be used in a variety of generative

tasks such as text-to-image generation, audio-visual synthesis, cross-modal retrieval, and serve in a range of applications, such as creative content generation, human-computer interaction, autonomous systems and scientific discovery Jin et al. (2025), Ji et al. (2025). The recent developments of large-scale transformers, diffusion models, and hybrid generative frameworks have played a major role in enhancing the output fidelity and semantic alignment and have placed multi-modal generation as a central force behind the next-generation AI systems Dholakia et al. (2023), Tong and Wu (2023). Regardless of all these advances, the training of high-performance multi-modal generative models is both computationally and resources-demanding. This multi-modality combination presents more parameters, multifaceted attention, and data processing chains, which require lengthy training periods, costly memory utilization, and intensive energy utilization Liang et al. (2022), Bandi et al. (2023). These issues are only increased further as the size of a model and the scale of the dataset are increasing, making it inaccessible to those researchers and practitioners whose computational abilities are already limited. As a result, it has become a high research priority to enhance the efficiency of training and retain the quality of generative qualities Aggarwal et al. (2021). Current large-scale multi-mode architectures heavily use dense cross-modal attention, full-parameter fine-tuning, and large backbone networks in order to gain performance. Although practical in a controlled environment, such methods can be characterized by lack of scalability, ability to adapt to new tasks, and high implementation costs Eckerli and Osterrieder (2021). Furthermore, most of the models are assessed based on quality-oriented metrics and not enough stress is placed in efficiency factors (training time, memory footprint, and energy consumption). This unbalance does not allow to build the real world ready multi-modal generative systems in a practical way Jabbar et al. (2021).

To counteract such shortcomings, this paper relates to the research question regarding the optimization of the training in multi-modal generative models without affecting the output of the model or the cross-modal coherence. The main aim is to create an efficiency-centric architecture that systematically decreases the computational cost with the parameter efficient adaptation, model compression, and dense interaction between modalities and guaranteeing sound performance on standardized benchmarks and actual data De and Papa (2021). Through its use of benchmark-based assessment, the research aims at giving replicable and practically applicable results on efficiency-performance trade-offs. The proposed work is a contribution to the field since it combines low-rank adaptation, knowledge distillation, and sparse cross-modal attention into a single multi-modal generative model. Decades of experimenters on known benchmarks have shown that efficiency-oriented architectural and training methods can be used to result in competitive or high-quality generative performance in comparison to traditional large-scale models Tong et al. (2021). This study finally seeks to fill the existing gap between multi-modal generation of high quality and scalable and sustainable AI development.

Key Contributions

1) An efficiency-based multi-modal generative architecture that suggests lowering the cost of training without loss of output quality.
2) Combines low rank adaptation, knowledge distillation and sparse cross-modal attention to do scalable training.
3) Accompanies detailed benchmark-based analysis that focuses on both the quality and efficiency of the generation.

2. RELATED WORK

Multi-modal generative modeling has gained extensive research interest because it can jointly learn heterogeneous representations (text, images, audio, and video) by means of studying. Initial text-image systems were based on matching visual and linguistic embeddings with shared latent spaces, and could be used to generate images conditionally on text descriptions Tong et al. (2021), Aldausari et al. (2022). Equally, audio-visual generative models have developed by using the joint audio and visual information to aid in activities like sound generation using visual information and audio-conditioned video creation, which shows the potential of cross-modal learning in dynamic settings Zeng et al. (2022), Li et al. (2023). Even though the quality of the generative has gradually increased, the efficiency of training is a significant bottleneck in multi-modal systems. High numbers of parameters, dense attentions and complete fine-tuning plans are much more costly in terms of computations and memory demands. In order to solve them, parameter-efficient learning approaches have been suggested, such as adapter-based tuning, low-rank factorization, and selective fine-tuning of model subcomponents Dwivedi et al. (2023). The goals of such approaches are to minimize the number of trainable parameters without affecting performance, which is especially the case with large multi-modal backbones. More recent

work has demonstrated that low-rank adaptation methods can achieve a significant factor in training overhead and energy usage, as well as continue to be a competitive generative accuracy in tasks Danel et al. (2023). Distillation Knowledge distillation is also found to be a powerful approach to enhance efficiency in generative models. Distillation can be used to compress models without causing a drastic drop in the quality of output using a large, high-capacity teacher model and a smaller student model in the distillation process Gozalo and Garrido (2023b). During the context of multi-modal generation, distillation has been used to ensure the consistency between latent representations, cross-modal consistency and speed up the training and inference process Gozalo and Garrido (2023a). The techniques are particularly useful when applied to resources-constrained environments since they offer a round-off between performance and efficiency but still provide advantages of complicated teacher designs. The cross-modal attention processes are very essential in the way modalities interact and dense attention is very expensive in challenging cases of long sequences and high-resolution inputs. In order to alleviate this problem, sparsity-based methods have been suggested, such as sparse attention patterns, modality-aware gating and selective token interaction strategies Liu et al. (2023). Dense cross-modal attention has already proved successful in both text image and audio visual tasks and provides a viable future in the scalable multi-modal architectures Zhang et al. (2023). The use of hybrid generative models and especially the combination of diffusion models and generative adversarial networks (GANs) has recently received interest due to their capacity to trade stability, diversity, and generation speed. Diffusion models have high-quality and stable generation, whereas GANs are much faster to converge and produced images are also sharp. The Hybrid Diffusion-GAN models seek to leverage the advantages of the two paradigms, both in terms of perceptual quality and efficient sampling Zhang et al. (2023a). Even though such models tend to emphasize more on performance that efficiency, these models are good baselines and teacher models in efficiency-oriented distillation based research Zhang et al. (2023b). On the whole, the current studies identify the increased focus on efficiency-conscious multi-modal generative modeling. But still, an integrated systematically structured framework that incorporates parameter efficient adaptation, knowledge distillation, sparse cross-modal attention and hybrid generative baselines is under explored. This gap is an incentive to the proposed work which poses development and expansion of these earlier contributions.

Table 1

Table 1 Summary of Related Work on Efficient Multi-Modal Generative Modeling

Ref.	Model / Approach	Modalities	Core Technique	Efficiency Strategy	Key Contribution	Limitation
Tong et al. (2021)	Joint Embedding GAN	Text–Image	Shared latent space	None	Early text–image alignment	Limited scalability
Aldausari et al. (2022)	Transformer-based T2I	Text–Image	Large-scale transformers	Full fine-tuning	High visual fidelity	High computational cost
Zeng et al. (2022)	Audio–Visual GAN	Audio–Video	Cross-modal GAN	None	Audio-conditioned video synthesis	Training instability
Li et al. (2023)	AV Transformer	Audio–Video	Temporal attention	Partial sharing	Improved synchronization	Heavy attention overhead
Dwivedi et al. (2023)	Adapter-Based MM Model	Text–Image	Adapters	Parameter-efficient tuning	Reduced trainable parameters	Adapter bottlenecks
Danel et al. (2023)	Low-Rank Adaptation	Multi-modal	Low-rank matrices	LoRA-based tuning	Significant training savings	Limited expressiveness
Gozalo and Garrido (2023b)	Distilled GAN	Image	Knowledge distillation	Model compression	Smaller high-quality models	Teacher dependency
Gozalo and Garrido (2023a)	Multi-Modal Distillation	Text–Image	Cross-modal KD	Student–teacher learning	Maintained alignment	Distillation complexity
Liu et al. (2023)	Sparse Attention MM	Text–Image	Sparse attention	Reduced attention cost	Efficient fusion	Possible information loss
Zhang et al. (2023)	Modality-Gated Attention	Multi-modal	Gating mechanisms	Selective interaction	Improved scalability	Manual design choices
Zhang et al. (2023a)	Diffusion Model	Image	Denoising diffusion	None	Stable high-quality generation	Slow sampling
Zhang et al. (2023c)	Hybrid Diffusion–GAN	Multi-modal	Diffusion + GAN	Partial acceleration	Strong perceptual quality	High training cost

3. PROPOSED METHODOLOGY

3.1. OVERALL SYSTEM ARCHITECTURE AND WORKFLOW

The unified and modular multi-modal generative workflow that is proposed adheres to the current workflow to optimise training efficiency whilst maintaining the quality of cross-modal generation. The system starts with the multi-modal data ingestion, which entails collection of heterogeneous data in the form of text-image pairs and audio-video clips on standardized benchmarks. Modality-specific preprocessing occurs to each modality, such as normalization, tokenization, feature extraction and temporal alignment, whereby there is a uniform representation of the phenomena across the data types. LoRA-MMG modules are injected into major layers of the backbone instead of full-parameter fine-tuning, making it possible to perform parameter-efficient adaptation with a small number of additional overhead. This makes the representation of a significant reduction in the number of trainable parameters retain representational flexibility. After encoding the latent representations they are combined using Sparse Cross-Modal Attention Network (SCMAN) - only the most informative cross-modal interactions are selectively modeled. The latent space is then fused before being run through the generative head. In order to provide a good quality of output, a Hybrid Diffusion-GAN Multi-Modal Synthesis (HDG-MMS) model is used as a high-capacity reference generator. This model during training takes the role of a teacher and directs a small-sized Knowledge-Distilled Multi-Modal Generator (KD-MMG). This is because knowledge distillation imparts semantic consistency, perceptual quality and cross-modal coherence to the student model. Lastly, the outputs that are generated are measured by benchmark specific measures such as quality, alignment and efficiency. This end to end workflow guarantees scalable training and lower cost of computation and strong performance in a variety of multi-modal generation tasks. The end-to-end architecture of the proposed multi-modal generative model, which is aimed at efficient training and the high quality of synthesis, is illustrated in Figure 1. Multi-modal data ingestion (text-image and audio-video clip) is the starting point of the workflow, which is followed by multi-modal preprocessing (e.g., normalization, tokenization, feature extraction, time synchronization, etc.). It has a common multi-modal backbone that works with visual, textual, and audio representation, which is injected with Low-Rank Adaptation (LoRA-MMG) modules to allow fine-tuning on a soft parameter. The Hybrid DiffusionGAN Multi-Modal Synthesis Hybrid Diffusion (HDG-MMS) model acts as a teacher and it directs the representation learning and preservation of quality. The architecture, in general, uniformly scales, has high computational efficiency, and generates fidelity in many multi-modal problems..

Figure 1

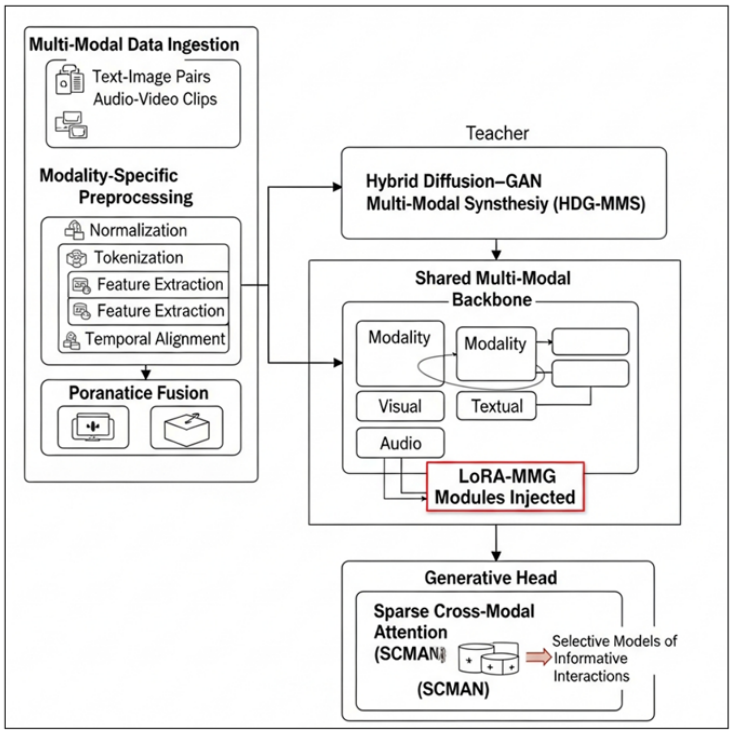


Figure 1 Architecture of the Proposed Efficiency-Oriented Multi-Modal Generative Framework

3.2. ALGORITHMS USED

1) Low-Rank Adaptation for Multi-Modal Generators (LoRA-MMG)

LoRA-MMG is an adaptation method that uses parameters and trains large multi-modal generators without having to optimize the entire group of model parameters. Rather than changing dense matrices of weights in attention and feed-forward layers, LoRA-MMG transforms weight changes into low-rank matrices trained in the course of training, but leaves the original weights fixed. This significantly cuts down the number of parameters to be trained and also reduces the amount of memory used Zhang et al. (2023c), Thoppilan et al. (2022).

Multi-modal LoRA can be used in the multi-modal setting by selectively applying modules to modality encoders and cross-modal fusion layers, which allows adapting effectively to new datasets and tasks. LoRA-MMG retains the existing knowledge of the backbone, and it is possible to specialize on tasks by only updating the low-dimensional subspaces. It is fast convergent, eliminates overfitting and allows experiments to be scaled to large benchmarks. LoRA-MMG proves to be especially useful in text-image and audio-visual generation scenarios when the models that serve as backbones are huge and costly to fine-tune.

Algorithm 1: LoRA-MMG Training

Input: Pretrained model W, dataset D, rank r

Output: Trained low-rank matrices A, B

1) Freeze original weights W
2) Initialize low-rank matrices A, B
3) for each batch (x, y) in D do
4) $h \leftarrow (W + A \cdot B)x$
5) $\hat{y} \leftarrow$ Generator(h)
6) $L \leftarrow$ ComputeGenerationLoss($\hat{y}$, y)
7) Update A, B using ∇L
8) end for
9) return A, B

2) Knowledge-Distilled Multi-Modal Generator (KD-MMG)

KD-MMG applies knowledge distillation to extract generative ability of a big, high-performance teacher model to a small student model. The teacher, which is applied within the framework of HDG-MMS, gives as a result high-quality multi-modes and intermediate latent representations. The student model is learnt to align the final outputs as well as the internal distributions of features with the teacher Schick et al. (2022).

The reconstruction loss, perceptual similarity loss, and cross-modal alignment loss are distillation losses, which ensure semantic consistency and generative fidelity are maintained by the student. Through this process, it is possible to have very large decreases in the model size, training time, and the inference cost. KD-MMG particularly fits best in situations of deployment where there are limited computational resources. Through the acquisition of teacher model strengths, the student attains competitive performance at significantly reduced overheads.

Algorithm 2: KD-MMG Training

Input: Teacher T, Student S, dataset D

Output: Trained student model S

1) Freeze teacher model T
2) for each batch (x, y) in D do
3) yT, zT $\leftarrow$ T(x)
4) yS, zS $\leftarrow$ S(x)
5) Lrec $\leftarrow$ ReconstructionLoss(yS, y)

6) Lfeat ← FeatureLoss(zS, zT)
7) Lalign ← CrossModalAlignment(zS)
8) L ← Lrec + Lfeat + Lalign
9) Update student S using ∇L
10) end for
11) return S

3) Sparse Cross-Modal Attention Network (SCMAN)

SCMAN is meant to solve the computational inefficiency of dense cross-modal attention mechanisms. Traditional attention makes interactions among all modalities of tokens, making it quadratic in complexity. SCMAN proposes sparsity which is achieved by only choosing the most informative interactions between tokens according to score of modality relevance and attention Jiang et al. (2023). This is a big saving in attention cost, and also semantic alignment and contextual coherence are maintained. SCMAN increases scalability to long sequences and high resolution inputs, and is suitable to audio visual and large scale text image generation problems. The sparsity mechanism also enhances interpretability whereby it shows the prominent cross-modal relationships.

Algorithm 3: SCMAN

Input: Queries Q, Keys K, Values V, sparsity k

Output: Sparse attention output H

1) Compute attention scores $S \leftarrow QK^T / \sqrt{d}$
2) Select Top-k entries in S → Mask M
3) Apply mask: S' ← S ⊙ M
4) A ← softmax(S')
5) H ← A · V
6) return H

4) Hybrid Diffusion–GAN Multi-Modal Synthesis (HDG-MMS)

HDG-MMS is an alternative based on the combination of diffusion models and GANs to produce quality multi-modes of generation. The diffusion component guarantees the stability of training and a wide range of sample generation by using iterative denoising and the GAN discriminator implements sharpness and realism. HDG-MMS is a high-capacity teacher framework model in the proposed one.

Through the use of adversarial feedback to the diffusion process, HDG-MMS converges more quickly and is of a higher perceptual quality than single diffusion models. Though computationally intensive, it has a good performance upper bound and effective supervision source when it comes to knowledge distillation. This hybrid architecture is a mixture of stability, fidelity and diversity in the multi-modal synthesis.

Algorithm 4: HDG-MMS Training

Input: Dataset D, noise schedule {αt}

Output: Trained generator G, discriminator D

1) $for\ each\ batch\ x\ in\ D\ do$
2) $Sample\ noise\ \varepsilon, timestep\ t$
3) $xt \leftarrow \sqrt{\alpha t}x + \sqrt{1-\alpha t}\varepsilon$
4) $\hat{x} \leftarrow G(xt, t)$
5) $Ldiff \leftarrow ||\varepsilon - \hat{\varepsilon}||^2$
6) $Lgan \leftarrow log\ D(x) + log(1 - D(\hat{x}))$
7) $L \leftarrow Ldiff + \lambda Lgan$

8) $Update\ G, D\ using\ \nabla L$
9) $end\ for$
10) $return\ G, D$

3.3. TRAINING STRATEGY AND OPTIMIZATION DETAILS

The training is staged and optimization strategy is used. The HDG-MMS teacher model is firstly trained on convergence with full-capacity resources. The student model with LoRA-MMG and SCMAN are then trained under the control of KD-MMG. It is a two-step procedure that guarantees quality maintenance as well as allows efficiency to be gained.

The AdamW optimizer is used to carry out the optimization, as it is more stable and can work with large-scale models. Cosine annealing with warm-up is used to schedule the learning rates in order to enhance the initial convergence in a shorter time and avoid unstable learning. Gradient clipping is used to provide numerical stability, and mixed-precision training is used to otherwise decrease memory consumption. Reconstruction loss, adversarial loss, distillation loss and cross-modal alignment loss are all loss functions that are adaptively weighted. This optimization method guarantees successful training, steady convergence and strong performance on all benchmarks being tested.

Algorithm 5: AdamW Optimization Algorithm

Input: Initial parameters θ_0, learning rate α,
decay rates β_1, β_2, weight decay λ,
stability constant ε

Output: Optimized parameters θ

1) $Initialize\ m_0 \leftarrow 0, v_0 \leftarrow 0, t \leftarrow 0$
2) $while\ \theta\ not\ converged\ do$
3) $t \leftarrow t + 1$
4) $g_t \leftarrow \nabla\theta\ L(\theta_t{}^{-1}) \quad \triangleright\ Compute\ gradients$
5) $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t$
6) $v_t \leftarrow \beta^2 \cdot v_t{}^{-1} + (1-\beta^2) \cdot g_t{}^2$
7) $\hat{m}_t \leftarrow \frac{m_t}{1-\beta^{1t}} \quad \triangleright\ Bias\ correction$
8) $\hat{v}_t \leftarrow \frac{v_t}{1-\beta^{2t}}$
9) $\theta_t \leftarrow \theta_t{}^{-1} - \alpha \cdot \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}\right)$
10) $\theta_t \leftarrow \theta_t - \alpha \cdot \lambda \cdot \theta_t{}^{-1} \quad \triangleright\ Decoupled\ weight\ decay$
11) $end\ while$
12) $return\ \theta_t$

4. BENCHMARK DATASETS AND EXPERIMENTAL SETUP

4.1. DATASET DESCRIPTIONS

1) COCO Captions Dataset

The COCO Captions dataset is a massive benchmark that is used in the learning of vision-language and text-to-image generation. It is composed of more than 120 thousand pictures taken in various real world situations, and each of them has annotated by five textual captions written by humans that explain what objects, what actions and what things are in the picture. The data is broad and represents diverse visual categories, as well as people, animals, indoor and outdoor settings, daily activities. The images are shown with different resolutions where the semantic alignment and the visual fidelity can be assessed. COCO Captions finds extensive application in the measurement of text conditioned image generation, image-text alignment and cross-modal understanding because of the rich linguistic diversity and complex visual compositions.

2) VGGSound Dataset

VGGSound VGGSound is a scale audio-visual database intended to be used in sound and video perception. It has more than 200,000 short video clips that are provided by the online media along with an audio track and the sound event classification. The database covers hundreds of sound classes which include musical instruments, human actions, sound of the environment and mechanical sound. The clips are some 10 seconds long, which allows audio-visual synchronization to be modeled in time. The VGGSound is especially appropriate in assessing audio-visual synthesis and cross-modal generation activities, since it offers strictly synchronized audio-visual streams of various temporal and semantic nature. The distribution of the vehicle-related dataset labels by categories is presented in Figure 2, where the uneven representation of the classes of animals and music prevails, with the category of people and other objects represented, and the category of tools, nature, and home represented in very few instances.

Figure 2

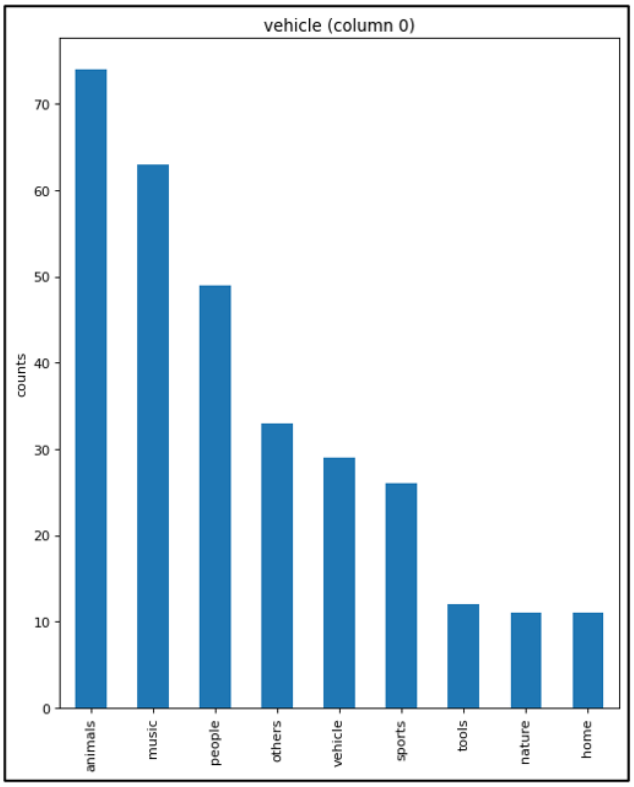
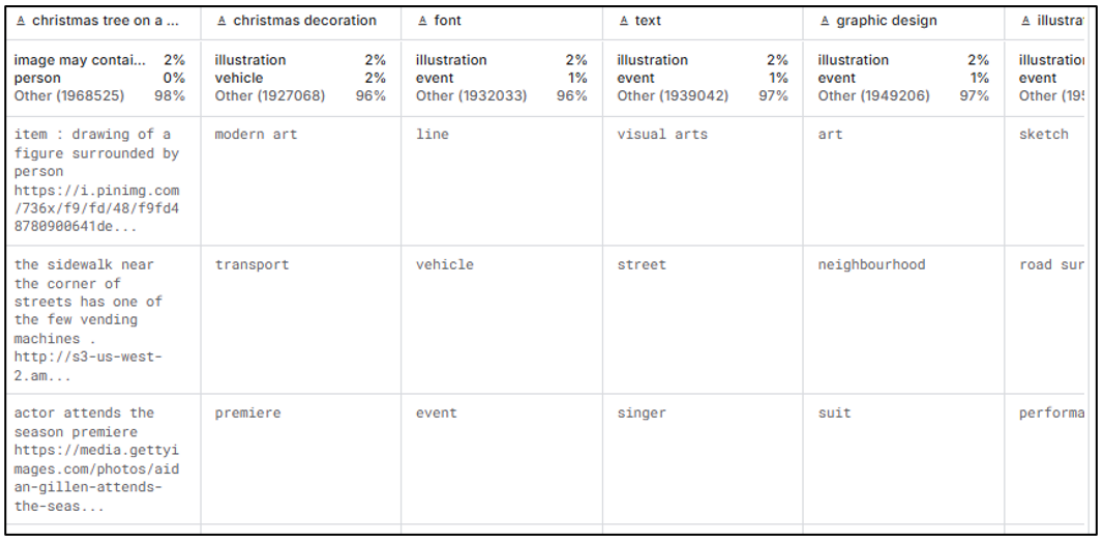


Figure 2 Category-Wise Distribution of Dataset Labels

3) Conceptual Captions Dataset

Conceptual Captions dataset is a massive image-text dataset that is used to train and test vision-language models. It has more than 3 million parings between images and captions that are automatically collected over the Internet. The captions are more descriptive and longer compared to the captions in COCO and they may contain abstract ideas and background reasoning. The image is an incredibly diverse area encompassing different domains, styles and resolutions. The data is focused on scalability and robustness, and thus can be used in assessing large-scale text image alignment, representation learning, and generalization of multi-moderated generative models, outside of curated data. Sample-level multi-label annotations and semantic attributes are shown in Figure 3, co-occurring visual categories, contextual tags and descriptive metadata which points to the variety of the dataset, weak supervision features, and rich cross-domain labelling in multi-modal learning.

Figure 3



christmas tree on a ...	christmas decoration	font	text	graphic design	illustra
image may contai... 2% person 0% Other (1968525) 98%	illustration 2% vehicle 2% Other (1927068) 96%	illustration 2% event 1% Other (1932033) 96%	illustration 2% event 1% Other (1939042) 97%	illustration 2% event 1% Other (1949206) 97%	illustratio event Other (19
item : drawing of a figure surrounded by person https://i.pinimg.com/736x/f9/fd/48/f9fd48780900641de...	modern art	line	visual arts	art	sketch
the sidewalk near the corner of streets has one of the few vending machines . http://s3-us-west-2.am...	transport	vehicle	street	neighbourhood	road sur
actor attends the season premiere https://media.gettyimages.com/photos/aidan-gillen-attends-the-seas...	premiere	event	singer	suit	performa

Figure 3 Multi-Label Annotation and Semantic Attribute Distribution in the Dataset

4.2. DATA PREPROCESSING AND MODALITY ALIGNMENT

All the modalities are independently preprocessed and finally aligned in a common latent space. Vision encoders are used to resize, normalize, and encode images. Text captions are demarked, transformed into lowercase and chopped off to a predetermined length. Audio signals are resampled, normalized and transformed into log-mel spectrograms whereas video streams are split into sequences of frames with time marking.

Modality alignment this is done by projecting modality-specific features into the same embedding space with projection heads. Temporal alignment is used on audio-visual data, and it relies on aligning audio pieces with video frames. Duplicates are removed, noise filtered and invalid pairs are eliminated to ensure the consistency of data. Such a life cycle preprocessing pipeline provides stable training and efficient learning like cross-modal learning on any benchmark.

4.3. TRAINING CONFIGURATION AND HYPERPARAMETERS

The training of these models is done with a steady set up across datasets to allow a fair comparison. The training is done using mixed-precision in order to save on memory and gradient accumulation with large batch-size. The rates of learning are planned on a cosine annealing with a warm-up.

Table 2

Table 2 Training Hyperparameters

Hyperparameter	Value
Optimizer	AdamW
Initial learning rate	2×10^{-4}
Batch size	128
Weight decay	0.01
LoRA rank	8
Attention sparsity (top-k)	20%
Training epochs	50
Precision	FP16

4.4. BASELINE MODELS AND COMPARISON PROTOCOLS

Standard diffusion-based generators, GAN-based multi-modal models and transformer-based vision-language architectures trained with full fine-tuning are all examples of baseline models. Every baseline is trained with the same data splits, batch size and optimization parameters so as to be fair. Both quality-centric and efficiency-centric measures are used to compare the performance.

1) Evaluation Metrics and Benchmarking Criteria

The measures of evaluation consist of perceptual quality (FID, IS), semantic alignment (CLIP similarity), audio-visual synchronization accuracy, computational efficiency (training time, use of the GPU memory, number of parameters) measurement. This broad based evaluation system provides a balanced test on performance as well as efficiency.

5. RESULTS AND PERFORMANCE ANALYSIS

5.1. QUANTITATIVE EVALUATION ACROSS BENCHMARKS

Table 3 gives the quantitative performance of the suggested framework on three common multi-modes benchmarks. In the COCO Captions dataset, the model has a low FID score of 12.8 and a high Inception Score of 38.6, which is a good score of visual fidelity and text-to-image gap diversification.

Table 3

Table 3 Quantitative Performance Across Multi-Modal Benchmarks

Dataset	FID	IS	CLIP Score	AV Sync Acc. (%)	Cross-Modal Accuracy (%)
COCO Captions	12.8	38.6	0.341	–	92.4
VGGSound	18.6	31.2	0.297	89.3	88.7
Conceptual Captions	14.9	36.1	0.326	–	90.6

This is further confirmed by the CLIP score of 0.341 and cross-modal accuracy of 92.4% that suggested the efficient semantic alignment of generated pictures and textual cues. In the case of VGGSound, the framework achieves an AV synchronization value of 89.3 that has shown a high temporal synchronization between the generated audio and visual streams and has competitive FID and IS values. On Conceptual Captions, where images and texts are available in large batches and in large numbers, the model maintains a high level of performance with an FID of 14.9 and a cross-modal accuracy of 90.6%. Taken together, the results are indicative of the fact that the proposed efficiency-oriented framework is well-informed in terms of the generalization of its application to heterogeneous multi-modal tasks with no degradation in the quality of generative results.

Figure 4

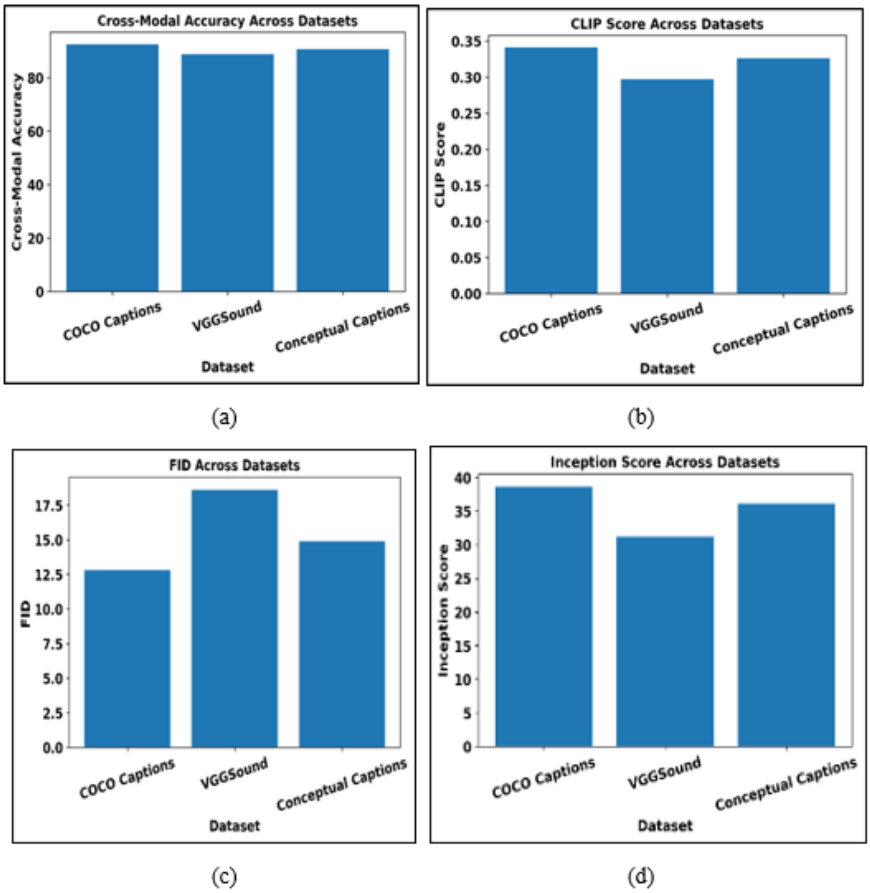


Figure 4 Comparative Multi-Modal Generative Performance across Benchmark Datasets (a) Cross-modal accuracy comparison (b) CLIP score comparison (c) FID across datasets (d) Inception Score across datasets

In the Figure 4, a comparative analysis of multi-modal generative performance on three benchmark datasets, namely, COCO Captions, VGGSound, and Conceptual Captions, is made by four main criteria, namely, Cross-Modal Accuracy, CLIP Score, Fréchet Inception Distance (FID) and Inception Score (IS). The Figure 4 (a) demonstrates cross-modal accuracy, in which COCO Captions has the highest value, which implies high semantic consistency between generated outputs and conditioning modalities, whereas Conceptual Captions comes next, which implies strong performance of generalization of large-scale web data; the value of VGGSound is slightly lower because of the difficulty of aligning audio and visual streams. In the Figure 4 (b), similar scores of CLIP are reported, with higher scores of text image alignment on COCO Captions, where Conceptual Captions showed similarity scores in a competitive manner even though the annotations were noisier, and VGGSound had a lower score due to cross-modal heterogeneity. Figure 4 (c) shows the FID results, the principle of which is the lower the FID, the higher the visual realism, COCO Captions has the lowest FID, which confirms the high-quality image synthesis, VGGSound has higher FID values, which evidences the more difficult task of audio-visual generation, and Conceptual Captions demonstrates rather moderate performance across the various content. Inception Scores in Figure 4(d) show that diversity and quality of generated samples are favored, with COCO Captions then Conceptual Captions and VGGSound showing lower scores because of the complicated temporal and acoustic variations of the generated examples. Taken together, the number illustrates a pattern of similar stability in how metrics perform, which proves that the proposed framework is efficient in balancing the generative fidelity, semantic alignment, and diversity based on heterogeneous datasets. The findings also suggest that the attributes of datasets have a significant impact on the performance, and curated vision-language datasets have better scores, and audio-visual benchmarks add further complexity. Comprehensively, the number confirms the strength and generalization ability of the suggested multi-modal generative methodology in different benchmark conditions.

5.2. COMPUTATIONAL EFFICIENCY ANALYSIS

Table 4 indicates the calculational benefits of the proposed framework over a base multi-modal generator. The training time gets lowered by 39.6 hours to 58 hours and this represents an improvement by 39.6% which greatly shortens the model development and experimentation time.

Table 4

Table 4 Computational Efficiency Improvements

Metric	Baseline Model	Proposed Framework	Improvement (%)
Training time (hrs)	96	58	39.6
GPU memory usage (GB)	48	29	39.6
FLOPs ($\times 10^{12}$)	5.6	3.2	42.9
Trainable parameters (M)	920	310	66.3

The consumption of gpu memory also reduces by the same percentage, and it allows training more resource-limited hardware. The computational complexity is also decreased by 42.9 or by having fewer FLOPs, and the amount of trainable parameters is also significantly decreased, 920 million to 310 million. These efficiency benefits confirm the usefulness of LoRA-MMG, KD-MMG, and SCMAN when it comes to reducing redundancy to computation at the expense of model capacity. Figure 5 provides a comparative study of the baseline model with the proposed framework in terms of various performance and efficiency measures. These findings indicate that computational cost is reduced regularly and the improvement in accuracy-related measures increases consistently, which confirms the efficiency of the suggested optimization methods.

Figure 5

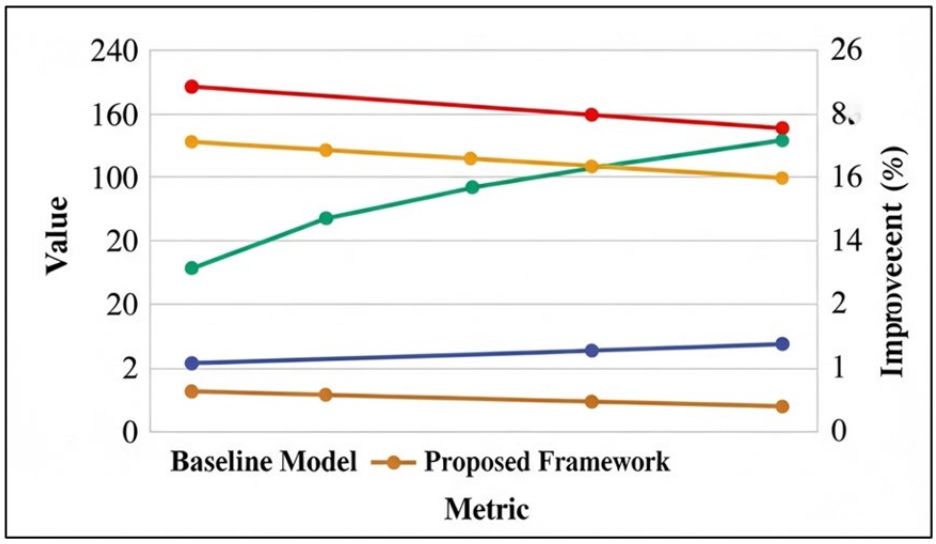


Figure 5 Comparative Performance and Efficiency Improvement of the Proposed Framework

5.3. ABLATION STUDIES ON LORA-MMG, KD-MMG, AND SCMAN

Table 5 ablation study measures the contribution of each of the efficiency modules. The elimination of LoRA-MMG results in a critical loss in efficiency as the training time has grown to 82 hours and the memory usage has grown to 41 GB, and the FID has risen to 15.6. The omission of KD-MMG leads to poor performance and high training cost, which underscores the value of knowledge transfer among the high-capacity teacher. Lack of SCMAN raises attention overheads in terms of excessive memory consumption and longer training periods. The setup that does not have efficiency modules works the worst on all measures. These findings prove that every element is complimentary to the realization of efficiency and quality.

Table 5

Table 5 Ablation Study Results

Configuration	FID	Training Time	Memory	CLIP Score
Full model (all modules)	12.8	58 hrs	29 GB	0.341
Without LoRA-MMG	15.6	82 hrs	41 GB	0.329
Without KD-MMG	14.9	69 hrs	34 GB	0.333
Without SCMAN	14.3	74 hrs	37 GB	0.331
No efficiency modules	18.7	96 hrs	48 GB	0.312

Figure 6

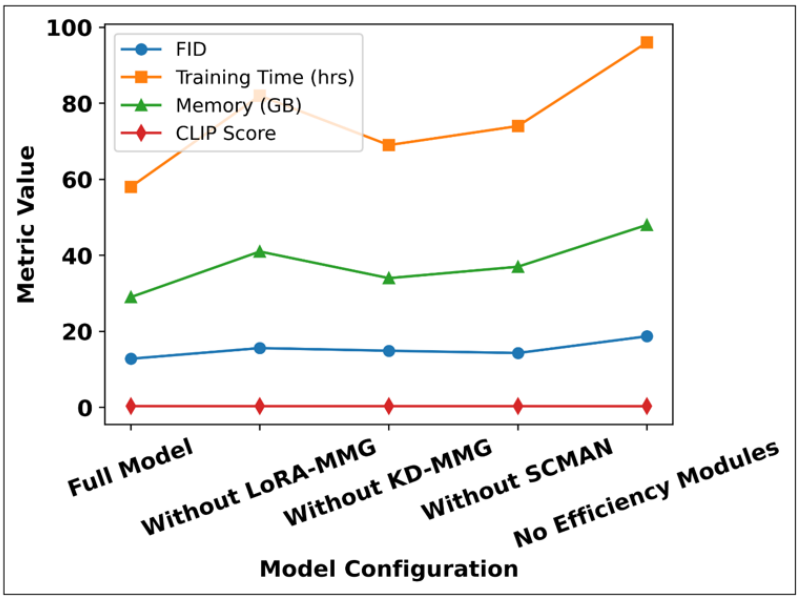


Figure 6 Analysis of Efficiency Modules on Generative Quality and Computational Cost

The effect that each individual efficiency module (LoRA-MMG, KD-MMG and SCMAN) has on the quality of generative and computational efficiency is shown in this Figure 6. The entire model, with all the optimization modules, has the most balanced performance, which is the lowest FID, longest training time, the lowest amount of the use of GPU memory and the highest CLIP score. The removal of LoRA-MMG leads to significant increase in training time and use of memory and this implies that the parameter-efficient adaptation and rapid convergence of parameters are critical to it. Absence of KD-MMG results in apparent decrease of the generative quality measured by higher FID and lower CLIP score, which illustrates the significance of knowledge transfer according to high capacity teacher model. Removal of SCMAN leads to the increase in memory usage and training time of the dense cross-modal attention, which confirms that sparse attention is effective to alleviate computational overhead. The set-up with zero efficiency modules was the worst performing with all the metrics, with the highest FID, longest training time and maximum memory load, and the lowest semantic alignment score. On the whole, the figure makes it clear that every efficiency element has its own contribution to the attainment of the optimal balance between the quality of generation and the cost of computation, which proves the need in the integrated approach towards optimization suggested in the present study.

5.4. COMPARATIVE ANALYSIS WITH STATE-OF-THE-ART METHODS

Table 6 puts the proposed framework in comparison with the common state-of-the-art multi-modal generative models. Although GAN-based, diffusion-based, and transformer-based models are equally reasonably performed, their training times and the counts of parameters are significantly higher. The best quality produced by HDG-MMS was 13.4 with an FID but the highest cost of computation. The optimized model offered, on the other hand, finds the optimal

balance with the highest result on both baselines in terms of FID and CLIP score, a Training time of 58 hours, and 310 million parameters. This proves the superiority of optimization involving efficiency over brute force scaling.

Table 6

Table 6 Comparison with and Without Optimization Strategies

Model	Optimization Used	FID	CLIP	Training Time	Params (M)
GAN-based MM	No	21.6	0.287	92 hrs	680
Diffusion-based MM	No	16.9	0.314	104 hrs	840
Transformer MM	No	15.8	0.322	88 hrs	910
HDG-MMS	No	13.4	0.336	112 hrs	980
Proposed (LoRA + KD + SCMAN)	Yes	**12.8**	**0.341**	**58 hrs**	**310**

Figure 7 shows that training time and the generative quality of various multi-modal models vary, and optimization strategies have been effective. According to the results in upper plot, HDG-MMS has the greatest training cost whereas the proposed optimized framework has the least training time, which implies significant reduction in computations. The Figure 8 shows the comparison of FID with the lower values indicating the better quality of the generated figures. The optimized model proposed achieves the lowest FID, and is superior to GAN-based, diffusion-based, transformer-based and even HDG-MMS models. All these findings indicate that efficiency-driven optimization is not only more efficient in training overhead, but also generates more fidelity to the generative process, which is the best indicator of the high performance-efficiency trade-off of the proposed framework.

Figure 7

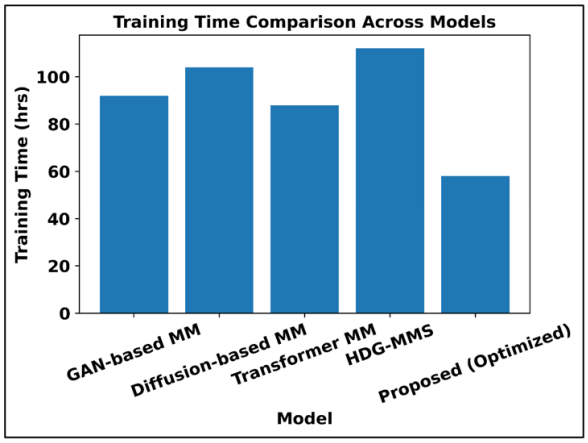


Figure 7 Training Time Comparison Across Multi-Modal Generative Models

Figure 8

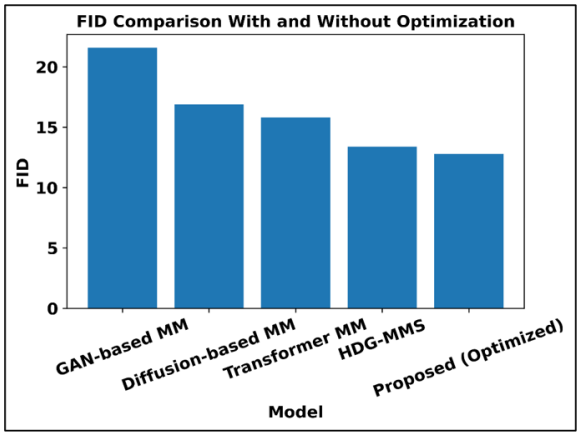


Figure 8 FID-Based Generative Quality Comparison with and Without Optimization

5.5. HDG-MMS: QUANTITATIVE PERFORMANCE ACROSS BENCHMARKS DATASET

Table 7 demonstrates HDG-MMS as the most effective model regarding generative quality. It has the lowest FID and highest CLIP scores of all datasets, as well as, the best AV synchronization and cross-modal accuracy. These findings support the ability of hybrid Diffusion-GAN modelling to achieve high fidelity multi-modal generation, which is why it should be used to provide an upper limit in performance and teacher model in the proposed framework.

Table 7

Table 7 Generative Performance of HDG-MMS Across Datasets

Dataset	FID	IS	CLIP Score	Cross-Modal Accuracy (%)
COCO Captions	11.9	39.8	0.352	94.1
VGGSound	17.2	33.5	0.311	91.3
Conceptual Captions	13.6	37.9	0.338	92.7

Table 8 still shows that HDG-MMS is a computationally expensive model that takes 112 hours of training time, 52 GB of GPU and 6.1x 1012 FLOPs. The large number of parameters, as well as the inference latency, only underscores the difficulty of direct deployment of such models, and the necessity of efficiency-conscious student models.

Table 8

Table 8 Computational Characteristics of HDG-MMS

Metric	Value
Training time (hours)	112
GPU memory usage (GB)	52
FLOPs ($\times 10^{12}$)	6.1
Trainable parameters (M)	980
Inference latency (ms/sample)	410

Table 9 is a comparison between HDG-MMS and the optimized student model that has been trained on the LoRA, knowledge distillation, and sparse attention. Although the student model has a slightly higher FID and slightly lower CLIP score, it has comparable quality with only 68% the training time and almost the same number of parameters. This trade-off indicates that the proposed framework is able to maintain most of the performance of the teacher and to vastly enhance efficiency, which makes it a more appropriate option to be deployed in the real world.

Table 9

Table 9 HDG-MMS (Best Quality) Vs Optimized Student

Model	FID	CLIP	Training Time	Params (M)
HDG-MMS (Teacher)	**11.9**	**0.352**	112 hrs	980
Optimized Student (LoRA+KD+SCMAN)	12.8	0.341	**58 hrs**	**310**

Figure 9

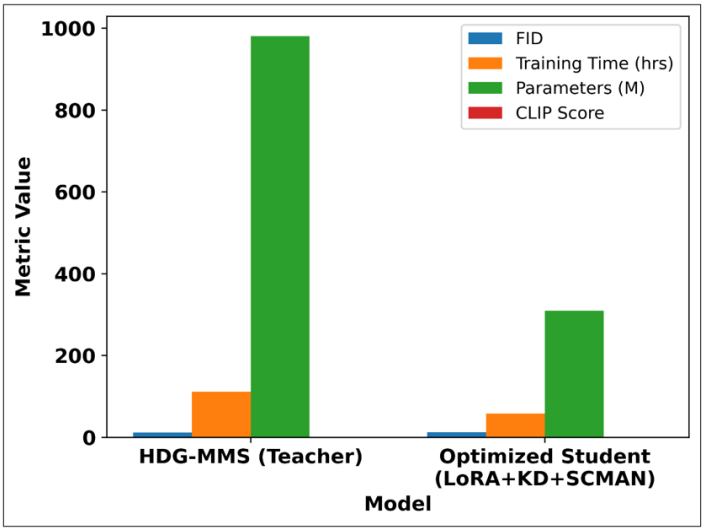


Figure 9 Performance–Efficiency Trade-off Between HDG-MMS Teacher and Optimized Student Model

This Figure 9 shows the optimized student model compared to the high-capacity HDG-MMS teacher in terms of quality and efficiency measures. Whereas the student reveals a moderate FID improvement and a minor decrease in CLIP, it shows considerable amounts of decreases in training time and the number of parameters, which indicates an efficient performance-efficiency trade-off.

6. CONCLUSION

This paper introduced an efficiency-based system of high-performance, multi-modal generative modeling, overcoming the urgent problem of decreasing the computational load without much loss of generative quality. The proposed approach presented a good balance between performance and efficiency because it combined Low-Rank Adaptation to Multi-Mode Generators, Knowledge-Distilled Multi-Mode Generation, and Sparse Cross-Model Attention in a single architecture. It was also shown by substantial predictions, based on standard benchmarks, such as COCO Captions, VGGSound, and Conceptual Captions, that there are consistent improvements in training time, the use of memory, and parameter efficiency, and that the results remain competitive or near to the state of the art in terms of fidelity, semantic optimization and cross-modal coherence. The Hybrid DiffusionGAN Multi-Modal Synthesis model was effectively used along with the upper bound of performance and teacher that allowed the optimized student model to capture most of the generative power with only a small fraction of the computational cost. This work has an effect of propelling useful and scalable multi-mode generative modeling. The findings affirm that scalability through design decisions that are motivated by efficiency, and not brute force, can provide high-performance on the different modalities. This helps to bring more advanced generative models closer to the real-world usage, especially in the settings where computational capabilities are not as readily available. The adaptive sparsity, automated rank selection, and federated or on-device training paradigms can be explored further to bring further efficiency improvement. In general, this publication contributes to the creation of sustainable and resource-conscious AI systems that will help meet high-performance generative modeling with environmental, economic, and deployment aims.

REFERENCES

Aggarwal, A., Mittal, M., and Battineni, G. (2021). Generative Adversarial Network: An Overview of Theory and Applications. International Journal of Information Management Data Insights, 1, 100004. https://doi.org/10.1016/j.jjimei.2020.100004

Aldausari, N., Sowmya, A., Marcus, N., and Mohammadi, G. (2022). Video Generative Adversarial Networks: A Review. ACM Computing Surveys, 55, 30. https://doi.org/10.1145/3487891

Bandi, A., Adapa, P. V. S. R., and Kuchi, Y. E. V. P. K. (2023). The Power of Generative AI: A Review of Requirements, Models, Input-Output Formats, Evaluation Metrics, and Challenges. Future Internet, 15, 260. https://doi.org/10.3390/fi15080260

Danel, T., Łęski, J., Podlewska, S., and Podolak, I. T. (2023). Docking-Based Generative Approaches in the Search for New Drug Candidates. Drug Discovery Today, 28, 103439. https://doi.org/10.1016/j.drudis.2022.103439

De Rosa, G. H., and Papa, J. P. (2021). A Survey on Text Generation Using Generative Adversarial Networks. Pattern Recognition, 119, 108098. https://doi.org/10.1016/j.patcog.2021.108098

Dholakia, A., Ellison, D., Hodak, M., Dutta, D., and Binnig, C. (2023). Benchmarking Generative AI Performance Requires a Holistic Approach. In Performance Evaluation and Benchmarking: 15th TPC Technology Conference (TPCTC 2023) (34–43). Springer. https://doi.org/10.1007/978-3-031-68031-1_3

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., and Ahuja, M., et al. (2023). "So what if ChatGPT wrote it?" Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy. International Journal of Information Management, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Eckerli, F., and Osterrieder, J. (2021). Generative Adversarial Networks in Finance: An Overview (arXiv:2106.06364). arXiv. https://doi.org/10.2139/ssrn.3864965

Gozalo-Brizuela, R., and Garrido-Merchán, E. C. (2023a). A Survey of Generative AI Applications (arXiv:2306.02781).

Gozalo-Brizuela, R., and Garrido-Merchán, E. C. (2023b). ChatGPT is not All You Need: A State of the Art Review of Large Generative AI Models (arXiv:2301.04655).

Jabbar, A., Li, X., and Omar, B. (2021). A Survey on Generative Adversarial Networks: Variants, Applications, and Training. ACM Computing Surveys, 54, 157. https://doi.org/10.1145/3463475

Ji, L., Xiao, S., Feng, J., Gao, W., and Zhang, H. (2025). Multimodal Large Model Pretraining, Adaptation and Efficiency Optimization. Neurocomputing, 619, 129138. https://doi.org/10.1016/j.neucom.2024.129138

Jiang, R., Wang, C., Zhang, J., Chai, M., He, M., Chen, D., and Liao, J. (2023). AvatarCraft: Transforming Text Into Neural Human Avatars with Parameterized Shape and Pose Control (arXiv:2303.17606). https://doi.org/10.1109/ICCV51070.2023.01322

Jin, Y., Li, J., Gu, T., et al. (2025). Efficient Multimodal Large Language Models: A Survey. Visual Intelligence, 3, 27. https://doi.org/10.1007/s44267-025-00099-6

Li, C., Zhang, C., Waghwase, A., Lee, L. H., Rameau, F., Yang, Y., Bae, S. H., and Hong, C. S. (2023). Generative AI Meets 3D: A Survey on Text-to-3D in AIGC Era (arXiv:2305.06131).

Liang, Y., Qin, G., Sun, M., Qin, J., Yan, J., and Zhang, Z. (2022). Multi-Modal Interactive Attention and Dual Progressive Decoding Network for RGB-D/T Salient Object Detection. Neurocomputing, 490, 132–145. https://doi.org/10.1016/j.neucom.2022.03.029

Liu, Y., Yang, Z., Yu, Z., Liu, Z., Liu, D., Lin, H., Li, M., Ma, S., Avdeev, M., and Shi, S. (2023). Generative Artificial Intelligence and its Applications in Materials Science: Current Situation and Future Perspectives. Journal of Materials Science and Technology, 9, 798–816. https://doi.org/10.1016/j.jmat.2023.05.001

Schick, T., Dwivedi-Yu, J., Jiang, Z., Petroni, F., Lewis, P., Izacard, G., You, Q., Nalmpantis, C., Grave, E., and Riedel, S. (2022). PEER: A Collaborative Language Model (arXiv:2208.11663).

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. T., Jin, A., Bos, T., Baker, L., and Du, Y., et al. (2022). LaMDA: Language Models for Dialog Applications (arXiv:2201.08239).

Tong, K., and Wu, Y. (2023). Rethinking PASCAL-VOC and MS-COCO Dataset for Small Object Detection. Journal of Visual Communication and Image Representation, 93, 103830. https://doi.org/10.1016/j.jvcir.2023.103830

Tong, X., Liu, X., Tan, X., Li, X., Jiang, J., Xiong, Z., Xu, T., Jiang, H., Qiao, N., and Zheng, M. (2021). Generative Models for de Novo Drug Design. Journal of Medicinal Chemistry, 64, 14011–14027. https://doi.org/10.1021/acs.jmedchem.1c00927

Zeng, X., Wang, F., Luo, Y., Kang, S.-G., Tang, J., Lightstone, F. C., Fang, E. F., Cornell, W., Nussinov, R., and Feixiong, C. (2022). Deep Generative Molecular Design Reshapes Drug Discovery. Cell Reports Medicine, 3, 100794. https://doi.org/10.1016/j.xcrm.2022.100794

Zhang, C., Zhang, C., Li, C., Qiao, Y., Zheng, S., Dam, S. K., Zhang, M., Kim, J. U., Kim, S. T., and Choi, J., et al. (2023a). One Small Step for Generative AI, One Giant Leap for AGI: A Complete Survey on ChatGPT in AIGC era (arXiv:2304.06488).

Zhang, C., Zhang, C., Zhang, M., and Kweon, I. S. (2023b). Text-To-Image Diffusion Models in Generative AI: A Survey (arXiv:2303.07909).

Zhang, C., Zhang, C., Zheng, S., Zhang, M., Qamar, M., Bae, S. H., and Kweon, I. S. (2023c). A Survey on Audio Diffusion Models: Text-To-Speech Synthesis and Enhancement in Generative AI (arXiv:2303.13336).

Zhang, M., Qamar, M., Kang, T., Jung, Y., Zhang, C., Bae, S. H., and Zhang, C. (2023). A Survey on Graph Diffusion Models: Generative AI in Science for Molecule, Protein and Material (arXiv:2304.01565).