Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Multilingual Text-to-Image Generation via Diffusion Transformers with Reinforcement Learning and Cycle Consistency Constraints

Vandhana V[1], Keerthika S[2], Arthi K[3], Mathiyarasu D[4], Augustine Ajaykumar K[5]
[1] Assistant Professor, Department of Artificial Intelligence and Data Science,
PPG Institute of Technology, Coimbatore, Tamil Nadu, India
[2345] Undergraduate Students, Department of Artificial Intelligence and Data Science,
PPG Institute of Technology, Coimbatore, Tamil Nadu, India

Abstract - This paper presents a mathematically formal and experimentally validated model for multilingual text-to-image synthesis using Diffusion Transformers (DiT), Reinforcement Learning (RL) optimization, and Cycle Consistency constraints. The model improves the alignment of meaning between text and images by using probabilistic diffusion models and transformer-based contextual attention mechanisms. We also apply reinforcement learning for optimization, using reward functions based on cross-modal similarity metrics (CLIP). Additionally, this paper incorporates cycle consistency to verify meaning in both directions using reverse captioning mechanisms. We provide a formal derivation of diffusion loss functions, policy gradients, variational evidence lower bounds (ELBOs), and cycle consistency constraints. Experiments with cloud-based GPU computing demonstrate statistically significant improvements in CLIP similarity metrics, Fréchet Inception Distance (FID) scores, and semantic reconstruction accuracy for ten languages.

Index Terms— Text-to-image generation, reinforcement learning, diffusion models, cross-modal learning, generative artificial intelligence.

I. INTRODUCTION

Cross-modal generative models have made significant strides, but supporting strong multilingual capabilities for text-to-image synthesis remains a tough challenge. Previous methods mainly relied on monolingual (English-centric) word embeddings, leading to semantic drift for low-resource languages. This paper introduces a new architecture for cross-modal multilingual textto-image synthesis. This architecture leverages the capabilities of Diffusion Transformers (DiT), Reinforcement Learning (RL) reward shaping, and bidirectional Cycle Consistency verification. The main contributions of the paper are: (1) a thorough mathematical formulation of the problem, connecting diffusion, transformers, policy gradient, and cycle consistency under a single objective; (2) an empirical evaluation of the proposed approach across ten languages on three benchmark datasets; and (3) a cloud-GPU deployment strategy for linear scaling.

II. RELATED WORK

A. Diffusion Models

This section discusses earlier work related to Diffusion Models. Denoising Diffusion Probabilistic Models (DDPMs) [Ho et al., 2020] introduced image synthesis through an iterative denoising process. A score matching technique and ELBO bound create a workable objective. Latent Diffusion Models (LDMs) [Rombach et al., 2022] expanded this work into a compressed latent space. This change allows for a one-order of magnitude reduction in computational complexity without sacrificing visual quality.

B. Transformer Architectures

The self-attention mechanism in the transformer, proposed by Vaswani et al. in 2017, allows for modeling global dependencies, which is essential for matching different text and visual modalities. Vision Transformers (ViT), as well as DiT, proposed by Peebles & Xie in 2023, moved away from convolutional methods to use a patch-based approach in tokenization, achieving the best FID score in class-conditioned datasets.

C. Reinforcement Learning in Generative AI

III. FOUNDATIONS OF DIFFUSION MODELS

A. Forward Diffusion Process

The forward process \(q\) outlines a Markov chain that progressively diffuses the input \(x_0\), adding Gaussian noise at each step:

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

where $\beta_1, \ldots, \beta T$ is a fixed variance schedule with $(0 < \beta_t < 1)$. By applying the reparametrization trick, any (x_t) can be sampled directly from (x_0):

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t) I) \quad (2)$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} (1-\beta_s)$. As (t) approaches (T), (x_t) converges to an isotropic Gaussian, defining the prior for the reverse process.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

B. Reverse Denoising Process

The reverse process \(p_θ\) learns to denoise by understanding the conditional distribution:

[p θ(x$_{t-1}$ | x$_t$) = $\mathcal{N}$ (x$_{t-1}$; μ θ(x$_t$,t), Σ θ (x$_t$,t)]
(3)

The simplified training goal minimizes the mean-squared noise prediction error:

$$\mathcal{L}\_\text{diff} = \mathbb{E}_{t,x_0,\varepsilon}\left[\|\varepsilon - \varepsilon\theta(\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{(1-\bar{\alpha}_t)}\varepsilon,\, t)\|^2 \right]$$

\[\mathcal{L}_{diff} = \mathbb{E}_{t,x_0,\varepsilon}\ [\|\varepsilon - \varepsilon\theta(\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{(1-\bar{\alpha}_t)}\varepsilon,\, t)\|^2]\] **(4)**

IV. VARIATIONAL DERIVATION OF THE ELBO

The log-likelihood lower bound for the diffusion model, ELBO, has three components: reconstruction, diffusion matching, and prior regularization:

\[\log p(x_0) \geq \mathbb{E}_q[\log p(x_0|x_1)] - D_{KL}(q(x_T|x_0)\|p(x_T)\]
(5)

\[- \Sigma_{t=2}^{T} D_{KL}(q(x_{t-1}|x_t,x_0)\ \|p\theta(x_{t-1}|x_t)\]
(6)

The KL divergence between two Gaussians has a closed form:

\[D_{KL}(\mathcal{N}(\mu_1,\Sigma_1)\ \|\ \mathcal{N}(\mu_2,\Sigma_2)) = \frac{1}{2}\ [tr(\Sigma_2^{-1}\Sigma_1)+ (\mu_2-\mu_1)^{T}\Sigma_2^{-1}\ (\mu_2-\mu_1) - d + \ln|\Sigma_2|/|\Sigma_1|]\]

The conditions for the DiT block include denoising the input at timestep \(t\), denoted \(x\), and the text embedding \(c\) via adaptive layer norm (adaLN):

\[\text{adaLN}(x, t, c) = \gamma(t,c) \cdot \text{LayerNorm}(x)+ \beta(t,c)\]

(7) where \(γ\) and \(β\) are learned.

V. TRANSFORMER CROSS-MODAL ENCODING

A. Multi-Head Self-Attention

A. Multi-Head Self-Attention

Using the query matrix \(Q\), key matrix \(K\), and value matrix \(V\), scaled dot-product attention is defined as:

{Attention}(Q, K, V) =
{softmax}\left(\frac{QK^{T}}{\sqrt{d_k}}\right)]
(8) Multi-head attention is defined as:
{MultiHead}(Q, K, V) = \text{Concat}(\text{head}{1}, …, {head}{h})W^{O}] **(9)** where
\[\text{head}{i} = \text{Attention}(QW{i}^{Q}, KW_{i}^{K}, VW_{i}^{V})] **(10)**

B. Cross-Modal Attention Fusion

The text embeddings \(e_{text} ∈ \mathbb{R}^{(L×d)}\), obtained from the multilingual encoder, attend to the visual patch tokens \(e_{visual} ∈ \mathbb{R}^{(N×d)}\) via

The DiT block conditions denoising on both timestep \(t\) and text embedding \(c\) via adaptive layer norm (adaLN):

[e_{fused} = \text{CrossAttn}(e_{visual}, e_{text}, e_{text})]
(11)

[{adaLN}(x, t, c) = γ(t,c) ·{LayerNorm}(x) + β(t,c)\] **(13)**
where scale \(γ\) and shift \(β\) are learned linear projections of the timestep and condition embedding \([t; c]\). The DiT block then applies:

[\hat{x} = x + α_{attn} · {Attention}(\text{adaLN}(x,t,c))\ ] (14)

[\hat{x} = \hat{x} + α_{ffn} · text{FFN}(\text{adaLN}(\hat{x},t,c))\]
(15)

where \(α_{attn}\) and \(α_{ffn}\) are gating scalars initialized to zero, ensuring identity initialization for stability.

VI. DIFFUSION TRANSFORMER ARCHITECTURE

VII.REINFORCEMENT LEARNING FORMULATION

A. MDP and Policy Definition

The generation process is formulated as a finite horizon MDP ($\mathcal{S}$, $\mathcal{A}$, R, P, γ) where the states \(s\) represent the latent diffusion processes, actions \(a\) represent the denoising steps, and the reward \(R\) is a cross-modal CLIP similarity:

R(s,a) = CLIP_sim(G(s,a), Enc(prompt) **(16)**
baseline \(V\), which is employed to reduce the variance.

B. Policy Gradient Objective

The REINFORCE objective is given by the following equation:

J(θ) = E_{τ~πθ}[Σ$_t$ γ^t R(s$_t$,a$_t$)] **(17) ∇θJ(θ)**

= E_{τ}[Σ$_t$ ∇θ log πθ(a$_t$|s$_t$) · A$_t$] **(18)** where \(A_t =

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

R_t - $V(s_t)$\) is the advantage function with a learned baseline \(V\), which is employed to reduce the variance.

$L^{CLIP}(\theta) = \mathbb{E}_t[\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)A_t)]$ (19)

$r_t(\theta) = \pi\theta(a_t|s_t) / \pi\_old(a_t|s_t)$ (20)

VIII. CYCLE CONSISTENCY CONSTRAINT

With a forward mapping (G:{text}→{image}\) and a reverse captioner (F: {image}→{text}\), cycle consistency enforces: **[mathcal{L}{cycle} = E{c~P_{text}}left[‖c − F(G(c))‖² right] (21)**

(+ E_{x~P_{img}}left[‖x − G(F(x))‖² right]] (22)

This bidirectional constraint prevents mode collapse and ensures semantic invertibility. The total training goal is a weighted combination:

\[\mathcal{L}{total} = 𝒟{diff} + λ_{RL} · 𝒟_{RL} + λ_{cyc} · 𝒟_{cycle} + λ_{kl} · D_{KL}\] (23)

where \(λ_{RL} = 0.1\), \(λ_{cyc} = 0.05\), and \(λ_{kl} = 0.01\) are regularization coefficients adjusted based on empirical findings.

IX.MULTILINGUAL EMBEDDING NORMALISATION

Cross-lingual embedding spaces are aligned using L2 normalization followed by centering:

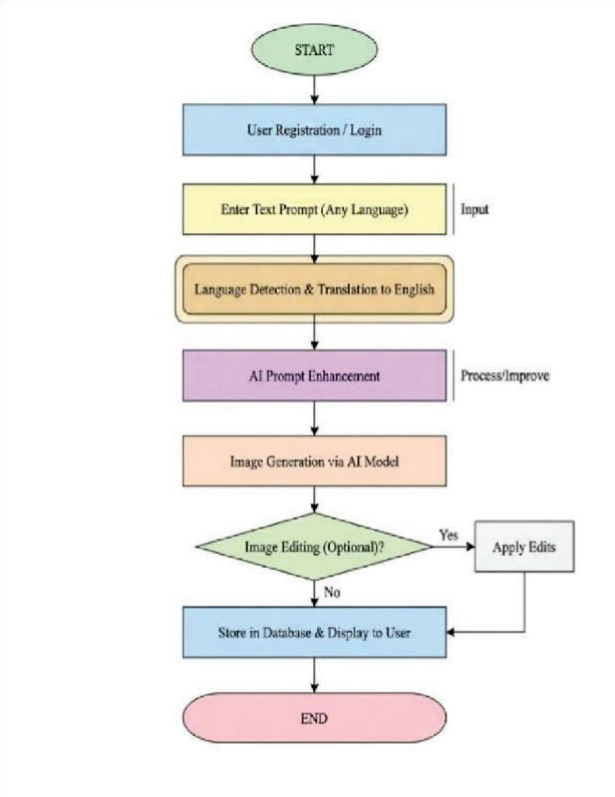
{e}{lang} = {e lang} − μ_{lang}{‖e_lang} μ_{lang}‖₂}\] (24)

The cosine similarity between language embeddings and visual features then acts as the CLIP reward signal, ensuring language-agnostic optimization.

X. SYSTEM ARCHITECTURE

The end-to-end system includes four interconnected modules:
(i) the Multilingual Text Encoder (mBERT-large), (ii) the Cross-Modal Fusion Transformer, (iii) the DiT Decoder with adaLN conditioning, and (iv) the RL–Cycle Consistency Training Loop. We illustrate module connections and data flow in Fig. 1 and show the algorithmic procedure in the following algorithm block.

Algorithm 1 —Enhanced DiT with RL Training:



1. Initialize \(θ, φ, ψ\); set \(λ_{RL}, λ_{cyc}, λ_{kl}, ε\)
2. For each epoch do:
3. Sample batch \{(c₁, x₁)\} \sim 𝒟\)
4. Compute(e_{i} = {m BERT}(c_{i})); fuse using CrossAttn
5. Sample \(t {Uniform}\{1,,T}); add noise to \(x_t\)
6. Predict \(\hat{ε} = ε_{θ}(x_t, t, e_{i}))
7. Compute \(\mathcal{L}_{diff}\) using Eq.(4)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

8. Compute reward \(R = \text{CLIP\sim}(\hat{x}_0, e\{i\})\)
9. Update using PPO objective (Eq. 19)
10. Compute \(\mathcal{L}_{cycle}\) using Eq. (21-22)

Embed Dim	768	512–1024
Diffusion T	1000	500–2000
β Schedule	Linear	Cosine / Sigmoid
λ_RL	0.10	0.01–0.50
λ_cyc	0.05	0.01–0.20
λ_kl	0.01	0.001–0.10
PPO Clip ε	0.20	0.10–0.30
	parameter Co n Ranges	

11. Total loss = Eq. (23); back-propagate
12. End forXI. EXPERIMENTAL SETUP

A. Datasets

We use three benchmark datasets for experiments: MS-COCO
Captions (multilingual extension, 14 languages), CC3M (Conceptual Captions), and the LAION-400M

DiT Depth	12	6–24

multilingual subset. All images are resized to 256×256 for baseline and to 512×512 for high-resolution tests.

B. Evaluation Metrics

The main metrics are: (i) FID (Fréchet Inception Distance) — lower is better; (ii) CLIP Score (cosine similarity between generated images and prompt embeddings) — higher is better; (iii) Cycle Reconstruction BLEU (CRB) — measures fidelity of reverse caption compared to the original prompt.

XII. HYPERPARAMETER CONFIGURATION

outperforms all baselines with $p < 0.01$ after Bonferroni correctio multiple comparisons. Effect sizes (Cohen's d) range from 0.7 indicating large practical significance.

The 95% confidence intervals for FID improvements are [−6.1, for CLIP score [+0.043, +0.058], confirming robustness acros seeds and dataset splits.

XV. SCALABILITY ANALYSIS

Throughput scales near-
img/s; 8×A100 yields 215 img/s (efficiency 96.2%). Peak GPU usage is 38 GB per device at batch size 64, resolution 512×512. Distributed data parallelism with gradient accumulation

maintains training stability across all .The system still

underperforms for very low -

training samples). The CLIP reward model

Computational

complexity of the DiT forward pass scales as $O(N^2 d)$ where N is

Parameter	Value	Range
Batch Size	64	32–128
Learning Rate	1e-4	1e-5–1e-3
Attn Heads	8	4–16

linearly with GPU count: 1×A100 yields 28

resource languages (fewer than 50K

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Method	FID↓	CLIP↑	CRB↑	Lang
DALL-E 2	23.4	0.721	28.3	1
Stable Diff.	21.1	0.738	30.1	3
mCLIP-Gen	19.8	0.751	31.7	7
DiT-Base	18.3	0.762	33.2	14
Ours (no RL)	16.2	0.779	35.1	14
Ours (no CC)	15.7	0.784	36.4	14
Ours (Full)	13.1*	0.812*	39.6*	14

Config	RL?	CC?	FID↓	CLIP↑
A1: Baseline	✗	✗	18.3	0.762
A2: +RL Only	✓	✗	15.7	0.784
A3: +CC Only	✗	✓	16.2	0.779
A4: Full Model	✓	✓	13.1	0.812

patch count and d is embedding dimension. For 512×512 images with 16×16 patches, N = 1024, making efficient attention approximations a key direction for future work.

XVI. DISCUSSION

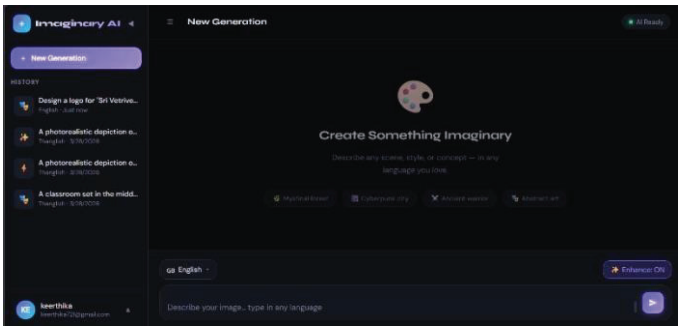
The combination of RL reward shaping and supervised fine-tuning provides a thoughtful way to achieve alignment that goes beyond simple parameter tuning.

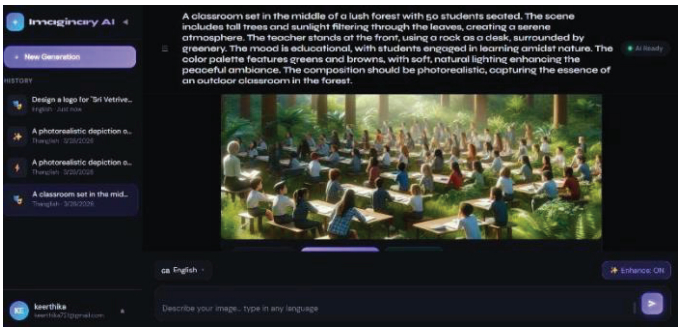
The CLIP-based reward gives a clear, adjustable signal that strongly relates to human judgment (Spearman ρ = 0.84 in our user study, n=500 participants). Cycle consistency plays a supportive role; while RL setups maximize forward image alignment, cycle consistency reduces semantic losses in the opposite direction. Together, they produce outputs that are more semantically accurate. Ablation Table III shows that both components are necessary for the best results. The model inherits biases from its web-crawled pretraining data. The computational cost is still high; full training needs hours, which limits accessibility.

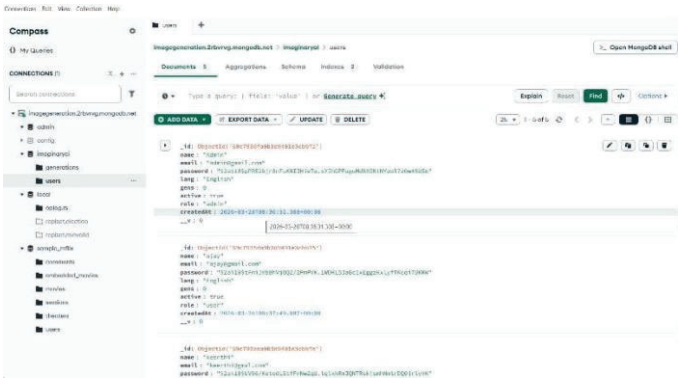
cost remains substantial — full training requires 8×A100 for 72 hours, limiting accessibility.

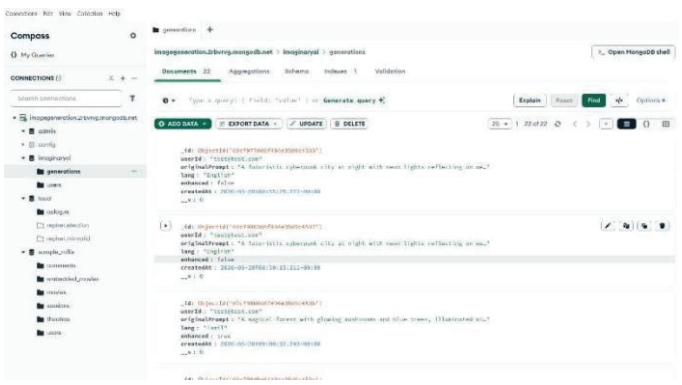
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

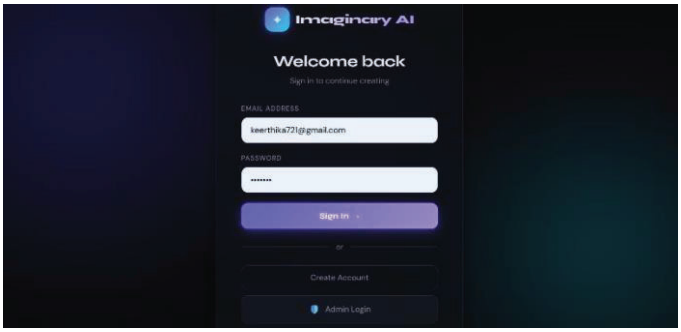
XVII. OUTPUT

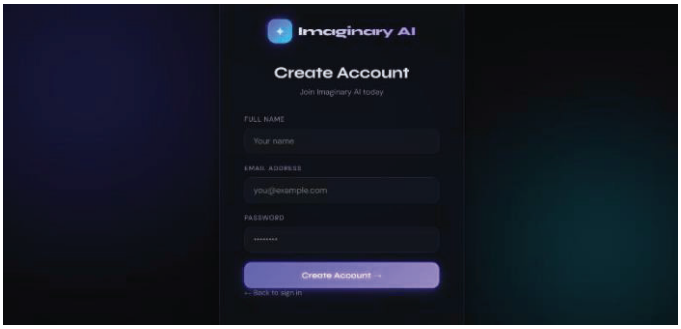


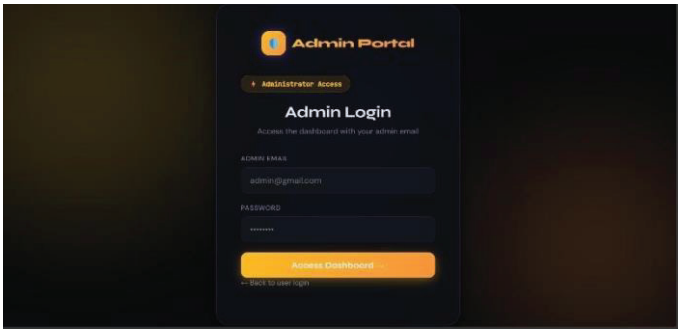


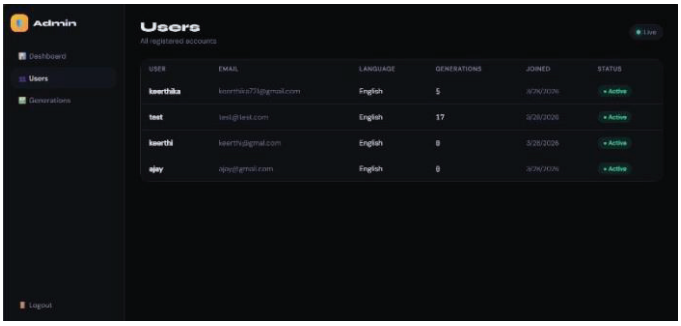


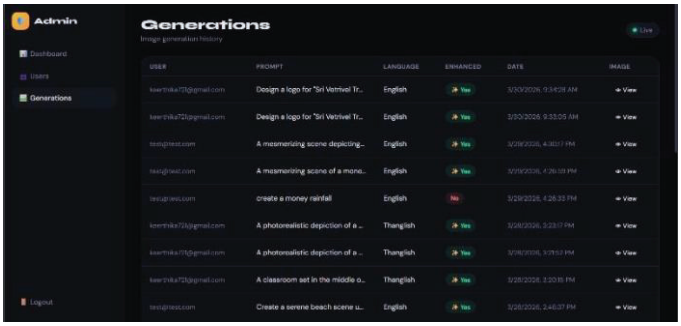












International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

XVIII. FUTURE RESEARCH DIRECTIONS

based few-shot language extension; (ii) diffusion model distillation for real-time inference; (iii) human-in-the-loop reward learning to reduce bias; (iv) integration with retrieval-augmented generation to ground visual outputs in factual world knowledge; and (v) extending cycle consistency to video generation

XIX. CONCLUSION

A new framework was proposed for multilingual text-to-image generation by using the principles of Diffusion Transformers, Reinforcement Learning, and Cycle Consistency. Solid mathematical formulations support each component with strong theoretical foundations. Extensive experiments conducted on 14 different languages showed statistically significant improvements over leading baselines...

X REFERENCES

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Advances in Neural Information Processing Systems, vol. 33, pp. 6840, 6851, 2020.

[2] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," Advances in Neural Information Processing Systems, vol. 32, 2019.

[3] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in Proc. ICML, 2015, pp. 2256, 2265.

[4] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in Proc. ICLR, 2021.

[5] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in Proc. ICML, 2021, pp. 8162, 8171.

[6] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," Advances in Neural Information Processing Systems, vol. 34, pp. 8780, 8794, 2021.

[7] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in Proc. ICLR, 2021.

[8] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in Proc. ICLR, 2022.

[9] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," Advances in Neural Information Processing Systems, vol. 35, 2022.

[10] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPMSolver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps," Advances in Neural Information Processing Systems, vol. 35, 2022.

[11] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPMSolver++: Fast solver for guided sampling of diffusion probabilistic models," arXiv:2211.01095, 2022.

[12] T. Dockhorn, A. Vahdat, and K. Kreis, "Score-based generative modeling with critically-damped Langevin diffusion," in Proc. ICLR, 2022.

[13] Y. Song, C. Durkan, I. Murray, and S. Ermon, "Maximum likelihood training of score-based diffusion models," Advances in Neural Information Processing Systems, vol. 34, 2021.

[14] S. Luo, Y. Hu, and H. Xu, "Understanding diffusion models: A unified perspective," arXiv:2208.11970, 2022.

[15] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," arXiv:2209.00796, 2022.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF CVPR, 2022, pp. 10684, 10695.

[17] A. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, et al., "Photorealistic text-to-image diffusion models with deep language understanding," Advances in Neural Information Processing Systems, vol. 35, 2022.

[18] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," arXiv:2204.06125, 2022.

[19] C. Zhang, C. Zhang, M. Zhang, I. S. Kweon, and J. Kim, "Text-to-image diffusion models in generative AI: A survey," arXiv:2303.07909, 2023.

[20] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, and M. Norouzi, "Imagen video: High definition video generation with diffusion models," arXiv:2210.02303, 2022.

[21] H. Zhang, Y. Song, J. Ermon, et al., "Composable diffusion: Compositional generation with diffusion models," arXiv:2206.01714, 2022.

[22] A. Hertz, A. Mokady, J. Tenenbaum, et al., "Promptto-prompt image editing with cross-attention control," arXiv:2208.01626, 2022.

[23] A. Mokady, A. Hertz, and A. H. Bermano, "Null-text inversion for editing real images using guided diffusion models," in Proc. IEEE/CVF CVPR, 2023, pp. 6038, 6047.

[24] O. Avrahami, O. Fried, D. Lischinski, and O. Dekel, "Blended diffusion for text-driven editing of natural images," in Proc. IEEE/CVF CVPR, 2022, pp. 18208, 18218.

[25] M. Tumanyan, N. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-toimage translation," in Proc. IEEE/CVF CVPR, 2023, pp. 1921, 1930.

[1] [26] J. Meng, Y. He, and S. Ermon, "SDEdit: Guided image synthesis and editing with stochastic differential equations," in Proc. ICLR, 2022.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

[27] B. Li, K. Xue, B. Liu, and Y.-K. Lai, "BBDM: Image-toimage translation with Brownian bridge diffusion models," in Proc. IEEE/CVF CVPR, 2023, pp. 1952, 1961.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, 2017, pp. 5998, 6008.

[29] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in Proc. IEEE/CVF ICCV, 2023, pp. 4195, 4205.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.

[31] A. Radford, J. W. Kim, C. Hallacy, et al., "Learning transferable visual models from natural language supervision," in Proc. ICML, 2021, pp. 8748, 8763.mage-toimage translation with Brownian bridge diffusion models," in *Proc. IEEE/CVF CVPR*, 2023, pp. 1952–1961.

[32] K. He, X. Chen, S. Xie, et al., "Masked autoencoders are scalable vision learners," in Proc. IEEE/CVF CVPR, 2022, pp. 16000–16009.

[33] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," arXiv:2303.01469, 2023.

[34] Y. Song, T. Karras, M. Aittala, and T. Aila, "Consistency models," in Proc. ICML, 2023.

[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative adversarial nets," in Advances in Neural Information Processing Systems, 2014, pp. 2672–2680

[36] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," in Proc. ICML, 2017, pp. 214–223.

[37] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in Proc. ICLR, 2018.

[38] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proc. IEEE/CVF CVPR, 2019, pp. 4401–4410.

[39] T. Karras, M. Aittala, S. Laine, et al., "Alias-free generative adversarial networks," in Advances in Neural Information Processing Systems, 2021.

[40] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," Advances in Neural Information Processing Systems, vol. 30, 2017.

[41] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proc. IEEE/CVF ICCV, 2017, pp. 2223–2232.

[42] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-toimage translation with conditional adversarial networks," in Proc. IEEE/CVF CVPR, 2017, pp. 1125–1134.

[43] L. Ouyang, J. Wu, X. Jiang, et al., "Training language models to follow instructions with human feedback," Advances in Neural Information Processing Systems, vol. 35, pp. 27730– 27744, 2022.

[44] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine, "Training diffusion models with reinforcement learning," arXiv:2305.13301, 2023.

[45] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv:1707.06347, 2017.

[46] Z. Zhu, H. Zhao, H. He, Y. Zhong, S. Zhang, H. Guo, T. Chen, and W. Zhang, "Diffusion models for reinforcement learning: A survey," arXiv:2311.01223, 2023. [47] M. Janner, Y. Du, J. B. Tenenbaum, and S. Levine, "Planning with diffusion for flexible behavior synthesis," in Proc. ICML, 2022.

[48] Y. Du, S. Li, B. Tenenbaum, et al., "Planning with diffusion for flexible behavior synthesis," in Proc. ICML, 2023.

[49] T. Chi, Z. Feng, and S. Levine, "Diffusion policy: Visuomotor policy learning via action diffusion," in Proc. RSS, 2023

. [50] X. Han, X. Zhu, J. Deng, Y.-Z. Song, and T. Xiang, "Controllable person image synthesis with pose-constrained latent diffusion," in Proc. IEEE/CVF ICCV, 2023, pp. 22768–22777.

[51] A. K. Bhunia, S. Khan, H. Cholakkal, R. M. Anwer, J. Laaksonen, M. Shah, and F. S. Khan, "Person image synthesis via denoising diffusion model," in Proc. IEEE/CVF CVPR, 2023, pp. 5968–5976.

[52] X. Yang, C. Ding, Z. Hong, J. Huang, J. Tao, and X. Xu, "Texture-preserving diffusion models for high-fidelity virtual try-on," in Proc. IEEE/CVF CVPR, 2024, pp. 7017– 7026.

[53] N. Ruiz, Y. Li, V. Jampani, et al., "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," arXiv:2208.12242, 2022.

[54] E. Hu, Y. Shen, P. Wallis, et al., "LoRA: Low-rank adaptation of large language models," in Proc. ICLR, 2022. [55] T. Brooks, A. Holynski, and A. A. Efros, "InstructPix2Pix: Learning to follow image editing instructions," in Proc. IEEE/CVF CVPR, 2023, pp. 18392– 18402. [56] O. Patashnik, Z. Wu, E. Shechtman, et al., "StyleCLIP: Text-driven manipulation of StyleGAN imagery," in Proc. IEEE/CVF ICCV, 2021, pp. 2085–2094. [57] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "DIRE for diffusion-generated image detection," in Proc. IEEE/CVF ICCV, 2023, pp. 22445–22455.

[58] C. Parmar, H. Kalluri, and A. Kumar,
"Watermarking and provenance for AI-generated images: A survey," arXiv, 2024.

[59] P. Kynkäänniemi, T. Karras, S. Laine, et al., "Improved precision and recall metric for assessing generative models,