# Latent Diffusion Prior Feature Matching and Zero-Shot Text-to-Image Benchmark Performance

Assignee Research

June 12, 2026

**Abstract**

There has been tremendous progress in large-scale text-to-image synthesis driven by diffusion models enabling versatile downstream applications such as 3D object synthesis from texts, image editing, and customized generation. We present a generic approach using latent diffusion models as powerful image priors for various visual synthesis tasks. Existing methods that utilize such priors fail to use these models' full capabilities. To improve this, our core ideas are 1) a feature matching loss between features from different layers of the decoder to provide detailed guidance and 2) a KL divergen

## 1 Introduction

This paper examines: Text-driven Visual Synthesis with Latent Diffusion Prior. Research question: How does feature matching loss in latent diffusion priors impact CLIP score and FID metrics compared to pixel-space reconstruction losses on zero-shot text-to-image benchmarks?.

## 2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.4/10.

## 3 Results

10 papers retrieved. 21 claims extracted; 16 independently verified. Quality review score: 7.4/10.

## 4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

## 5 Extracted Claims

| Claim | Verified | Confidence |
|---|---|---|
| The proposed method produces a significantly better FID score compared to StyleGANFusion. | × | 0.12 |
| The proposed method produces a competitive CLIP score compared to StyleGANFusion. | × | 0.13 |
| Pioneering image manipulation methods utilize GANs to achieve editing of appearances while preserving the shape. | ✓ | 0.16 |
| GAN-based approaches are often restricted by training image domains and text expression constraints. | ✓ | 0.16 |
| Advanced methods have leveraged embeddings from a pretrained CLIP model to update generative models with test-time optim | ✓ | 0.24 |
| Text-to-image diffusion models have shown exceptional success in manipulation tasks. | × | 0.13 |
| The proposed method manipulates images using test-time optimization with the proposed diffusion guidance. | ✓ | 0.22 |
| The proposed method produces more detailed results than latent diffusion-guided baselines. | ✓ | 0.17 |
| The proposed method produces more detailed results than the CLIP-guided method Text2LIVE. | ✓ | 0.16 |
| The proposed feature matching loss (LFM) uses the decoder of the stable diffusion autoencoder. | ✓ | 0.24 |
| The feature matching loss LFM is inspired by the GAN discriminator feature matching loss proposed in pix2pixHD. | ✓ | 0.29 |
| The proposed approach measures similarity of extracted features from a pretrained and fixed decoder, not from an additio | ✓ | 0.16 |
| The proposed approach uses latent code with added noise residual and the clean latent as decoder input, as opposed to re | ✓ | 0.25 |
| Direct gradient computation involving the UNet Jacobian is computationally expensive. | × | 0.15 |
| CLIP models are not generative models. | ✓ | 0.16 |
| The contrastive objective of CLIP models may not preserve visual information useful for synthesis-oriented tasks. | ✓ | 0.23 |
| Diffusion models can achieve competitive or better performance compared to CLIP-based approaches in synthesis tasks. | ✓ | 0.24 |
| Score Distillation Sampling uses an image diffusion model as a prior to train a NeRF model without backpropagating throu | ✓ | 0.18 |
| Latent score distillation (LSD) computes the loss in the latent space. | × | 0.11 |
| Methods using score distillation with trained diffusion models compute the loss in a limited spa- | ✓ | 0.20 |

## References


- http://arxiv.org/abs/2412.01496v2
- http://arxiv.org/abs/2411.12832v1
- http://arxiv.org/abs/2302.08510v2