Impact of Domain-Specific Fine-Tuning on Zero-Shot Accuracy of Text-to-Image Models

Assignee Research

June 11, 2026

Abstract

Human motion generation is essential for fields such as animation, robotics, and virtual reality, requiring models that effectively capture motion dynamics from text descriptions. Existing approaches often rely on Contrastive Language-Image Pretraining (CLIP)-based text encoders, but their training on text-image pairs constrains their ability to understand temporal and kinematic structures inherent in motion and motion generation. This work introduces MoCLIP, a fine-tuned CLIP model with an additional motion encoding head, trained on motion sequences using contrastive learning and tethering lo

1 Introduction

This paper examines: MoCLIP: Motion-Aware Fine-Tuning and Distillation of CLIP for Human Motion Generation. Research question: To what extent does fine-tuning the semantic alignment module on domain-specific datasets (e.g., medical or scientific images) improve the zero-shot accuracy of text-to-image models on specialized benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.9/10.

3 Results

15 papers retrieved. 11 claims extracted; 8 independently verified. Quality review score: 6.9/10.

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
MoCLIP improves Top-1, Top-2, and Top-3 accuracy compared to baselines.	×	0.11
MoCLIP maintains competitive FID (Frchet Inception Distance) scores.	×	0.06
The method selects M2T-Interpretable [34] as the foundation for the motion encoder.	✓	0.17
The method introduces cross-limb attention connections that extend beyond conventional skeletal adjacency constraints.	✓	0.18
Direct attention connections are introduced between both hands and both feet.	×	0.15
Temporal attention mechanisms are applied to encoded motion features before pooling along the temporal dimension.	✓	0.19
The model uses a symmetric cross-entropy loss for contrastive alignment between motion and text embeddings.	✓	0.17
A feature distillation loss (Tethering Loss) is used to preserve original semantic representations.	✓	0.17
The distillation loss constrains the student text encoder using the pre-trained teacher text encoder.	✓	0.23
HumanML3D [14] and KIT-ML [31] are the two major datasets used for motion generation tasks in this study.	✓	0.16
The proposed model relies on pre-trained weights from chosen baseline models on HumanML3D and KIT-ML datasets.	✓	0.18

References

- http://arxiv.org/abs/2605.04772v1
- http://arxiv.org/abs/2304.03119v1
- http://arxiv.org/abs/2505.10810v1