New chat
Search chats
Library
Codex
Atlas

GPTs
Explore
Canva

Projects
New project
MA-Paper
Write abstract and intr...
Math and Code Implem...
Final project discussion
User study execution cl...
Modular quality metrics...
See All
MA-Pro
MA
IG & Ad Marketing
Uni

Your chats
BRISQUE deep learning me...

Daniel Sauer
Plus

By contrast, SSIM, LPIPS, and COVER displayed weak or negative correlations, ranging from –0.60 to 0.30. These metrics largely measure spatial similarity, perceptual content fidelity, or aesthetic composition. However, in generative video style transfer, subjective quality is rarely determined by spatial distortions alone. Participants evaluated the videos primarily through their temporal coherence—smoothness of motion, absence of flicker, and stability of stylization—and not through the preservation of pixel-level structure. This explains why spatially focused metrics overfit or misalign: they penalize stylization intensity or reward image sharpness even when viewers prefer more expressive or consistent visual motion. COVER, despite being a state-of-the-art neural VQA system, similarly underperforms because its semantic and technical branches assess content correctness and compression-like artifacts, neither of which dominate failure modes in modern diffusion- or transformer-based video models.

The P.1401 polynomial mapping results (predicted vs. actual MOS) provide the final validation. Motion-based metrics maintain tight clustering around the identity line, with most predictions falling within ±0.5 MOS. This is considered strong predictive performance for small-sample subjective studies. Importantly, prediction stability persists across the full MOS spectrum, demonstrating that motion descriptors do not collapse at high or low quality levels. Conversely, LPIPS, SSIM, CLIP-Score, and COVER predictions show high variance, systematic bias, and large deviations from the ideal line, making them unsuitable for reliable MOS prediction.

Collectively, these results converge on a clear conclusion: **objective evaluation of neural video style transfer requires temporally aware metrics**, and classical image-based or semantics-based quality indicators fail to capture the dominant perceptual characteristics of generative video artifacts. TV-L1–derived temporal features provide the strongest, most generalizable, and most interpretable predictors of subjective quality, establishing them as the most appropriate metrics for future benchmarking of AI-generated stylized videos.