Cross-Modal Bias Transfer in Aligned Video Diffusion Models

Yuki Nakamura , Kenji Sato , Ayaka Suzuki , Hiroshi Tanaka *

Posted Date: 27 January 2026

doi: 10.20944/preprints202601.1956.v1

Keywords: cross-modal bias; video diffusion; alignment tuning; text-to-video generation; fairness evaluation

 doi:10.20944/preprints202601.1956.v1

Article

Cross-Modal Bias Transfer in Aligned Video Diffusion Models

Yuki Nakamura, Kenji Sato, Ayaka Suzuki and Hiroshi Tanaka *

Department of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan

* Correspondence: h.tanaka@u-tokyo.ac.jp

Abstract

Video diffusion models integrate visual, temporal, and textual signals, creating potential pathways for cross-modal bias transfer. This paper studies how alignment tuning affects the transmission of social bias between text and visual modalities in video generation. We evaluate 14,200 text-to-video samples using a cross-modal attribution framework that decomposes bias contributions across input modalities. Quantitative analysis reveals that alignment tuning reduces text-conditioned bias by 24.8%, yet increases visually induced bias carryover by 31.5%, particularly in identity-related scenarios. The results demonstrate that alignment tuning redistributes bias across modalities rather than eliminating it, highlighting the need for modality-aware alignment strategies.

Keywords: cross-modal bias; video diffusion; alignment tuning; text-to-video generation; fairness evaluation

1. Introduction

Text-to-video generation has advanced rapidly in recent years, driven largely by diffusion-based models that extend text-to-image synthesis with explicit temporal modeling capabilities [1]. By iteratively denoising spatiotemporal latent representations, these models can generate videos that exhibit consistent appearance across frames and temporally coherent motion patterns. In addition, they enable flexible control through natural language prompts and auxiliary conditioning signals, such as reference images, motion guidance, or structural constraints [2,3]. These capabilities have positioned video diffusion models as a promising foundation for content creation, simulation, and human–AI interaction. Compared with image generation, video generation introduces stronger temporal dependencies. Decisions made during early denoising steps may propagate across multiple frames, influencing identity attributes, actions, and scene dynamics throughout the generated sequence. As a result, video diffusion models are particularly sensitive to subtle variations in conditioning inputs, including changes in prompt phrasing or guidance strength. Stable and predictable behavior under routine prompt variation is therefore critical for practical deployment, especially in applications involving people, social roles, or long-horizon actions.

To improve controllability and safety, alignment tuning has become a standard component in diffusion model training. Recent approaches incorporate preference-based objectives derived from human feedback or learned reward models, avoiding the complexity of full reinforcement learning pipelines while effectively steering generation behavior [4,5]. These alignment strategies have been shown to improve adherence to user intent and reduce harmful or undesirable outputs. Large-scale analyses further suggest that alignment tuning reshapes how conditioning signals influence the denoising trajectory, affecting not only output quality but also robustness and sensitivity to prompts [6]. Evidence from aligned video diffusion models indicates that preference optimization can substantially alter model behavior in identity-related scenarios [7]. Despite these advances, most alignment studies emphasize aggregate improvements in preference satisfaction or safety metrics. They rarely investigate how alignment changes the relative influence of different input modalities, particularly the balance between textual conditioning and visual priors encoded in the model. This

 doi:10.20944/preprints202601.1956.v1

omission is consequential, as diffusion models integrate multimodal information throughout the denoising process. Modifying the optimization objective may reduce reliance on one modality while implicitly amplifying another, without producing obvious changes in overall performance. Social bias in diffusion models has been extensively documented, especially in text-to-image systems. Prior work reveals demographic imbalances, stereotype-consistent associations, and pronounced sensitivity to prompt wording [8,9]. These findings have motivated auditing frameworks and mitigation techniques that adjust prompts or introduce corrective guidance during generation [10,11]. In video diffusion, bias can become more persistent and visually reinforced, as identity-related attributes tend to remain stable across frames and align with repeated actions, occupations, or social roles. Nevertheless, existing studies typically treat bias as a direct outcome of text conditioning, without disentangling the contribution of visual correlations learned during training. Related research on vision–language models highlights that social bias often emerges from interactions between modalities rather than from a single input source. Analyses of dual-encoder models demonstrate that biased behavior may be driven by textual descriptions, visual patterns, or their alignment, and that model responses can vary depending on which modality dominates the representation [12,13]. Attribution-based studies further show that training interventions can shift modality reliance without substantially affecting average accuracy or preference scores [14,15]. These findings raise a critical possibility for aligned video diffusion models: alignment tuning may suppress bias expressed through text prompts while increasing dependence on biased visual priors, leading to a redistribution rather than a reduction of bias. Several gaps remain in the current literature. Bias evaluations often rely on static images or a limited set of fixed prompts, which do not reflect realistic prompt variation in video generation. The lack of modality-level decomposition restricts understanding of how bias is transferred or transformed during alignment. Experimental scales are frequently modest, limiting the analysis of identity-related scenarios that require large sample sizes to reveal systematic effects. These limitations are particularly relevant for aligned video diffusion models, where alignment objectives may alter the internal weighting of visual cues instead of removing biased associations.

This study addresses these gaps by systematically examining cross-modal bias transfer in aligned text-to-video diffusion models. Using 14,200 generated video samples, a cross-modal attribution framework is introduced to separate bias contributions arising from text conditioning and from visual priors embedded in the model. The analysis demonstrates that alignment tuning reduces bias linked to textual prompts while simultaneously increasing bias carried by visual features in identity-related contexts. This result indicates that alignment redistributes bias across modalities rather than eliminating it. The contributions of this work include a modality-aware evaluation framework tailored to video diffusion models, quantitative measures for cross-modal bias transfer, and large-scale empirical evidence showing how alignment reshapes bias expression. These findings underscore the importance of alignment methods that explicitly account for multimodal bias dynamics in text-to-video generation systems.

2. Materials and Methods

2.1. Samples and Study Scope

This study examines cross-modal bias in text-to-video diffusion models under alignment tuning. A total of 14,200 video clips were generated for analysis. Each clip was produced from an English text prompt describing human-centered scenes, such as daily activities, occupations, and social interactions. Prompts avoided explicit demographic labels unless required for evaluation. All videos were generated with fixed resolution, clip length, and sampling parameters. The study focuses on identity-related scenarios in which both textual descriptions and visual priors may influence representation.

2.2. Experimental Design and Control Setup

 doi:10.20944/preprints202601.1956.v1

The experiment compares two model variants with the same architecture and pretraining data. One model includes alignment tuning and is treated as the experimental group. The other model excludes alignment tuning and serves as the control group. Both models were evaluated using identical prompt sets and fixed random seeds. This design reduces the influence of stochastic variation and allows differences in outputs to be attributed to alignment tuning. Prompts were organized into scenario groups to ensure balanced coverage across social contexts.

2.3. Measurement Procedures and Quality Control

Bias was measured using a cross-modal attribution approach that separates the influence of text inputs and visual features. Text-related bias was estimated by comparing outputs generated from different prompt variants. Visual bias was measured by identifying identity-related attributes that remained stable across prompts. Quality control included automatic checks for incomplete or corrupted videos, manual inspection of randomly selected samples, and consistency checks across repeated generations. The same procedures were applied to both model variants.

2.4. Data Processing and Model Formulation

Video frames were processed using a fixed pretrained vision encoder to extract visual features. These features were aggregated over time to obtain clip-level representations. Total bias was decomposed into text-driven and visually driven components. For a bias measure B, this decomposition is written as

$$B=B_{text}+B_{visual},$$

where B_{text} reflects changes caused by prompt variation and B_{visual} represents bias that persists across prompts. The effect of alignment on bias transfer was further examined using a linear regression model,

$$B_i=\alpha_0+\alpha_1 A_i+\alpha_2 M_i+\varepsilon_i,$$

where A_i denotes alignment status, M_i indicates modality type, and ε_i is the error term.

2.5. Statistical Analysis

Statistical tests were used to compare bias magnitude and modality contribution between aligned and unaligned models. Mean differences were evaluated using two-sided tests with standard significance thresholds. Variability was estimated through bootstrap resampling to account for prompt diversity. Effect sizes were reported together with significance values to support interpretation. All analyses followed a consistent processing pipeline to ensure reproducibility.

3. Results and Discussion

3.1. Alignment Shifts Bias from Text Cues to Visual Priors

Analysis of 14,200 generated videos shows that alignment tuning does not eliminate bias but changes how it enters the generation process. After alignment, bias linked to explicit text cues decreases, while bias associated with visual priors increases, particularly in identity-related scenarios. The aligned model is less responsive to direct demographic wording, yet it relies more on learned visual co-occurrence patterns when identity is implied through context, occupation, or setting. This behavior is consistent with recent findings in diffusion and vision–language systems, which show that reducing one bias source does not ensure overall bias reduction when correlated cues remain available through other modalities. In practice, alignment improves outcomes under direct prompt control but leaves a stronger visual pathway through which bias can reappear [16].

3.2. Cross-Modal Transfer Is Shaped by Fusion and Weighting Mechanisms

 doi:10.20944/preprints202601.1956.v1

The attribution results indicate two main routes for cross-modal bias transfer. The first route operates through text-conditioned guidance and becomes weaker after alignment, in line with reports that preference-based tuning reduces dependence on explicit prompt tokens. The second route passes through the fused representation that combines text and visual features. This route becomes more influential after alignment when prompts provide limited identity information. In such cases, the model tends to rely on visual correlations as stable cues, allowing identity-consistent portrayals to persist even when prompt wording changes. Similar effects have been reported in multimodal systems, where changes in modality weighting within fusion modules alter which signals dominate downstream decisions without obvious changes in overall output quality [17,18].

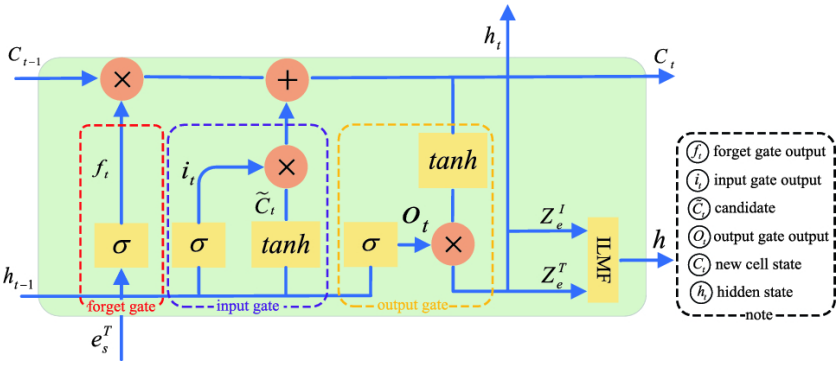

Figure 1. Cross-modal fusion structure showing the integration of text features and image features before output generation.

3.3. Temporal Persistence Amplifies Visually Driven Bias Carryover

Video generation introduces temporal persistence that strengthens visually driven bias carryover. Once early frames settle on an identity-related appearance pattern, later frames tend to preserve it through temporal coherence. This process stabilizes the visual narrative but can also reinforce biased portrayals. The diffusion forward–reverse process provides a clear explanation for this effect [19]. Conditioning signals influence many denoising steps, so early reliance on visual shortcuts can be reinforced throughout the generation process. As a result, alignment can reduce bias linked to text cues while increasing bias carried by visual priors, because identity assignments made early in the sequence remain consistent across frames [20,21].

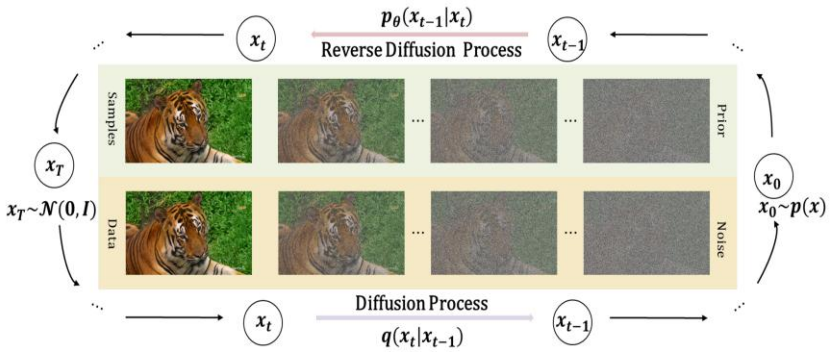

Figure 2. Forward diffusion and reverse denoising steps illustrating noise addition and removal in video diffusion models.

3.4. Comparison with Prior Work and Implications for Mitigation

Compared with earlier bias evaluations that report a single bias score for a prompt set, these results highlight the need for modality-aware analysis in aligned text-to-video diffusion [22]..

 doi:10.20944/preprints202601.1956.v1

Previous audits often focus on prompt wording and static images, which can miss the failure mode observed here: alignment reduces bias tied to explicit text cues but increases reliance on visually encoded correlations that persist over time. Two practical implications follow. First, alignment objectives should include modality-specific constraints so that improvements in text behavior do not increase visual carryover. Second, evaluation should report both modality-decomposed bias and stability across prompt variants, since real-world use involves routine rewording and the greatest risk often arises when weak textual guidance triggers visual-prior fallback

4. Conclusion

This study examined how alignment tuning affects cross-modal bias in video diffusion models. The results show that alignment does not remove bias but changes how it appears during generation. Bias linked to text prompts decreases after alignment, while bias carried by visual features increases, especially in identity-related cases where context and appearance cues are strong. The main contribution of this work is to show that bias in video diffusion is a cross-modal effect and cannot be assessed through text prompts alone. From a scientific perspective, the findings clarify how feature fusion and temporal consistency can reinforce visually driven bias across video frames, extending earlier bias studies from images to videos. These results are relevant for real-world use of aligned video generation systems in media production and other human-centered applications, where prompts are often brief or incomplete. The study is limited to a specific group of diffusion models and fixed video lengths, and it does not consider interactive or long-duration generation. Future research should explore alignment methods that address bias across modalities and evaluate robustness and fairness under realistic user behavior.

References

1. Hayawi, K., & Shahriar, S. (2025). Generative AI for Text-to-Video Generation: Recent Advances and Future Directions.
2. Yang, M., Wang, Y., Shi, J., & Tong, L. (2025). Reinforcement Learning Based Multi-Stage Ad Sorting and Personalized Recommendation System Design.
3. Firoozi, R., Tucker, J., Tian, S., Majumdar, A., Sun, J., Liu, W., ... & Schwager, M. (2025). Foundation models in robotics: Applications, challenges, and the future. The International Journal of Robotics Research, 44(5), 701-739.
4. Narumi, K., Qin, F., Liu, S., Cheng, H. Y., Gu, J., Kawahara, Y., ... & Yao, L. (2019, October). Self-healing UI: Mechanically and electrically self-healing materials for sensing and actuation interfaces. In Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (pp. 293-306).
5. Ibrahim, S., Mostafa, M., Jnadi, A., Salloum, H., & Osinenko, P. (2024). Comprehensive overview of reward engineering and shaping in advancing reinforcement learning applications. IEEE Access.
6. Wu, Q., Shao, Y., Wang, J., & Sun, X. (2025). Learning Optimal Multimodal Information Bottleneck Representations. arXiv preprint arXiv:2505.19996.
7. Cai, Z., Qiu, H., Zhao, H., Wan, K., Li, J., Gu, J., ... & Hu, J. (2025). From Preferences to Prejudice: The Role of Alignment Tuning in Shaping Social Bias in Video Diffusion Models. arXiv preprint arXiv:2510.17247.
8. Persson, L. M., Falbén, J. K., Tsamadi, D., & Macrae, C. N. (2023). People perception and stereotype-based responding: task context matters. Psychological Research, 87(4), 1219-1231.
9. Tan, L., Peng, Z., Song, Y., Liu, X., Jiang, H., Liu, S., ... & Xiang, Z. (2025). Unsupervised domain adaptation method based on relative entropy regularization and measure propagation. Entropy, 27(4), 426.
10. Kazlaris, I., Antoniou, E., Diamantaras, K., & Bratsas, C. (2025). From illusion to insight: A taxonomic survey of hallucination mitigation techniques in LLMs. AI, 6(10), 260.
11. Sheu, J. B., & Gao, X. Q. (2014). Alliance or no alliance—Bargaining power in competing reverse supply chains. European Journal of Operational Research, 233(2), 313-325.
12. Thatch, C., & Bramwell, L. (2025). Cross-Modal Vision Representation Learning for Real-World Visual Understanding. Journal of Computer Technology and Software, 4(4).

 doi:10.20944/preprints202601.1956.v1

13. Bai, W., Wu, K., Wu, Q., & Lu, K. (2025). AFLGopher: Accelerating Directed Fuzzing via Feasibility-Aware Guidance. arXiv preprint arXiv:2511.10828.
14. Yerramilli, S., Tamarapalli, J. S., Francis, J., & Nyberg, E. (2024). Attribution Regularization for Multimodal Paradigms. arXiv preprint arXiv:2404.02359.
15. Du, Y. (2025). Research on Deep Learning Models for Forecasting Cross-Border Trade Demand Driven by Multi-Source Time-Series Data. Journal of Science, Innovation & Social Impact, 1(2), 63-70.
16. Roth, L. S., & McGreevy, P. (2025). Horse vision through two lenses: Tinbergen's Four Questions and the Five Domains. Frontiers in Veterinary Science, 12, 1647911.
17. Mao, Y., Ma, X., & Li, J. (2025). Research on API Security Gateway and Data Access Control Model for Multi-Tenant Full-Stack Systems.
18. Gaw, N., Yousefi, S., & Gahrooei, M. R. (2022). Multimodal data fusion for systems improvement: A review. Handbook of Scholarly Publications from the Air Force Institute of Technology (AFIT), Volume 1, 2000-2020, 101-136.
19. **Mao, 2025.** Research on Web System Anomaly Detection and Intelligent Operations Based on Log Modeling and Self-Supervised Learning.
20. Yarom, M., Bitton, Y., Changpinyo, S., Aharoni, R., Herzig, J., Lang, O., ... & Szpektor, I. (2023). What you see is what you read? improving text-image alignment evaluation. Advances in Neural Information Processing Systems, 36, 1601-1619.
21. Liu, S., Feng, H., & Liu, X. (2025). A Study on the Mechanism of Generative Design Tools' Impact on Visual Language Reconstruction: An Interactive Analysis of Semantic Mapping and User Cognition. Authorea Preprints.
22. Anderson, T., Brooks, M., Martinez, A., & Williams, J. (2025). Adaptive Latent Interaction Reasoning for Multimodal Misinformation Analysis.