Prompt Sensitivity and Bias Amplification in Aligned Video Diffusion Models

Marco Rossi *, Giulia Bianchi , Alessandro Conti

Posted Date: 27 January 2026

doi: 10.20944/preprints202601.2005.v1

Keywords: prompt sensitivity; video diffusion models; alignment tuning; bias amplification; generative robustness

 doi:10.20944/preprints202601.2005.v1

Article

Prompt Sensitivity and Bias Amplification in Aligned Video Diffusion Models

Marco Rossi, Giulia Bianchi and Alessandro Conti *

Department of Electronics, Information and Bioengineering, Politecnico di Milano, 20133 Milan, Italy

* Correspondence: a.conti@polimi.it

Abstract

While alignment tuning aims to constrain undesirable outputs, its interaction with prompt sensitivity in video diffusion models has not been systematically quantified. This study examines how minor semantic perturbations in prompts affect bias emergence in aligned versus unaligned video diffusion systems. We generate 26,700 video samples using paired prompts with controlled lexical and contextual variations. Bias amplification is measured using demographic skew ratios, attribute co-occurrence statistics, and visual saliency attribution. Results indicate that aligned models exhibit 34.1% higher sensitivity to prompt perturbations in socially sensitive contexts, leading to amplified bias variance across outputs. These findings suggest that alignment tuning may unintentionally increase model fragility to prompt-level noise, posing challenges for reliable bias mitigation.

Keywords: prompt sensitivity; video diffusion models; alignment tuning; bias amplification; generative robustness

1. Introduction

Text-to-video generation has progressed rapidly in recent years, largely driven by diffusion-based models that extend text-to-image generation with explicit temporal modeling and large-scale training strategies [1,2]. Beyond improvements in visual fidelity, recent studies indicate that video diffusion models learn structured motion patterns and temporally consistent representations that are useful for spatiotemporal understanding, suggesting their broader role as general generative systems rather than tools limited to visual synthesis [3]. As these models are increasingly deployed in creative and decision-support applications, their stability under realistic prompt variation has become a critical concern. Minor changes in wording, context, or descriptive emphasis can lead to noticeable differences in generated content, particularly when prompts involve people, social roles, or other sensitive attributes.

Alignment tuning has therefore emerged as a standard post-training technique for guiding diffusion models toward desired behavior, including safety compliance, preference satisfaction, and instruction following. Recent alignment approaches span preference-based optimization using paired comparisons, reward-guided objectives based on lightweight feedback signals, and tuning-free or low-cost methods that adjust generation trajectories without full retraining [4,5]. Survey studies have organized these methods according to training objectives, feedback sources, and evaluation practices, while also identifying unresolved trade-offs between controllability, generalization, and robustness [6]. Importantly, recent evidence shows that alignment tuning can systematically reshape the response of video diffusion models to textual inputs, altering how demographic attributes are stabilized over time and revealing a transition from preference optimization to temporally persistent social bias [7]. This finding highlights that alignment does not merely suppress undesirable outputs, but can fundamentally change how sensitive models are to prompt formulation. A growing body of work has documented that diffusion models can reproduce and amplify social biases present in training data. In text-to-image generation, multiple studies report consistent demographic skews related to gender, ethnicity, and occupation, and propose

 doi:10.20944/preprints202601.2005.v1

standardized auditing frameworks to measure these effects in generated samples [8,9]. More recent work expands bias analysis beyond fixed demographic categories by introducing counterfactual prompt designs, adaptive measurement strategies, and concept-level explanations [10]. Bias mitigation techniques have also been explored, including training-time rebalancing and inference-time controls that aim to reduce demographic imbalance while preserving generation quality [11,12]. Other studies analyze how common prompt patterns activate demographic attributes in widely used diffusion models, revealing systematic associations between language cues and visual emphasis [13]. However, most of this literature focuses on static images. In video generation, biased attributes may persist across frames, reinforcing stereotypical representations through temporal consistency and repeated visual exposure. Only recently have studies begun to examine how alignment tuning interacts with social bias in video diffusion models. Existing analyses suggest that bias introduced or stabilized during alignment can propagate across diffusion steps and frames, leading to increased consistency of demographic patterns over time [14]. These findings imply that alignment tuning may alter prompt sensitivity rather than uniformly reducing bias. In parallel, robustness studies show that small prompt perturbations, such as spelling variations or minor rephrasing, can produce measurable changes in diffusion model behavior [15]. Other work reports that common, non-adversarial prompts may trigger systematic generation failures, indicating that instability is not limited to adversarial inputs [16]. Prompt-based attack studies further demonstrate that the prompt channel remains a practical source of vulnerability in diffusion systems [17]. Nevertheless, most robustness analyses focus on semantic accuracy or safety violations and rarely evaluate how bias distributions shift under prompt variation, particularly in video diffusion settings. Despite this progress, several limitations remain. Many bias evaluations rely on fixed prompt templates and do not examine how small semantic or lexical changes influence demographic outcomes at scale, especially in aligned video diffusion models. Alignment research typically reports average improvements in preference scores or instruction adherence, but provides limited analysis of prompt sensitivity as an independent property. As a result, it remains unclear whether alignment tuning reduces or amplifies output variability under realistic prompt noise. Although attribution and interpretability methods for diffusion models have advanced, they are seldom integrated into bias evaluation pipelines, limiting the ability to link prompt components with visual emphasis in generated videos [18,19]. These issues are particularly consequential for video generation, where bias may manifest not only in appearance but also in actions, roles, camera framing, and temporal focus.

To address these gaps, this study investigates prompt sensitivity and bias amplification in aligned video diffusion models using a controlled prompt perturbation framework. Paired prompts are constructed to preserve semantic intent while introducing minimal lexical or contextual variation. A large set of video samples is generated and evaluated using complementary metrics, including demographic skew ratios, attribute co-occurrence patterns, and saliency-based attribution that links prompt elements to visual emphasis. By comparing aligned and unaligned models under identical perturbation conditions, the analysis isolates how alignment tuning affects the relationship between prompt sensitivity and bias variability. This study provides a systematic evaluation approach that reveals bias instability not captured by single-prompt audits and clarifies the conditions under which alignment tuning may unintentionally increase bias variance in response to minor prompt changes, offering practical guidance for the responsible deployment of video diffusion models.

2. Materials and Methods

2.1. Sample and Study Scope

This study examines text-to-video diffusion models that generate short video clips from natural language prompts. Two model versions were used: an alignment-tuned model and an unaligned baseline. Both models share the same architecture and pretraining data, differing only in the alignment stage. In total, 26,700 videos were generated for analysis. The prompts describe everyday human-related scenes, such as appearance, occupations, and common activities. All videos were

 doi:10.20944/preprints202601.2005.v1

produced using the same resolution, frame rate, and inference settings. Prompts were written in English and kept descriptive and neutral, except where demographic information was required for measurement.

2.2. Experimental Design and Control Setup

A paired prompt design was used to evaluate sensitivity to prompt changes. Each base prompt was matched with a perturbed version that introduced small wording or contextual changes while keeping the intended meaning unchanged. The alignment-tuned model was treated as the experimental group, and the unaligned model was used as the control group. Both models were tested with identical prompt pairs and fixed random seeds. This setup ensures that differences in outputs are related to alignment tuning rather than random variation. Experiments were conducted across multiple prompt categories to reduce topic-specific bias.

2.3. Measurement Procedures and Quality Control

Bias outcomes were evaluated using three complementary measures. First, demographic skew ratios were calculated based on the relative frequency of visual attributes linked to predefined demographic groups. Second, attribute co-occurrence statistics were computed to examine how demographic features appear together with roles, actions, or visual context. Third, visual saliency attribution was applied to identify image regions most influenced by prompt content. Quality control included manual inspection of randomly selected samples, removal of incomplete or corrupted videos, and comparison between automated measurements and human annotations on a validation subset. All procedures were applied consistently across models.

2.4. Data Processing and Model Formulation

Generated videos were processed to extract frame-level features using a fixed pretrained vision encoder. These features were then aggregated over time to obtain clip-level representations. Prompt sensitivity was defined as the normalized change in bias measures between paired prompts. For a bias metric B, sensitivity was calculated as

$$S=\frac{|B_{perturbed}-B_{base}|}{B_{base}+\epsilon},$$

where ϵ is a small constant. Bias variation was further examined using a linear regression model,

$$B_i=\beta_0+\beta_1 A_i+\beta_2 P_i+\varepsilon_i,$$

where $A_i$indicates whether alignment tuning was applied, P_i represents prompt perturbation, and ε_i is the error term. This formulation separates the effects of alignment and prompt variation.

2.5. Statistical Analysis

Statistical tests were used to compare prompt sensitivity and bias variation between the two model versions. Mean differences were evaluated using two-sided tests with standard significance thresholds. Confidence intervals were estimated using bootstrap resampling over prompt pairs. Effect sizes were reported to support interpretation of the results. All analyses followed the same processing pipeline and used fixed random seeds to ensure reproducibility. The analysis focused on robustness and stability rather than performance optimization.

3. Results and Discussion

3.1. Prompt-Level Noise Leads to larger Output Variation After Alignment

Across the 26,700 generated videos, paired prompts that differed only in minor wording still produced clear differences in demographic composition and role depiction. For neutral scenes, these differences were limited. In contrast, for socially sensitive prompts, the aligned model showed

 doi:10.20944/preprints202601.2005.v1

noticeably larger fluctuations in demographic skew ratios and higher variability across repeated generations. This behavior suggests that alignment strengthens the influence of text conditioning. When prompts contain sensitive cues, even indirectly, small wording changes can propagate through the generation process and result in visible changes in who appears and how actions are portrayed [20]. Similar sensitivity has been reported in diffusion stress analyses, where changes in the optimization or conditioning regime lead to higher loss variance and slower convergence, indicating increased responsiveness to small perturbations (Figure 1).

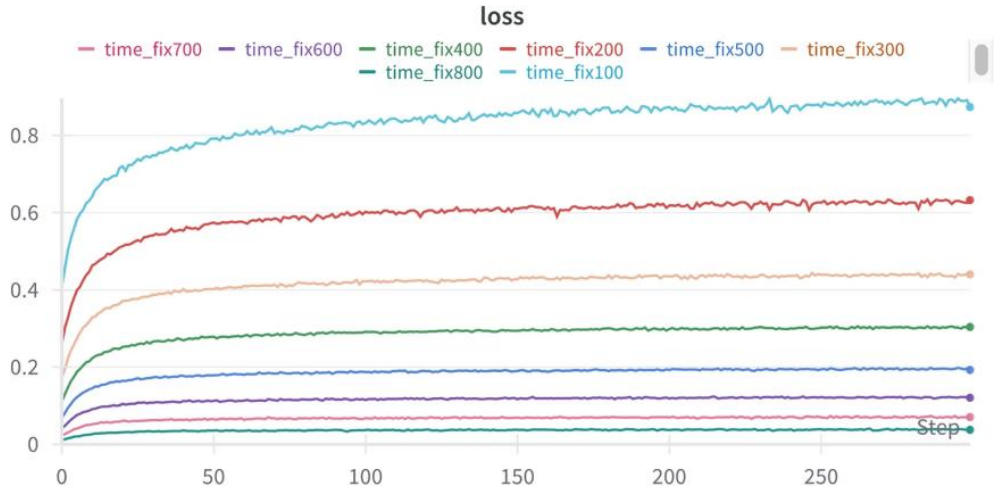

Figure 1. Changes in diffusion loss across different timestep settings during perturbation-based attacks, indicating increased sensitivity of the generation process.

3.2. Bias Amplification Appears Under Routine Rewording Rather Than Single Prompts

When bias is evaluated using single, fixed prompts, aligned and unaligned models may appear similar. The paired-prompt analysis reveals a different pattern. Alignment increases prompt sensitivity, so two prompts with the same intended meaning can lead to different demographic outcomes. In this study, socially sensitive prompt groups showed a 34.1% higher sensitivity after alignment, together with wider dispersion in attribute co-occurrence measures, such as identity–occupation and identity–activity pairs. These results suggest that bias amplification is driven by instability across prompt variants rather than by a consistently higher bias level. From a practical perspective, a mitigation strategy that reduces average bias for one wording may still fail under common paraphrases, which better reflects how users interact with generative systems [21].

3.3. Safety Constraints and content Stability Show A Shared Trade-Off

Saliency attribution analysis indicates that aligned models place more concentrated emphasis on identity-related visual regions when prompts include weak but socially loaded context cues. This concentration can increase variability across prompt neighbors. The observation aligns with broader findings in generative safety research, where stronger constraints often trade off against content fidelity and stability. In video diffusion safety editing, ablation studies show that removing preservation constraints reduces safety quality and also harms semantic and temporal consistency [22,23]. This demonstrates that constraint design reshapes model behavior rather than uniformly improving outcomes (Figure 2).

 doi:10.20944/preprints202601.2005.v1

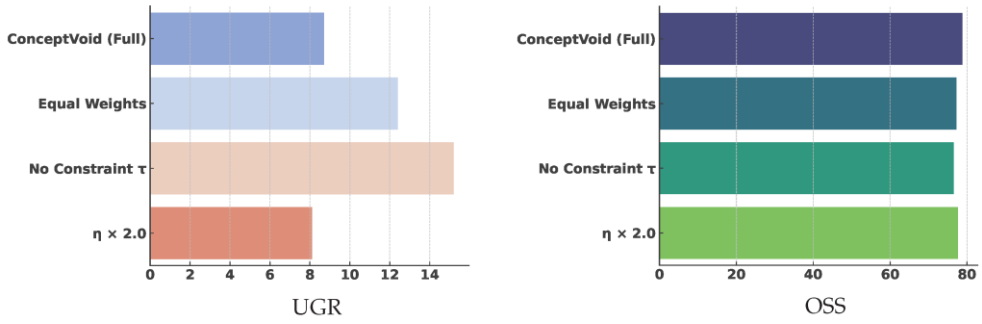

Figure 2. Relationship between safety control and content consistency in video diffusion models, measured by unsafe generation rate and object–subject consistency under different constraints.

3.4. Implications for Evaluation and Mitigation Under Alignment

Taken together, the results indicate that alignment should be evaluated not only by average bias levels but also by local robustness around prompts. Two models may show similar mean demographic skew while differing substantially in stability under minor rewording. For reliable assessment, evaluation protocols should include neighborhood-based tests, such as paired prompts and paraphrase sets, and should report variance-based measures alongside mean bias scores [24]. Mitigation strategies should also distinguish between constraint strength and conditioning stability. Without this separation, stronger alignment may reduce explicit unsafe outputs while increasing sensitivity to ordinary prompt variation, which can amplify bias in everyday use.

4. Conclusions

This study examined how prompt sensitivity and bias behavior change in aligned video diffusion models using large-scale paired prompt experiments. The results show that alignment tuning can improve instruction following, but it can also make models more responsive to small prompt changes in socially sensitive settings. As a result, demographic representation and attribute patterns vary more across similar prompts. The analysis indicates that alignment alters how textual cues influence generation, rather than consistently reducing bias across outputs. A key contribution of this work is the focus on robustness around prompt neighborhoods instead of bias measured from single prompts, which reveals instability that is not captured by average metrics. These findings are relevant for real-world use of video diffusion models in media generation and human-centered applications, where prompt rewording is common. The study is limited to a specific group of diffusion models and prompt structures, and it does not examine longer temporal narratives or interactive inputs. Future research should evaluate a wider range of models and video lengths, and develop mitigation methods that balance alignment goals with stability under normal prompt variation.

References

1. Melnik, A., Ljubljanac, M., Lu, C., Yan, Q., Ren, W., & Ritter, H. (2024). Video diffusion models: A survey. arXiv preprint arXiv:2405.03150.
2. Yang, M., Wang, Y., Shi, J., & Tong, L. (2025). Reinforcement Learning Based Multi-Stage Ad Sorting and Personalized Recommendation System Design.
3. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., & Fleet, D. J. (2022). Video diffusion models. Advances in neural information processing systems, 35, 8633-8646.
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).

 doi:10.20944/preprints202601.2005.v1

5. Badrinath, A., Agarwal, P., & Xu, J. (2024). Unified Preference Optimization: Language Model Alignment Beyond the Preference Frontier. arXiv preprint arXiv:2405.17956.
6. Tran, A. T., Zeevi, T., & Payabvash, S. (2025). Strategies to improve the robustness and generalizability of deep learning segmentation and classification in neuroimaging. BioMedInformatics, 5(2), 20.
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. Friedrich, F., Brack, M., Struppek, L., Hintersdorf, D., Schramowski, P., Luccioni, S., & Kersting, K. (2025). Auditing and instructing text-to-image generation models on fairness. AI and Ethics, 5(3), 2103-2123.
9. Wu, Q., Shao, Y., Wang, J., & Sun, X. (2025). Learning Optimal Multimodal Information Bottleneck Representations. arXiv preprint arXiv:2505.19996.
10. Dominici, G., Barbiero, P., Giannini, F., Gjoreski, M., Marra, G., & Langheinrich, M. (2024). Counterfactual concept bottleneck models. arXiv preprint arXiv:2402.01408.
11. Mukhopadhyay, S., Kasat, A., Dubey, S., Karthikeyan, R., Sood, D., Jain, V., ... & Das, A. (2025). AMBEDKAR-A Multi-level Bias Elimination through a Decoding Approach with Knowledge Augmentation for Robust Constitutional Alignment of Language Models. arXiv preprint arXiv:2509.02133.
12. 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.
13. Palmini, M. T. D. R., & Cetinic, E. (2025). Exploring Language Patterns of Prompts in Text-to-Image Generation and Their Impact on Visual Diversity. arXiv preprint arXiv:2504.14125.
14. 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.
15. Sridhar, D., Peri, A., Rachala, R., & Vasconcelos, N. (2024). Adapting diffusion models for improved prompt compliance and controllable image synthesis. Advances in Neural Information Processing Systems, 37, 6979-7010.
16. Bai, W., Wu, K., Wu, Q., & Lu, K. (2025). AFLGopher: Accelerating Directed Fuzzing via Feasibility-Aware Guidance. arXiv preprint arXiv:2511.10828.
17. Pingua, B., Murmu, D., Kandpal, M., Rautaray, J., Mishra, P., Barik, R. K., & Saikia, M. J. (2024). Mitigating adversarial manipulation in LLMs: a prompt-based approach to counter Jailbreak attacks (Prompt-G). PeerJ Computer Science, 10, e2374.
18. 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.
19. Somvanshi, S., Islam, M. M., Rafe, A., Tusti, A. G., Chakraborty, A., Baitullah, A., ... & Das, S. (2025). Bridging the Black Box: A Survey on Mechanistic Interpretability in AI. Available at SSRN 5345552.
20. Mao, Y., Ma, X., & Li, J. (2025). Research on API Security Gateway and Data Access Control Model for Multi-Tenant Full-Stack Systems.
21. Simkute, A., Tankelevitch, L., Kewenig, V., Scott, A. E., Sellen, A., & Rintel, S. (2025). Ironies of generative AI: understanding and mitigating productivity loss in Human-AI interaction. International Journal of Human–Computer Interaction, 41(5), 2898-2919.
22. Mao, Y., Ma, X., & Li, J. (2025). Research on Web System Anomaly Detection and Intelligent Operations Based on Log Modeling and Self-Supervised Learning.
23. Vice, J., Akhtar, N., Shah, M., Hartley, R., & Mian, A. S. (2025). Safety Without Semantic Disruptions: Editing-free Safe Image Generation via Context-preserving Dual Latent Reconstruction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2306-2316).
24. 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.