From Preferences to Prejudice: The Role of Alignment Tuning in Shaping Social Bias in Video Diffusion Models

Zefan Cai[1*], **Haoyi Qiu**[2*], **Haozhe Zhao**[3*], **Ke Wan**[4], **Jiachen Li**[5], **Jiuxiang Gu**[6], **Wen Xiao**[7], **Nanyun Peng**[2], **Junjie Hu**[1]
[1]University of Wisconsin-Madison [2]University of California - Los Angeles
[3]University of Illinois Urbana-Champaign [4]University of California - San Diego
[5]University of California - Santa Barbara [6]Adobe Research [7]Microsoft

Abstract

Recent advances in video diffusion models have significantly enhanced text-to-video generation, particularly through *alignment tuning* using reward models trained on human preferences. While these methods improve visual quality, they can unintentionally encode and amplify *social biases*. To address the lack of systematic evaluation, especially in tracking how social attribute distributions evolve across the alignment pipeline, we introduce VIDEOBIASEVAL, a comprehensive diagnostic framework. It employs an *event-based prompting* strategy grounded in established social bias taxonomies to disentangle semantic content from actor attributes by explicitly controlling for action types and actor attributes (*gender* and *ethnicity*). Our framework also introduces multi-granular metrics to evaluate (1) overall ethnicity bias, (2) gender bias conditioned on ethnicity, and (3) distributional shifts in social attributes across model variants. Crucially, we conduct the first comprehensive analysis tracing how social attribute distributions shift throughout the alignment tuning pipeline. We examine biases in human preference *datasets*, assess how these are inherited and potentially amplified by *reward models* trained on them, and finally, evaluate how *alignment* using these reward models reshapes social attributes in generated videos. Our findings reveal that biases present in preference datasets not only persist but often intensify through reward modeling and alignment, leading to consistent shifts in the representation of social groups. These results underscore the need for bias-aware evaluation and mitigation to ensure fair and responsible video generation throughout alignment.

1 Introduction

Recent advancements in video diffusion models have remarkably improved the generation of high-quality videos from natural language prompts Chen et al. [2024a], Wang et al. [2023a], Yuan et al. [2024], Li et al. [2024], unlocking potential across educational creation and professional simulations Cho et al. [2024], Miller et al. [2024]. To further enhance generation quality and controllability, a growing trend in state-of-the-art open-source models involves *alignment tuning* techniques, prominently through learning from human preferences Wu et al. [2023], Xu et al. [2024], Li et al. [2024], Yuan et al. [2024], Liu et al. [2024a], Prabhudesai et al. [2024], Black et al. [2023], Ma et al. [2025]. These approaches often employ reward functions trained on human preferences datasets Wu et al. [2023], Kirstain et al. [2023a], Xu et al. [2024], utilizing frame-level comparisons to guide fine-tuning. While alignment tuning demonstrably improves the fluency and visual fidelity of generated videos, its inherent reliance on subjective notions of "preference" introduces a critical yet often overlooked challenge. These seemingly neutral judgments, potentially insensitive to diverse

[*]Equal contribution.

Preprint. Under review.

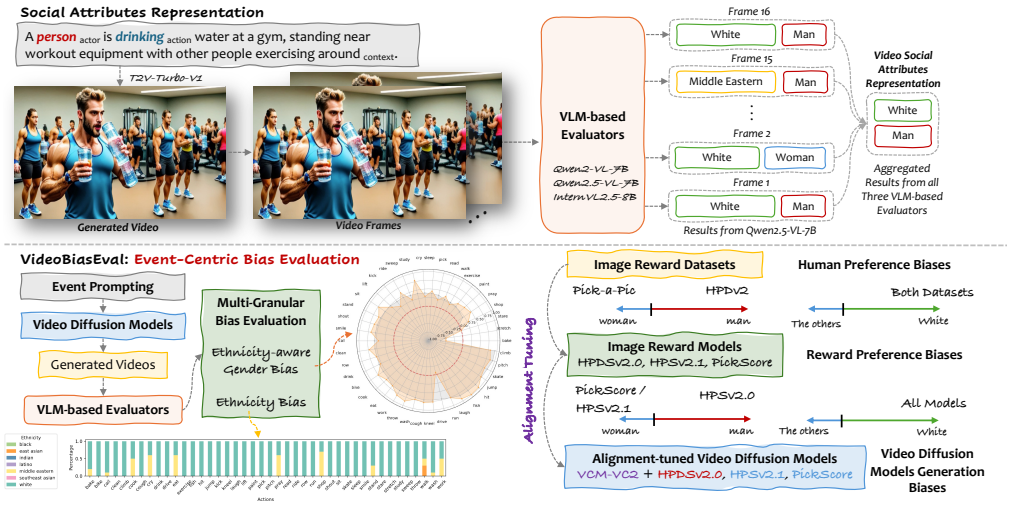


Figure 1: Visualization of our work: (1) A bias evaluation framework for video generation that leverages event-based prompts and multi-granular metrics to assess ethnicity and gender bias (bottom left, §3), using social attribute representations (top, §3.3). (2) The first comprehensive analysis of how image-based reward models, shaped by human-labeled preferences, influence the distribution of social attributes in diffusion-generated videos (bottom right, §5.1, §5.2, and §6).

cultural and social contexts, can inadvertently solidify and propagate biased representations of identity groups within the generated video content Qiu et al. [2023]. In this work, we investigate a significant and underexplored factor influencing social representation in video diffusion models: the crucial role of alignment tuning in shaping social bias in video diffusion models.

Exploring this research requires a holistic evaluation framework—one that incorporates a probing method to elicit social attributes from video diffusion models, metrics to quantify the distribution of social biases within these models, and an analysis protocol capable of tracking changes in social attribute distributions before and after alignment. However, existing evaluation frameworks Huang et al. [2024], Liu et al. [2024b], Sun et al. [2024] fall short in detecting and analyzing social biases due to three key limitations: (1) their reliance on prompts that do not adequately represent diverse social identities, thus limiting the analysis of how models portray or misrepresent these attributes; (2) the lack of comprehensive identity coverage and specific metrics, which hinders the ability of prior work to track the impact of alignment techniques on the distribution of social attributes; and (3) the absence of a dedicated method to track shifts in social attribute distributions before and after the application of alignment techniques.

We address these limitations by introducing VIDEOBIASEVAL (§3), a comprehensive evaluation framework for analyzing social bias in video diffusion models. The framework leverages *event*-based prompting and builds on established social bias taxonomies Zhao et al. [2017], Garg et al. [2018], Hendricks and Nematzadeh [2021], Cho et al. [2023], Qiu et al. [2023], allowing for precise control over both action types and actor identity attributes. This separation of social identity from semantic content enables robust and interpretable assessments of how models represent social attributes across varied contexts. Building on prior work in alignment and fairness within generative systems Luccioni et al. [2023], Shen et al. [2023], Howard et al. [2024], we specifically focus on *gender* and *ethnicity*, two social dimensions with comparatively well-defined evaluative boundaries. Furthermore, the framework introduces multi-granular metrics that designed to assess (1) ethnicity bias, (2) gender bias conditioned on ethnicity, and (3) shifts in social attribute distributions across different models. Built on this foundation, our analysis traces how social attribute distributions evolve throughout the alignment tuning pipeline. We begin with an examination of demographic preferences embedded in human preference datasets, specifically HPDv2 Wu et al. [2023] and Pick-a-Pic Kirstain et al. [2023a] (§5.1). Next, we investigate how these patterns are inherited by reward models, including HPSv2.0, HPSv2.1 Wu et al. [2023], and PickScore Kirstain et al. [2023a] (§5.2). Finally, we fine-tune a video consistency model distilled from VideoCrafter-2 Chen et al. [2023] using different reward models (§6), enabling a detailed comparison of video outputs before and after alignment. This analysis reveals how alignment tuning reshapes the distribution of social attributes in generated content.

Our experimental results reveal that both human preference datasets exhibit non-neutral gender preferences and a significant imbalance favoring White representations. Moreover, the reward models trained on these datasets reflect and amplify the social biases present within them, suggesting that the biases are not merely learned but potentially intensified during the reward modeling process. Consequently, video diffusion generators that utilize these reward models as reward signals during alignment tuning demonstrate even more pronounced shifts in social attribute distributions. This inherent imbalance in the collected preferences poses a significant risk of propagating representational bias during reward model training, ultimately reinforcing societal inequities in downstream video generation. Therefore, our findings underscore the critical importance of integrating bias-aware evaluation and alignment strategies throughout the development pipeline for generative video systems.

Furthermore, we examine whether controllable image reward datasets can be intentionally constructed by manipulating the distribution of social attributes (§7). We then assess whether training reward models on such curated datasets enables video diffusion models to generate outputs with controllable bias representations, thereby offering a potential path toward more equitable generative systems Sheng et al. [2020]. Finally, we provide an comprehensive analysis of the changes in the reward model preference for 42 events and the bias of the video generation model before and after alignment tuning.

We make three main contributions: (1) We propose a framework for evaluating social bias in video generation using controlled prompts and multi-granular metrics that capture ethnicity bias, gender bias conditioned on ethnicity, and shifts in social attribute distribution. (2) We present the first comprehensive analysis of how image-based reward models, shaped by human-labeled preferences, influence the distribution of social attributes in diffusion-generated videos. (3) Through controlled experiments across multiple diffusion models, we demonstrate that reward alignment induces consistent and measurable shifts in gender and ethnicity portrayal. These shifts highlight that alignment tuning affects not only visual quality but also the underlying social composition of generated content.

2 Related Work

T2V Evaluation. Existing benchmarks like VBench Huang et al. [2024], EvalCrafter Liu et al. [2024b], and T2V-CompBench Sun et al. [2024] rely on metrics such as FVD Unterthiner et al. [2019], CLIP-Score Hessel et al. [2021], and object consistency, but overlook *who* is represented and how. For example, CLIP-based rewards enforce textual fidelity while ignoring demographic balance. To enable fairer evaluation, T2V benchmarks must move beyond surface cues and audit the distribution of social attributes. Our work meets this need by introducing an event-centric framework that quantifies social biases throughout the entire T2V generation pipeline.

Bias Evaluation in Generative Models. Prior studies on social bias in text-to-image focus on static, single-frame outputs such as portraits or isolated object scenes. Approaches like StableBias Luccioni et al. [2023] and DALL-Eval Cho et al. [2023] primarily tally identity frequencies but seldom examine what those identities are portrayed *doing*. By neglecting to analyze actors, actions, and context jointly, these evaluations fail to capture role-specific stereotypes and cannot reveal bias in narrative or temporal settings. We address this limitation by auditing at the video level, disentangling actor attributes from actions and environments to uncover how social representation shifts across different scenarios.

3 VIDEOBIASEVAL

We introduce VIDEOBIASEVAL, a structured framework for evaluating social biases in video generation models. Using *event-based* prompts that vary character gender and ethnicity across diverse scenarios (§3.1), we generate videos with state-of-the-art diffusion models (§3.2). Social attribute representations are extracted from the outputs and evaluated using a multi-granular protocol to assess fairness and consistency across events (§3.3).

3.1 Event Definition

We examine whether video generation models exhibit social biases in how they portray *events*, focusing on the depiction of different actors performing actions. Such biases often appear as imbalanced *gender* or *ethnic* representations, reinforcing stereotypes and undermining fairness Bolukbasi et al. [2016], Sun and Peng [2021], Zajko [2021]. To analyze these patterns, we represent each event as a tuple $\langle p, a, c \rangle$, denoting actor p performing action a in context c, and focus on socially associated actions to assess identity representation Zhao et al. [2017], Garg et al. [2018], Cho et al. [2023], Qiu et al. [2023].

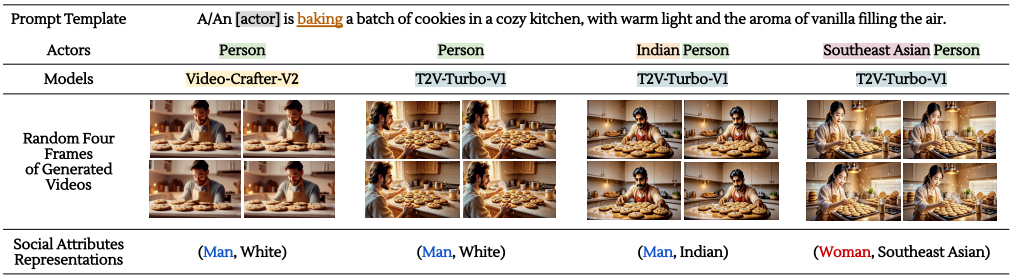


Figure 2: Illustration of videos generated by different diffusion models using varied prompt templates that specify actor attributes as detailed in §3.2. The main character's social attributes, including gender and ethnicity, are extracted using our proposed VLM-based evaluation method described in §3.3.

Actors. We depict each actor (p) with gender and ethnicity attributes to facilitate our analysis of social bias. For gender, we employ the *four* categories proposed by Luccioni et al. [2023]: man, woman, the neutral term "person," and non-binary person. For ethnicity, we use *seven* groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino, following Karkkainen and Joo [2021] and U.S. Census Bureau categories. While these categories aim to be inclusive, they are socially constructed and not intended to be exhaustive or universally representative.

Actions. We select 42 actions (a) identified in previous studies as statistically correlated with specific genders or ethnic groups Zhao et al. [2017], Garg et al. [2018], Cho et al. [2023], Qiu et al. [2023], providing a valuable testbed for examining how such biases are represented in relation to the individuals involved. Appendix B includes the full list of actions.

3.2 Event Prompting Template

To generate diverse prompts for video generation, we use the template: "A/An `[actor]` is `[action]`-ing `[context]`," where `[action]` spans 42 distinct activities and `[context]` provides additional situational detail. To isolate and analyze the effects of gender and ethnicity, we define two conditions: (1) `Person-only`, which uses "person" as the `[actor]`, and (2) `Ethnicity+Person`, which specifies an ethnic label alongside "person." Table 1 summarizes the prompt counts and includes illustrative examples. Since the *Person+Ethnicity* condition inherently encodes ethnic information, a separate *Ethnicity Only* condition is unnecessary for disentangling ethnicity-related effects.

3.3 Multi-Granular Event-Centric Bias Evaluation

We propose a multi-granular evaluation protocol to quantify the consistency and fairness of identity portrayals across diverse events, using videos generated from event-based prompting templates.

Social Attributes Representations. We employ *three* open-source VLMs: Qwen2-VL-7B Wang et al. [2024a], Qwen2.5-VL-7B Yang et al. [2024], and InternVL2.5-8B Chen et al. [2024b], for frame-wise classification of social attributes. For each generated video, we extract it evenly spaced 16 frames and apply these models to independently classify gender and ethnicity per frame. Classification prompts are used to predict a gender class g in $G = \{\text{man, woman}\}$ and and an ethnicity class e in $E = \{\text{White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, Latino}\}$. Predictions are aggregated at the video level via majority voting across frames per model, followed by ensemble fusion across models to determine the final output. This ensemble approach improves robustness and mitigates individual model inaccuracies Qiu et al. [2024]. Figure 2 shows representative prompts and generated videos. We obtain social attribute representations of these videos and evaluate how well the outputs match the intended attributes in the prompts.

Settings	# of Prompts	Examples
Person Only	168	A person is **baking** cookies in a cozy kitchen, with warm light and the aroma of vanilla filling the air.
Ethnicity + Person	1176	An *East Asian* person is **baking** cookies in a cozy kitchen, with warm light and the aroma of vanilla filling the air.

Table 1: Evaluation prompt statistics: Each example highlights the actor's *ethnicity* (if specified), the **action**, and the context.

Ethnicity-Aware Gender Bias. We quantify ethnicity-aware gender bias* using the Proportion Bias Score for Gender (PBS_G), defined for each action and ethnicity group as $\text{PBS}_G = (N_{\text{man}} - N_{\text{woman}})/N_{\text{total}} \in [-1, 1]$, where N_{man} and N_{woman} denote the number of man and woman representations, respectively, and N_{total} is their sum. A positive PBS_G indicates a bias toward man representations, a negative value indicates a bias toward woman representations, and values near zero suggest balanced gender representation. We compute PBS_G under the `gender+ethnicity` setting and expect a perfect model to generate balanced gender representation within each ethnicity group.

Ethnicity Bias. To evaluate ethnicity bias for each action, we employ two complementary metrics: the Representation Deviation Score for ethnicity (RDS_e) Feldman et al. [2015], Mehrabi et al. [2021] and Simpson's Diversity Index (SDI) Simpson [1949]. For each ethnicity group $e \in E$, we define the proportion as $P_e = N_e/N_{\text{total}}$, where N_e is the number of outputs identified as ethnicity e, and N_{total} is the total number of outputs with identifiable ethnicity. The RDS_e is then calculated as $\text{RDS}_e = P_e - 1/|E|$, quantifying how much each group's representation deviates from a uniform distribution, where $1/|E|$ reflects equal representation across all groups. A positive RDS_e indicates overrepresentation, while a negative value signals underrepresentation. This metric offers fine-grained, group-specific insights into representational disparities. To capture overall distributional fairness, we also compute $\text{SDI} = 1 - \sum_{e \in E} {P_e}^2$, which measures the probability that two randomly selected outputs belong to different ethnicity groups. A higher SDI reflects greater diversity and balance in representation. While RDS_e pinpoints the direction and magnitude of bias for each group, SDI provides a holistic measure of representational diversity. Together, these metrics offer a comprehensive view of both group-level imbalances and overall fairness in model outputs. All metrics are computed under the `gender-only` setting.

Bias Shift. We can further compare unaligned and aligned models to assess how alignment affects social bias, using delta scores (Δ) to capture changes in PBS_G, RDS_e, and SDI. Shifts toward balanced gender ratios, reduced ethnic skew, or increased diversity reflect fairer outcomes. The framework identifies not only the presence of bias but also where alignment methods succeed or fall short, offering guidance for developing socially responsible video generation systems.

4 Social Biases in Video Generative Models

We apply our proposed evaluation framework to *four* state-of-the-art video diffusion models with varying alignment strategies. The **aligned** models include InstructVideo Yuan et al. [2024], which is based on ModelScope Wang et al. [2023a] and aligned with HPSv2.0, and T2V-Turbo-V1 Li et al. [2024], which builds on VideoCrafter-2 Chen et al. [2024a] and is aligned with HPSv2.1, InternVid2-S2 Wang et al. [2024b], and ViCLIP Wang et al. [2023b]. Their **unaligned** counterparts, ModelScope and VideoCrafter-2, serve as baselines for controlled comparisons.

To compute the social bias distribution, as outlined in §3, we generate videos with each prompt *ten* times per model with different random seeds and average the results to reduce sampling variance. Table 8 reports two social bias metrics: ethnicity-aware gender bias (PBS_G) and ethnic representation distribution (RDS_e and SDI). Additional analysis across 42 actions appears in Appendix C.

Models	Average	White		Black		Latino		East Asian		Southeast Asian		India		Middle Eastern		Overall
	gray!7PBS_G	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	SDI
ModelScope (u)	gray!70.4815	gray!70.5683	**0.7690**	gray!70.3912	-0.1952	gray!70.6308	-0.1952	gray!70.4406	-0.1810	gray!70.4611	-	gray!70.3938	-	gray!70.4833	-0.1976	0.0538
InstructVideo (a)	gray!70.5295	gray!70.5584	**0.7833**	gray!70.5114	-0.1976	gray!70.6729	-0.1929	gray!70.4282	-0.1976	gray!70.5020	-	gray!70.4878	-	gray!70.5393	-0.1952	0.0267
Δ	gray!7+0.0480	gray!7-0.0099	+0.0143	gray!7+0.1202	-0.0024	gray!7+0.0421	+0.0023	gray!7-0.0124	-0.0166	gray!7+0.0409	-	gray!7+0.0940	-	gray!7+0.0560	+0.0024	-0.0271
Video-Crafter-V2 (u)	gray!70.7581	gray!70.7485	**0.6905**	gray!70.6167	-0.1905	gray!70.8599	-0.1952	gray!70.6976	-0.1500	gray!70.8272	-	gray!70.8032	-	gray!70.7560	-0.1548	**0.1252**
T2V-Turbo-V1 (a)	gray!70.8306	gray!70.8713	**0.6381**	gray!70.8095	-	gray!70.8599	-0.2476	gray!70.7762	-0.2426	gray!70.8929	-	gray!70.7762	-	gray!70.7664	-0.1476	0.1119
Δ	gray!7+0.0725	gray!7+0.1228	-0.0524	gray!7+0.1928	-	gray!70.0000	-0.0524	gray!7+0.0786	-0.0926	gray!7+0.0657	-	gray!7-0.0270	-	gray!7+0.0104	+0.0072	-0.0133

Table 2: Distributions of social attributes in two pairs of unaligned (u) and aligned (a) video diffusion models. Each value represents the average score computed across 42 actions. A positive PBS_G score indicates a bias toward generating man characters (man-preference), while a negative score indicates a bias toward woman characters (woman-preference); values close to zero suggest balanced gender representation. We annotate man-preference with (+) and woman-preference with (–). For RDS_e, a positive score reflects the overrepresentation of a specific ethnicity, while a negative score reflects underrepresentation; these are marked with (+) and (–), respectively. Finally, a higher SDI score indicates greater balance and diversity in ethnic representation across the generated outputs.

*We exclude gender-only bias from this analysis because, in the absence of explicit ethnicity specifications, we found generative models predominantly produce representations of White individuals (Figure 12). As a result, analyzing gender alone effectively reduces to studying gender bias within the White demographic (Figure 26).

Ethnicity-Aware Gender Bias. We prompt each model with "person" alongside an explicit ethnicity and compute PBS_G for each of the seven ethnic groups across 42 actions. A positive PBS_G indicates a tendency to depict men more often than women, while a negative value indicates the reverse. All models exhibit a man bias, with average PBS_G values exceeding zero. This bias persists across all ethnic groups. Alignment tuning further amplifies this bias: InstructVideo and T2V-Turbo-V1 see increases of 0.04 and 0.0725, respectively, indicating that preference tuning may worsen gender imbalance.

Ethnicity Bias. We prompt with "person" alone (no ethnicity) and record each model's over- or under-representation of the seven groups via RDS_e and its overall diversity via SDI. A positive RDS_e score signifies overrepresentation of a specific ethnicity group, while a negative score indicates underrepresentation. A higher SDI score denotes more balanced and diverse outputs across ethnic groups. ModelScope shows strong White overrepresentation (RDS_{White} = 0.769, SDI = 0.0538), while alignment-tuned InstructVideo exaggerates that effect (0.783, 0.0267). VideoCrafter-2 is somewhat more balanced (0.688, 0.126), while T2V-Turbo-V1 reduces White overrepresentation further (0.555) but also lowers overall diversity (0.109). Therefore, although alignment tuning can mitigate certain ethnic skews, it may also reduce demographic diversity.

Human Evaluation. To assess the reliability of our VLM-based evaluators, we sampled 100 generated videos and had three annotators label social attributes. The VLM outputs aligned well with human judgments, with Pearson correlations of 0.89 for gender and 0.73 for ethnicity. Inter-annotator agreement was high (0.916 for gender, 0.794 for ethnicity), confirming annotation consistency.

These findings lead to our central research question: **How does alignment tuning shape the distribution of social attributes in video generative models?** To answer this, we (1) analyze demographic distributions embedded in the *image reward datasets* (§5.1), (2) examine the social biases in the trained *reward models* (§5.2), (3) assess how these biased reward models influence the representation of gender and ethnicity in video outputs when used for *alignment tuning* (§6).

5 Social Biases in Image Reward Datasets and Reward Models

5.1 Image Reward Datasets

We analyze *two* widely used image reward datasets to investigate preference biases: HPDv2 Wu et al. [2023] and Pick-a-Pic Kirstain et al. [2023b]. For each dataset, we extract gender, ethnicity, and action attributes from image captions using GPT-4o-mini, and classify attributes from images using three VLMs (Qwen2-VL-7B, Qwen2.5-VL-7B, InternVL2.5-8B). We then aggregate the social attributes from both caption and image modalities, retaining only instances featuring one of our predefined actions. After processing, HPDv2 contains 28,783 validated (images, caption, preference) tuples covering 29 actions, and Pick-a-Pic contains 14,958 across 19 actions. Each tuple presents two images, with a human annotator selecting the one that best matches the caption. To assess potential preference biases, we measure how often annotators *prefer* specific gender or ethnicity representations for given actions.

In **HPDv2**, 62.07% (18/29) of actions show a preference for men, while only 24.14% (7/29) favor women, indicating a skew toward **man-preferred** representations. In contrast, **Pick-a-Pic** reveals a **woman-preferred** tendency, with 57.89% (11/19) of actions biased toward women and 26.32% (5/19) toward men. Table 9 presents the ethnicity preference distribution across the two image reward datasets. Notably, both datasets exhibit a strong preference for the **White** group, 43.34% in HPDv2 and 40.08% in Pick-a-Pic. This imbalance in collected preferences risks might propagate representational bias during reward model training, ultimately reinforcing societal inequities in downstream video generation. Appendix D includes more analysis across 42 actions.

5.2 Image Reward Models

Extending our analysis of gender and ethnicity preference biases in human preference datasets, we examine how such biases propagate through reward models. Therefore, we construct an evaluation benchmark and use it to systematically examine the distribution of social biases in reward models.

Benchmark Construction. We create a evaluation benchmark based on text-to-image (T2I) generation, inspired by HPDv2 Wu et al. [2023] and ImageRewardDB

Settings	# of Prompts	Examples
Ethnicity + Person	294	An *East Asian* person is **baking** cookies in a cozy kitchen, with warm light and the aroma of vanilla filling the air.
Ethnicity + Gender	1176	An *East Asian* woman is **baking** cookies in a cozy kitchen, with warm light and the aroma of vanilla filling the air.

Table 3: Evaluation prompt statistics: Each example highlights the actor's *ethnicity* (if specified), the **action**, and the context.

Xu et al. [2024]. Leveraging the event prompting templates introduced in Section 3.2, we use FLUX Labs [2023], a state-of-the-art T2I model, to generate diverse image sets that systematically vary across gender, ethnicity, and action dimensions. The benchmark includes two evaluation settings: (1) `Ethnicity+Person`, where prompts specify only the actor's ethnicity, and (2) `Gender+Ethnicity`, where both gender and ethnicity are explicitly indicated. Table 3 provides statistics on prompt coverage and includes representative examples. To ensure statistical robustness and minimize variance, we generate 100 images per prompt. Figure 3 showcases sample outputs from the benchmark. To validate generation quality, three human annotators independently reviewed 100 randomly sampled images. Of these, 77 were unanimously deemed to be of sufficient quality to accurately represent the social attributes specified in the generation prompts.

Figure 3: Image examples of our constructed benchmark with generation prompts: "A/An [ethnicity][gender] is baking." We only show the images with gender $\in$ {man, woman}.

Preference Bias Evaluation. We evaluate *four* image reward models: (1) HPSv2.0 Wu et al. [2023], trained on the HPDv2 dataset; (2) HPSv2.1 Wu et al. [2023], trained on the unreleased HPDv2.1 dataset; (3) PickScore Kirstain et al. [2023b], developed using the Pick-a-Pic dataset; and (4) CLIP Radford et al. [2021], which serves as the base model for HPSv2.0, HPSv2.1, and PickScore prior to fine-tuning on their respective image reward datasets. Table 10 reports two complementary bias metrics, ethnicity-aware gender bias (PBS_G) and ethnic representation distribution (RDS_e and SDI). Appendix E includes more analysis across 42 actions.

Models	Average	White		Black		Latino		East Asian		Southeast Asian		India		Middle Eastern		Overall
	gray!7PBS$_G$	gray!7PBS$_G$	RDS	gray!7PBS$_G$	RDS	gray!7PBS$_G$	RDS	gray!7PBS$_G$	RDS	gray!7PBS$_G$	RDS	gray!7PBS$_G$	RDS	gray!7PBS$_G$	RDS	SDI
CLIP	gray!7-0.0726	gray!70.0343	**0.0182**	gray!7-0.1198	0.0002	gray!7-0.0934	-0.0013	gray!7-0.1315	0.0141	gray!7-0.0865	0.0094	gray!7-0.0508	-0.0299	gray!7-0.0607	-0.0108	**0.8495**
HPSv2.0	gray!70.6039	gray!70.6090	-0.0423	gray!70.7341	-0.0069	gray!70.6512	0.0237	gray!70.4752	-0.0031	gray!70.5192	-0.0100	gray!70.5922	0.0070	gray!70.6464	**0.0315**	0.8492
Δ	gray!7+0.6765	gray!7+0.5747	-0.0605	gray!7+0.8539	-0.0071	gray!7+0.7446	+0.0250	gray!7+0.6067	-0.0172	gray!7+0.6057	-0.0194	gray!7+0.6430	+0.0369	gray!7+0.7071	+0.0423	-0.0003
HPSv2.1	gray!7-0.0984	gray!7-0.0833	-0.0189	gray!70.0257	-0.0321	gray!7-0.0031	**0.0382**	gray!7-0.3044	0.0091	gray!7-0.2181	-0.0099	gray!7-0.0006	-0.0077	gray!7-0.1053	0.0214	0.8470
Δ	gray!7-0.0258	gray!7-0.1176	-0.0371	gray!7+0.1455	-0.0323	gray!7+0.0903	+0.0395	gray!7-0.1729	-0.0050	gray!7-0.1316	-0.0193	gray!7+0.0502	+0.0222	gray!7-0.0446	+0.0322	-0.0025
PickScore	gray!7-0.1157	gray!70.0321	0.0069	gray!7-0.0777	0.0279	gray!7-0.3479	-0.0118	gray!7-0.2257	**0.0316**	gray!7-0.2163	0.0115	gray!70.1531	-0.0391	gray!7-0.1277	-0.0271	0.8483
Δ	gray!7-0.0431	gray!7-0.0022	-0.0113	gray!7+0.0421	+0.0277	gray!7-0.2545	-0.0105	gray!7-0.0942	+0.0175	gray!7-0.1298	+0.0021	gray!7+0.2039	-0.0092	gray!7-0.0670	-0.0163	-0.0012

Table 4: Preference bias of reward models. All values represent *average* scores across 42 actions.

Ethnicity-Aware Gender Bias. We construct preference evaluation prompts in the format "A/An `[ethnicity]` person is `[action]`-ing `[context]`", covering all combinations of ethnicity and action. For each preference prompt, we generate images using generation prompts in the format "A/An `[ethnicity]` `[gender]` is `[action]`-ing `[context]`", where gender, ethnicity, and action are explicitly specified. The reward scores assigned to these images by a reward model are standardized using their mean and standard deviation. We then compute the average standardized score across the 100 images for each generation prompt. To compute the final PBS_G, we fix the ethnicity and action, and subtract the average standardized score for women from that for men. Because of the adaptation, PBS_G score here can be greater than one. A positive PBS_G score indicates a preference for men, while a negative score reflects a preference for women. CLIP shows a slight woman-preference bias (–0.0726). Fine-tuning on HPDv2 shifts HPSv2.0 toward a strong man-preference (+0.6039), consistent across ethnic groups. In contrast, PickScore (–0.1157) and HPSv2.1 (–0.0984) show woman-preference biases, with the latter's training data undisclosed. These shifts align with each model's training data, revealing consistent gender preferences across ethnicities.

Ethnicity Bias. We use preference evaluation prompts in the form "A person is `[action]`-ing `[context]`". For each preference prompt, we have generated images using more specific generation prompts of the form "A/An `[ethnicity]` person is `[action]`-ing `[context]`", where the ethnicity and action are explicitly specified. The reward scores for these images provided by a reward model

are standardized with mean and standard deviation. We then compute the average standardized score across the 100 images for each generation prompt. To calculate RDS_e and SDI, we fix the action and apply softmax function Bridle [1990], Bishop [2006] to normalize the scores for each ethnicity, indicating ethnicity preference within each action context. A positive RDS_e indicates overrepresentation of an ethnicity, while a negative score indicates underrepresentation. A higher SDI score corresponds to more balanced and diverse outputs across all groups. The base model, CLIP, slightly favors White individuals (RDS = 0.0182) and achieves the highest SDI score (0.8495), indicating relatively balanced ethnic representation. After fine-tuning, HPSv2.0 shifts toward Middle Eastern (RDS = 0.0315), HPSv2.1 toward Latino (RDS = 0.0382), and PickScore toward East Asian individuals (RDS = 0.0352). All show reduced SDI, indicating decreased ethnic diversity post-alignment.

6 Social Biases in Preference Alignment

Building on our analysis of gender and ethnicity biases in image reward models, we examine how preference alignment tuning affects bias in video generation. We fine-tune a Video Consistency Model distilled from VideoCrafter-V2 (VCM-VC2) Li et al. [2024] using three image-text reward models, HPSv2.0, HPSv2.1, and PickScore, and compare social bias distributions before and after tuning to assess how each reward model shapes identity representation. Following the T2V-Turbo-V1 training protocol Li et al. [2024], we incorporate reward feedback into the Latent Consistency Distillation process Luo et al. [2023] by using single step video generation. During student model distillation from a pretrained teacher text to video model, we directly optimize the decoded video frames to maximize reward scores from the image-text alignment models, guiding each frame toward representations more aligned with human preferences.

We evaluate aligned video diffusion models using our bias framework (§4). Table 11 reports two metrics: PBS_G for gender imbalance across ethnic groups, and RDS_e and SDI for ethnicity representation disparity and overall output diversity. Appendix F includes more analysis across 42 actions.

Models	Average	White		Black		Latino		East Asian		Southeast Asian		India		Middle Eastern		Overall
	gray!7PBS_G	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	SDI
VCM-VC2	gray!70.8034	gray!70.7925	**0.6405**	gray!70.7758	-0.2381	gray!70.8090	–	gray!70.7115	-0.2333	gray!70.7945	–	gray!70.8634	–	gray!70.8071	-0.1690	0.1433
+ HPSv2.0	gray!70.9116	gray!70.9667	0.3667	gray!70.9214	–	gray!70.9667	–	gray!70.8500	–	gray!70.9214	–	gray!70.9000	–	gray!70.8548	-0.3667	0.1257
Δ	gray!7+0.1082	gray!7+0.1742	-0.2738	gray!7+0.1456	–	gray!7+0.1577	–	gray!7+0.1385	–	gray!7+0.1269	–	gray!7+0.0366	–	gray!7+0.0477	-0.1977	-0.0176
+ HPSv2.1	gray!70.2267	gray!70.1321	**0.4286**	gray!70.2381	–	gray!70.3738	–	gray!70.1571	–	gray!70.3619	–	gray!70.1452	–	gray!70.1786	-0.4286	0.0976
Δ	gray!7-0.5767	gray!7-0.6604	-0.2119	gray!7-0.5377	–	gray!7-0.4352	–	gray!7-0.5544	–	gray!7-0.4326	–	gray!7-0.7182	–	gray!7-0.6285	-0.2596	-0.0457
+ PickScore	gray!70.3714	gray!70.3429	**0.6833**	gray!70.3357	-0.1810	gray!70.7190	-0.1929	gray!70.1450	-0.1952	gray!70.4548	–	gray!70.2500	–	gray!70.3515	-0.1143	0.1467
Δ	gray!7-0.4320	gray!7-0.4496	+0.0428	gray!7-0.4401	+0.0571	gray!7-0.0900	-0.1929	gray!7-0.5665	+0.0381	gray!7-0.3397	–	gray!7-0.6134	–	gray!7-0.4556	+0.0547	+0.0034

Table 5: Social biases of aligned models. All values represent *average* scores across 42 actions.

Ethnicity-Aware Gender Bias. To assess gender representation across ethnic groups, we prompt each model with "person" specified by an explicit ethnicity and compute PBS_G across 42 actions for each of the seven ethnic groups. A positive PBS_G score indicates a tendency to depict men more frequently, while a negative score suggests a preference for women. The base model, VCM-VC2, demonstrates a strong man bias across all ethnicities, which becomes more pronounced with alignment using HPSv2.0. In contrast, alignment with HPSv2.1 and PickScore significantly reduces PBS_G, indicating a shift toward more balanced or woman-preferred outputs. This change reflects the underlying woman bias present in the HPSv2.1 and PickScore reward models, which steer the model away from the man-dominant bias of the base model.

Ethnicity Bias. Ethnic representation is evaluated by prompting each model with "person" while omitting explicit ethnicity, and calculating RDS_e for each group. Positive values indicate overrepresentation, and negative values indicate underrepresentation. Overall demographic balance is measured using SDI, where higher values reflect more equitable representation. The base model, VCM-VC2, strongly favors White individuals (RDS = 0.6405), while Black, East Asian, and Middle Eastern groups are underrepresented. Alignment with HPSv2.1 reduces some disparities by improving balance for White and Black groups, but significantly decreases Latino representation (RDS = –0.4352) and lowers SDI, indicating reduced diversity. In contrast, PickScore achieves the highest SDI and produces more balanced representation across most ethnic groups, resulting in the most demographically equitable outputs.

7 Controllable Preference Modeling for Video Diffusion Models

Building on prior findings, we observe that reward models trained on imbalanced image preference datasets inherit and amplify social biases. These biases are then reflected in video diffusion models fine-tuned with such reward signals, often leading to unbalanced outputs. In this section, we explore

whether manipulating the distribution of social attributes in image datasets allows for controllable bias in reward models, enabling video models to produce more equitable outputs Sheng et al. [2020].

7.1 Image Reward Dataset Construction

We construct two reward datasets: a man-preferred version and a woman-preferred version, using images from §5.1 to guide diffusion models toward gender-specific representations. Each dataset includes 2.94 million preference pairs from the `Gender+Ethnicity` set, where each pair depicts the same action and ethnicity but differs by gender (*e.g.*, M-1 vs. W-1 in Figure 3). Prompts follow the format "A/An `[ethnicity]` person is `[action]`-ing `[context]`." In the man-preferred dataset, male images are labeled 1 and female images 0; the opposite applies in the woman-preferred dataset. To enhance face-free diversity, we also include 537,660 additional image pairs from HPDv2. When applied to a base model with man-preference bias, the woman-preferred dataset helps correct this imbalance and promotes more equitable gender representation.

7.2 Image Reward Model Development & Alignment Tuning

Leveraging the man-preferred and woman-preferred image datasets introduced in §7.1, we fine-tune two reward models on top of a pre-trained CLIP vision encoder: the Man-Preferred Reward Model (RM_M) and the Woman-Preferred Reward Model (RM_W). Each model is trained to reflect gender-specific preferences based on its respective dataset. As shown in Table 12, RM_M consistently assigns greater PBS_G scores across all demographic groups, indicating a strong alignment with man-preferred representations. In contrast, RM_W exhibits an opposite trend, systematically favoring woman-preferred content. The clear divergence between these models highlights the effectiveness of reward tuning in capturing and reinforcing gendered preferences.

Models	Average	White	Black	Latino	East Asian	Southeast Asian	India	Middle Eastern
CLIP	-0.0726	0.0343	-0.1198	-0.0934	-0.1315	-0.0865	-0.0508	-0.0607
RM_M	1.5280$_{+1.60}$	1.6300$_{+1.60}$	1.5752$_{+1.70}$	1.5524$_{+1.65}$	1.4323$_{+1.56}$	1.4525$_{+1.54}$	1.5619$_{+1.61}$	1.4914$_{+1.55}$
RM_W	-0.7448$_{-2.27}$	-0.6318$_{-2.26}$	-0.7943$_{-2.37}$	-0.8279$_{-2.38}$	-0.6282$_{-2.06}$	-0.6429$_{-2.10}$	-0.8846$_{-2.45}$	-0.8042$_{-2.30}$

Table 6: Preference bias of reward models. All values represent *average* scores across 42 actions.

Building on our earlier reward model training, we applied RM_M and RM_W to guide alignment tuning of a base video diffusion model using the same preference-driven training strategy. These reward signals enabled the generation of two distinct variants: one aligned with man-preferred content and the other with woman-preferred content. As shown in Table 13, alignment with RM_M led to consistently greater PBS_G scores across all demographic groups, reinforcing man-preference bias. Conversely, alignment with RM_W resulted in substantially smaller scores, indicating a strong shift toward woman-preference bias. These results confirm that our controllable preference modeling approach can effectively modulate gender bias in video generation, offering a flexible mechanism to either amplify or reduce specific social tendencies in model outputs.

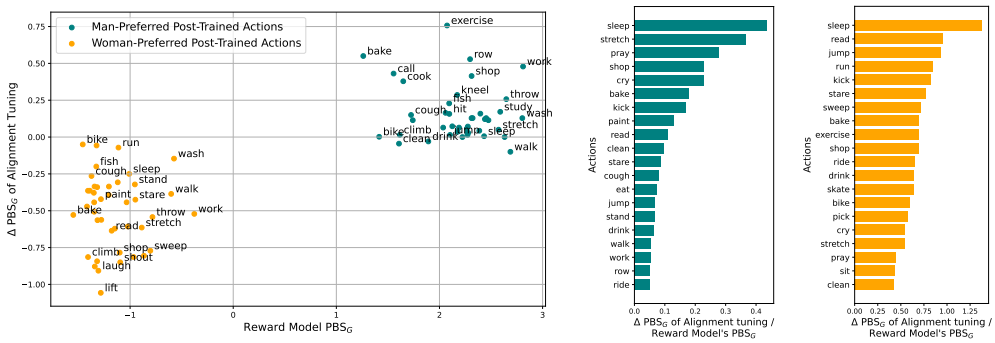
Models	Average	White	Black	Latino	East Asian	Southeast Asian	India	Middle Eastern
VCM-VC2	0.8034	0.7925	0.7758	0.8690	0.7115	0.7945	0.8634	0.8071
+ RM_M	0.9584$_{+0.16}$	0.9595$_{+0.17}$	0.9524$_{+0.18}$	0.9756$_{+0.11}$	0.9437$_{+0.23}$	0.9447$_{+0.15}$	0.9640$_{+0.10}$	0.9709$_{+0.16}$
+ RM_W	0.3082$_{-0.50}$	0.3341$_{-0.46}$	0.3913$_{-0.38}$	0.3314$_{-0.54}$	0.1008$_{-0.61}$	0.2639$_{-0.53}$	0.3446$_{-0.52}$	0.3894$_{-0.42}$

Table 7: Social biases of aligned models. All values represent *average* scores across 42 actions.

Which Actions Are Most Sensitive During Alignment Tuning? We analyze how different actions respond to gender-specific reward model tuning using PBS_G scores before and after alignment tuning (§7). In Figure 4a, we observe that actions like *exercise*, *row*, and *cook* become more biased toward men after RM_M tuning, while actions like *bake*, *sleep*, and *sweep* show strong shifts toward women after RM_W tuning. Figures Figure 4b and Figure 4c rank actions by sensitivity (*i.e.*, ΔPBS_G normalized by reward models' PBS_G), revealing that socially gendered actions, such as sleep, stretch, and read, are especially susceptible to alignment tuning bias shifts. These results underscore the effectiveness of our proposed event-centric evaluation framework in capturing fine-grained, action-specific shifts in gender bias during alignment tuning.

8 Conclusion

In summary, this work identifies and addresses critical blind spots in evaluating social bias within text-to-video generation. By introducing VIDEOBIASEVAL, we provide a structured framework that disentangles identity attributes from content semantics and captures how alignment tuning reshapes social representations. Our analysis reveals that reward-model-based alignment not only inherits but often amplifies existing biases in human preference data. These findings underscore the necessity of



(a) Δ PBS$_G$ of video generation model before and after alignment tuning by RM$_\text{M}$ and RM$_\text{W}$. Results are broken down into actions. Figure 4b and Figure 4c are based on this figure.

(b) Sensitive actions in man-preferred alignment tuning.

(c) Sensitive actions in woman-preferred alignment tuning.

Figure 4: Action-level impact of alignment tuning guided by RM$_\text{M}$ and RM$_\text{W}$.

incorporating bias auditing and mitigation at every stage of the video generation pipeline, paving the way toward more equitable and socially aware generative systems.

References

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

Alexander Black, Simon Jenni, Tu Bui, Md. Mehrab Tanjim, Stefano Petrangeli, Ritwik Sinha, Viswanathan Swaminathan, and John Collomosse. Vader: Video alignment differencing and retrieval, 2023. URL `https://arxiv.org/abs/2303.13193`.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

John S. Bridle. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In *Advances in Neural Information Processing Systems*, volume 2, pages 211–217. Morgan Kaufmann, 1990.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024a. URL `https://arxiv.org/abs/2401.09047`.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3043–3054, 2023.

Joseph Cho, Samuel Schmidgall, Cyril Zakka, Mrudang Mathur, Dhamanpreet Kaur, Rohan Shad, and William Hiesinger. Surgen: Text-guided diffusion model for surgical video generation. *arXiv preprint arXiv:2408.14028*, 2024.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 259–268. ACM, 2015.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16): E3635–E3644, 2018.

Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11975–11985, 2024.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. 2023a.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023b.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2023.

Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *arXiv preprint arXiv:2405.18750*, 2024.

Runtao Liu, Haoyu Wu, Zheng Ziqiang, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. *arXiv preprint arXiv:2412.14167*, 2024a.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024b.

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023. URL `https://arxiv.org/abs/2310.04378`.

Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, Yu Zhou, Deshan Sun, Deyu Zhou, Jian Zhou, Kaijun Tan, Kang An, Mei Chen, Wei Ji, Qiling Wu, Wen Sun, Xin Han, Yanan Wei, Zheng Ge, Aojie Li, Bin Wang, Bizhu Huang, Bo Wang, Brian Li, Changxing Miao, Chen Xu, Chenfei Wu, Chenguang Yu, Dapeng Shi, Dingyuan Hu, Enle Liu, Gang Yu, Ge Yang, Guanzhe Huang, Gulin Yan, Haiyang Feng, Hao Nie, Haonan Jia, Hanpeng Hu, Hanqi Chen, Haolong Yan, Heng Wang, Hongcheng Guo, Huilin Xiong, Huixin Xiong, Jiahao Gong, Jianchang Wu, Jiaoren Wu, Jie Wu, Jie Yang, Jiashuai Liu, Jiashuo Li, Jingyang Zhang, Junjing Guo, Junzhe Lin, Kaixiang Li, Lei Liu, Lei Xia, Liang Zhao, Liguo Tan, Liwen Huang, Liying Shi, Ming Li, Mingliang Li, Muhua Cheng, Na Wang, Qiaohui Chen, Qinglin He, Qiuyan Liang, Quan Sun, Ran Sun, Rui Wang, Shaoliang

Pang, Shiliang Yang, Sitong Liu, Siqi Liu, Shuli Gao, Tiancheng Cao, Tianyu Wang, Weipeng Ming, Wenqing He, Xu Zhao, Xuelin Zhang, Xianfang Zeng, Xiaojia Liu, Xuan Yang, Yaqi Dai, Yanbo Yu, Yang Li, Yineng Deng, Yingming Wang, Yilei Wang, Yuanwei Lu, Yu Chen, Yu Luo, Yuchu Luo, Yuhe Yin, Yuheng Feng, Yuxiang Yang, Zecheng Tang, Zekai Zhang, Zidong Yang, Binxing Jiao, Jiansheng Chen, Jing Li, Shuchang Zhou, Xiangyu Zhang, Xinhao Zhang, Yibo Zhu, Heung-Yeung Shum, and Daxin Jiang. Step-video-t2v technical report: The practice, challenges, and future of video foundation model, 2025. URL `https://arxiv.org/abs/2502.10248`.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Elijah Miller, Thomas Dupont, and Mingming Wang. Enhanced creativity and ideation through stable video synthesis. *arXiv preprint arXiv:2405.13357*, 2024.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.

Haoyi Qiu, Zi-Yi Dou, Tianlu Wang, Asli Celikyilmaz, and Nanyun Peng. Gender biases in automatic evaluation metrics for image captioning. *arXiv preprint arXiv:2305.14711*, 2023.

Haoyi Qiu, Wenbo Hu, Zi-Yi Dou, and Nanyun Peng. Valor-eval: Holistic coverage and faithfulness evaluation of large vision-language models. *arXiv preprint arXiv:2404.13874*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. Finetuning text-to-image diffusion models for fairness. *arXiv preprint arXiv:2311.07604*, 2023.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. Towards controllable biases in language generation. *arXiv preprint arXiv:2005.00268*, 2020.

Edward H Simpson. Measurement of diversity. *Nature*, 163(4148):688–688, 1949.

Jiao Sun and Nanyun Peng. Men are elected, women are married: Events gender bias on Wikipedia. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 350–360, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.45. URL `https://aclanthology.org/2021.acl-short.45/`.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. *arXiv preprint arXiv:2407.14505*, 2024.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023b.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024b.

Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. In *European Conference on Computer Vision*, pages 207–224. Springer, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: instructing video diffusion models with human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6463–6474, 2024.

Mike Zajko. Conservative ai and social inequality: conceptualizing alternatives to bias through social theory. *Ai & Society*, 36(3):1047–1056, 2021.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

A Related Work

T2V Evaluation. Recent evaluation benchmarks such as VBench Huang et al. [2024], EvalCrafter Liu et al. [2024b], and T2V-CompBenchSun et al. [2024] evaluate text-to-video models using metrics like Fréchet Video Distance Unterthiner et al. [2019], CLIP-Score Hessel et al. [2021], and object consistency, yet they overlook who is depicted and how identities are portrayed. GRiT-based metrics Wu et al. [2025] may verify that a "doctor" appears, but fail to flag when all doctors are white men. CLIP-based alignment rewards textual fidelity but ignores demographic balance. To ensure fair and trustworthy evaluation, T2V benchmarks must move beyond surface-level metrics and explicitly audit the distribution of social attributes across outputs. Our work meets this need by introducing an event-centric framework that quantifies gender and ethnicity-aware biases throughout the entire T2V generation pipeline.

Bias Evaluation in Generative Models. Most existing studies on social bias in text-to-image or language generation focus on static, single-frame outputs such as portraits or isolated object scenes. Approaches like StableBias Luccioni et al. [2023], DALL-Eval Cho et al. [2023], and SocialCounterfactuals Howard et al. [2024] primarily tally identity frequencies but seldom examine what those identities are portrayed *doing*. Even recent benchmarks that track demographic representation often evaluate each image independently, which conceals recurring patterns such as the tendency to depict men in authoritative roles and women in supportive ones. By neglecting to analyze actors, actions, and context jointly, these evaluations fail to capture role-specific stereotypes and cannot reveal bias in narrative or temporal settings. We address this limitation by auditing at the event level, disentangling actor attributes from actions and environments to uncover how social representation shifts across different scenarios.

B VIDEOBIASEVAL

B.1 Event Definition

We investigate whether video generation models exhibit social biases in their portrayal of *events*, particularly in how different actors are depicted performing actions within these events Sun and Peng [2021]. Such biases often manifest as imbalanced *gender* portrayals or the disproportionate representation of certain *ethnic groups* Zajko [2021], potentially reinforcing stereotypes and compromising fairness Bolukbasi et al. [2016]. To systematically analyze these patterns, we represent each event as a tuple $\langle p, a, c \rangle$: an actor p performing action a in context c. Building on prior work, we target socially associated actions to examine identity representation Zhao et al. [2017], Garg et al. [2018], Cho et al. [2023], Qiu et al. [2023]. Unlike existing benchmarks, our approach captures social dynamics in event generation through a controllable fairness evaluation framework.

Actors. We depict each actor (p) with gender and ethnicity attributes to facilitate our analysis of social bias. For gender, we employ the *four* categories proposed by Luccioni et al. [2023]: man, woman, the neutral term "person," and non-binary person. Though inclusive, this schema remains limited in capturing the full spectrum of gender identities. For ethnicity, we use *seven* groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino, following Karkkainen and Joo [2021] and U.S. Census Bureau categories. While these categories aim to be inclusive, they are socially constructed and not intended to be exhaustive or universally representative.

Actions. We select **42** actions (a): bake, bike, call, clean, climb, cook, cough, cry, drink, drive, eat, exercise, fish, hit, jump, kick, kneel, laugh, lift, paint, pick, pitch, pray, read, ride, row, run, shop, shout, sit, skate, sleep, smile, stand, stare, stretch, study, sweep, throw, walk, wash, work, identified in previous studies as statistically correlated with specific genders or ethnic groups Zhao et al. [2017], Garg et al. [2018], Cho et al. [2023], Qiu et al. [2023]. These actions, exhibiting a stronger correlation with gender or ethnicity than random in the studied corpus, provide a valuable testbed for examining how such biases are represented in relation to the individuals involved.

C Social Biases in Video Generative Models

To demonstrate the utility of our framework, we apply it to *four* state-of-the-art video diffusion models with varying alignment strategies. The **aligned** models include InstructVideo Yuan et al. [2024], which is based on ModelScope Wang et al. [2023a] and aligned with HPSv2.0, and T2V-Turbo-V1 Li et al. [2024], which builds on VideoCrafter-2 Chen et al. [2024a] and is aligned with HPSv2.1, InternVid2-S2 Wang et al. [2024b], and ViCLIP Wang et al. [2023b]. Their **unaligned** counterparts, ModelScope and VideoCrafter-2, serve as baselines for controlled comparisons. For implementation, we use the official code repositories provided by the respective papers and run inference on 1 to 8 NVIDIA A100 80GB GPUs.

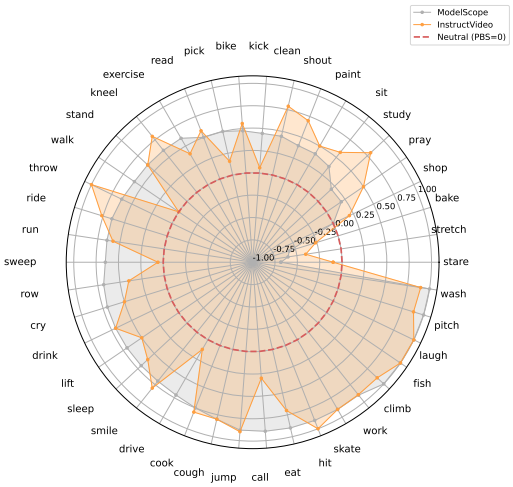
To compute the social bias distribution, as outlined in §3, we generate videos for each prompt using each model *ten* times with different random seeds. The final results are obtained by averaging the outcomes across these generations to account for variability introduced by stochastic sampling. Table 8 reports two complementary bias metrics, ethnicity-aware gender bias (PBS_G) and ethnic representation distribution (RDS_e and SDI).

Models	Average	White		Black		Latino		East Asian		Southeast Asian		India		Middle Eastern		Overall
	gray!7PBS_G	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	SDI
ModelScope (u)	gray!70.4815	gray!70.5683	**0.7690**	gray!70.3912	-0.1952	gray!70.6308	-0.1952	gray!70.4406	-0.1810	gray!70.4611	-	gray!70.3938	-	gray!70.4833	-0.1976	0.0538
InstructVideo (a)	gray!70.5295	gray!70.5584	**0.7833**	gray!70.5114	-0.1976	gray!70.6729	-0.1929	gray!70.4282	-0.1976	gray!70.5020	-	gray!70.4878	-	gray!70.5393	-0.1952	0.0267
Δ	gray!7+0.0480	gray!7-0.0099	+0.0143	gray!7+0.1202	-0.0024	gray!7+0.0421	+0.0023	gray!7-0.0124	-0.0166	gray!7+0.0409	-	gray!7+0.0940	-	gray!7+0.0560	+0.0024	-0.0271
Video-Crafter-V2 (u)	gray!70.7581	gray!70.7485	**0.6905**	gray!70.6167	-0.1905	gray!70.8599	-0.1952	gray!70.6976	-0.1500	gray!70.8272	-	gray!70.8032	-	gray!70.7560	-0.1548	**0.1252**
T2V-Turbo-V1 (a)	gray!70.8306	gray!70.8713	**0.6381**	gray!70.8095	-	gray!70.8599	-0.2476	gray!70.7762	-0.2426	gray!70.8929	-	gray!70.7762	-	gray!70.7664	-0.1476	0.1119
Δ	gray!7+0.0725	gray!7+0.1228	-0.0524	gray!7+0.1928	-	gray!70.0000	-0.0524	gray!7+0.0786	-0.0926	gray!7+0.0657	-	gray!7-0.0270	-	gray!7+0.0104	+0.0072	-0.0133

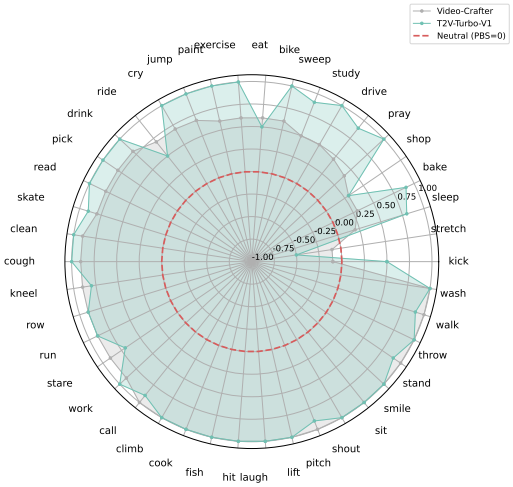
Table 8: Distributions of social attributes in two pairs of unaligned (u) and aligned (a) video diffusion models. Each value represents the average score computed across 42 actions. A positive PBS_G score indicates a bias toward generating man characters (man-preference), while a negative score indicates a bias toward woman characters (woman-preference); values close to zero suggest balanced gender representation. We annotate man-preference with (+) and woman-preference with (–). For RDS_e, a positive score reflects the overrepresentation of a specific ethnicity, while a negative score reflects underrepresentation; these are marked with (+) and (–), respectively. Finally, a higher SDI score indicates greater balance and diversity in ethnic representation across the generated outputs.

Ethnicity-Aware Gender Bias. We prompt each model with "person" alongside an explicit ethnicity and compute PBS_G for each of seven ethnic groups across 42 actions. A positive PBS_G indicates a tendency to depict men more often than women, while a negative value indicates the reverse. All models exhibit a man bias, with average PBS_G values exceeding zero. This bias persists across all ethnic groups. The Δ rows show how alignment-tuned models shift relative to their baselines: InstructVideo's average PBS_G increases by 0.04 and T2V-Turbo-V1's by 0.0725, suggesting that preference-based alignment tuning can inadvertently amplify gender imbalance. Figures 11 and 26 to 31 presents the PBS_G scores across 42 actions for each ethnicity group.

Ethnicity Bias. We prompt with "person" alone (no ethnicity) and record each model's over- or under-representation of the seven groups via RDS_e and its overall diversity via SDI. A positive RDS_e score signifies overrepresentation of a specific ethnicity group, while a negative score indicates underrepresentation. A higher SDI score denotes more balanced and diverse outputs across ethnic groups. ModelScope shows strong White overrepresentation (RDS_{White} = 0.769, SDI = 0.0538). Alignment-tuned InstructVideo exaggerates that effect (RDS_{White} = 0.783, SDI = 0.0267). VideoCrafter-2 is somewhat more balanced (RDS_{White} = 0.688, SDI = 0.126), while T2V-Turbo-V1 reduces White overrepresentation further (RDS_{White} = 0.555) but also lowers overall diversity (SDI = 0.109). Thus, although alignment tuning can mitigate certain ethnic skews, it may also reduce demographic diversity. Figure 12 show the ethnicity bias across 42 actions.

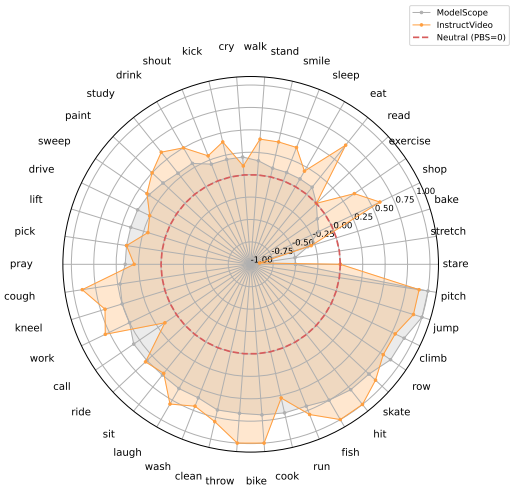


(a) ModelScope vs. InstructVideo

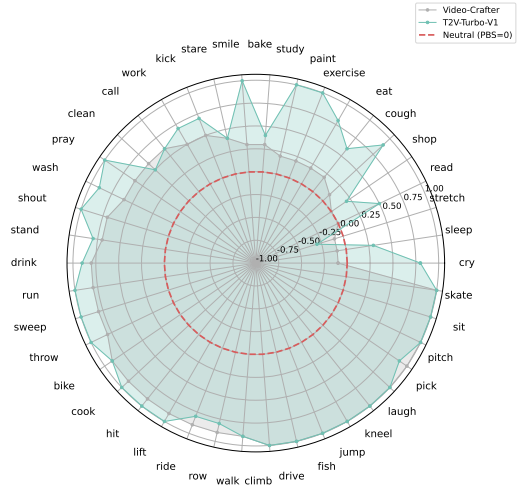


(b) Video-Crafter-2 vs. T2V-Turbo-V1

Figure 5: Ethnicity-aware gender bias (White).

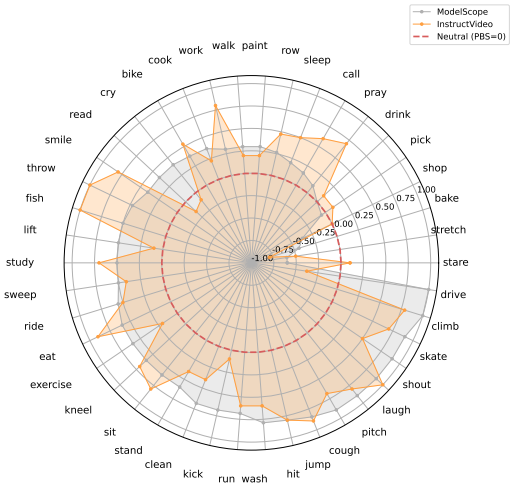


(a) ModelScope vs. InstructVideo

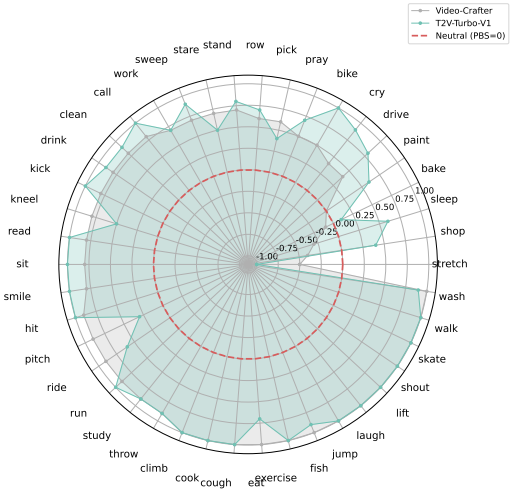


(b) Video-Crafter-2 vs. T2V-Turbo-V1

Figure 6: Ethnicity-aware gender bias (Black).

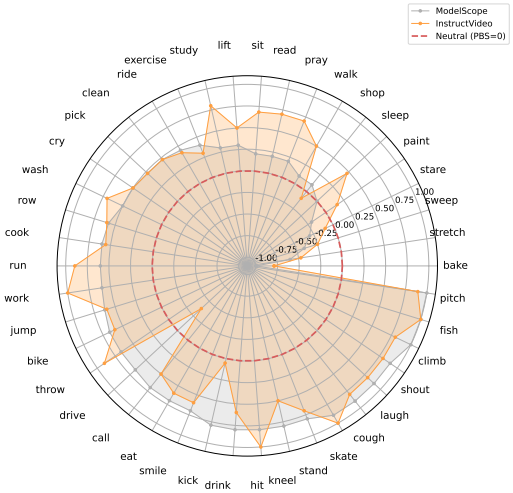


(a) ModelScope vs. InstructVideo (orange)

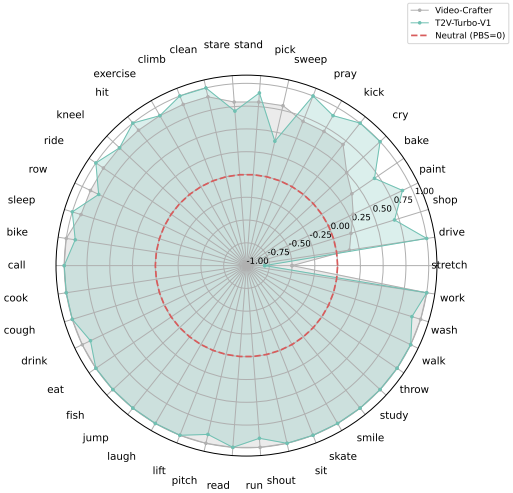


(b) Video-Crafter-2 vs. T2V-Turbo-V1

Figure 7: Ethnicity-aware gender bias (East Asian).

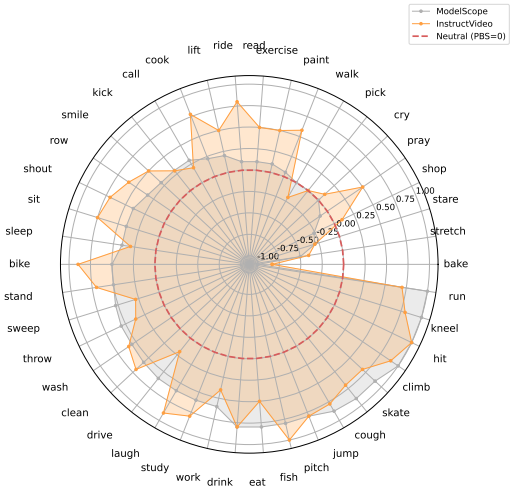


(a) ModelScope vs. InstructVideo

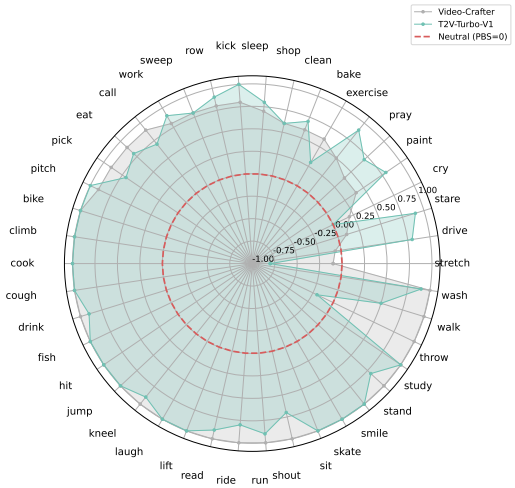


(b) Video-Crafter-2 vs. T2V-Turbo-V1

Figure 8: Ethnicity-aware gender bias (Southeast Asian).

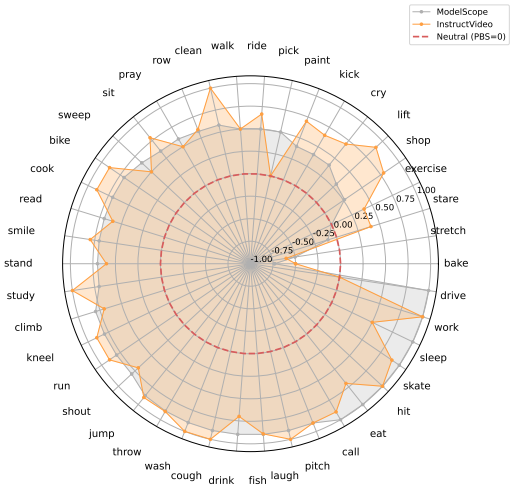


(a) ModelScope vs. InstructVideo

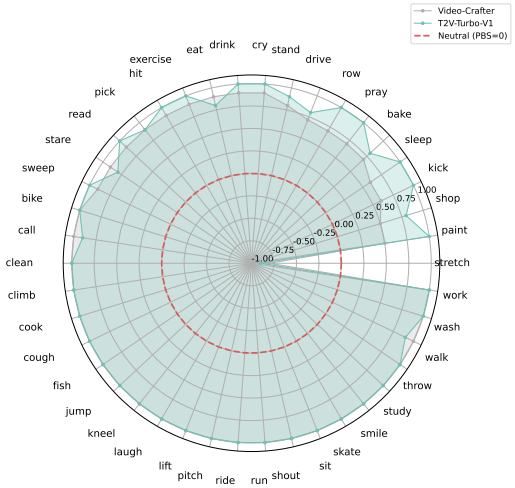


(b) Video-Crafter-2 vs. T2V-Turbo-V1

Figure 9: Ethnicity-aware gender bias (Indian).

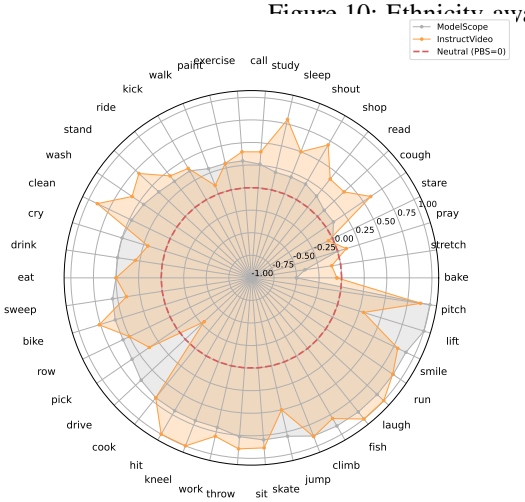


(a) ModelScope vs. InstructVideo

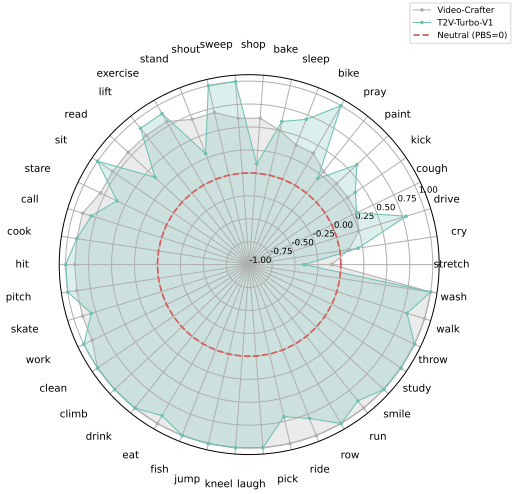


(b) Video-Crafter-2 vs. T2V-Turbo-V1

Figure 10: Ethnicity-aw



(a) ModelScope vs. InstructVideo



(b) Video-Crafter-2 vs. T2V-Turbo-V1

Figure 11: Ethnicity-aware gender bias (Middle Eastern).

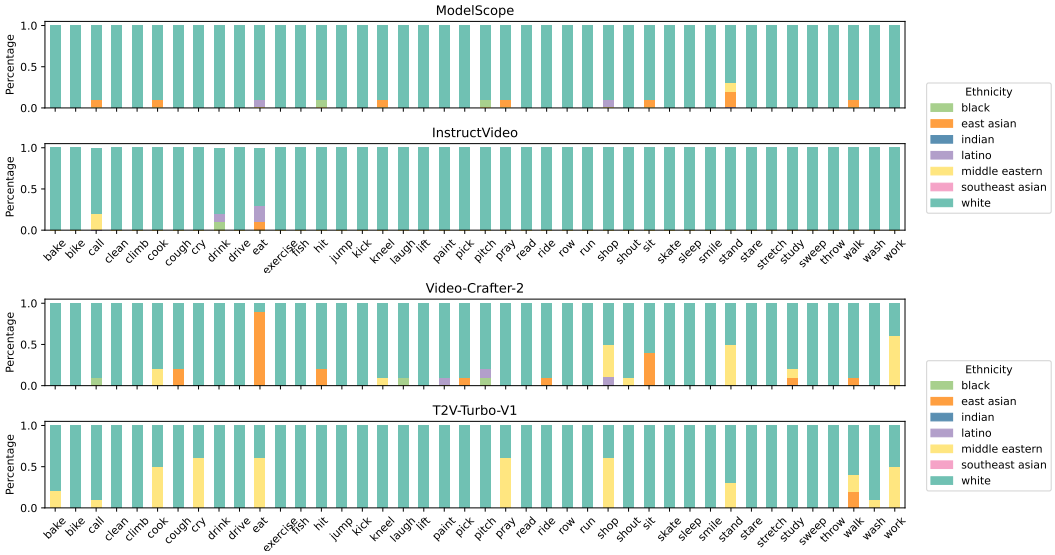


Figure 12: Ethnicity bias distribution.

D Social Biases in Image Reward Datasets

We analyze *two* widely used image reward datasets to investigate preference biases: HPDv2 Wu et al. [2023] and Pick-a-Pic Kirstain et al. [2023b]. For each dataset, we extract gender, ethnicity, and action attributes from image captions using GPT-4o-mini, and classify attributes from images using three VLMs (Qwen2-VL-7B, Qwen2.5-VL-7B, InternVL2.5-8B). We then aggregate the social attributes from both caption and image modalities, retaining only instances featuring one of our predefined actions. After processing, HPDv2 contains 28,783 validated (images, caption, preference) tuples covering 29 actions, and Pick-a-Pic contains 14,958 across 19 actions. Each tuple presents two images, with a human annotator selecting the one that best matches the caption. To assess potential preference biases, we measure how often annotators *prefer* specific gender or ethnicity representations for given actions.

Figure 13 shows the gender preference bias in the two datasets. Values greater than zero (outside the red circle) indicate a man-preferred bias, while values less than zero (inside the red circle) indicate a woman-preferred bias. Points on the red circle represent more neutral preference. In **HPDv2**, 62.07% (18/29) of actions show a preference for men, while only 24.14% (7/29) favor women, indicating a skew toward **man-preferred** representations. In contrast, **Pick-a-Pic** reveals a **woman-preferred** tendency, with 57.89% (11/19) of actions biased toward women and 26.32% (5/19) toward men. These patterns highlight that both datasets exhibit non-neutral gender preferences, though in opposing directions, potentially shaping downstream alignment in different ways.

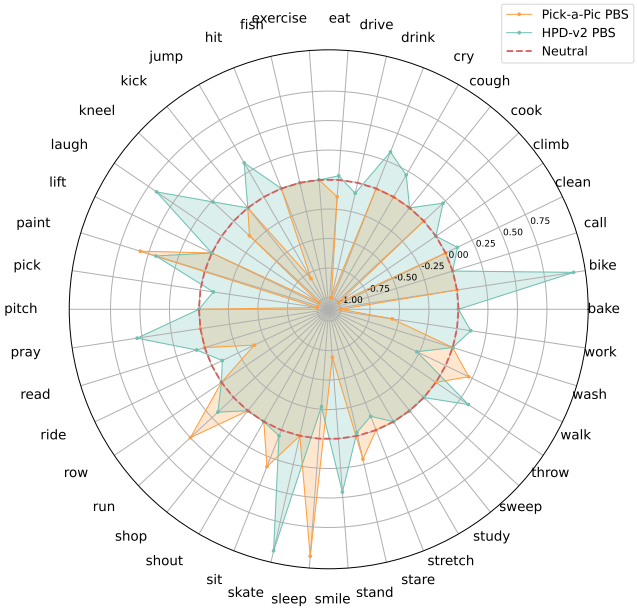


Figure 13: Image reward datasets gender preference distribution.

Table 9 presents the ethnicity preference distribution across the two image reward datasets, while Figure 14 provides a fine-grained breakdown across 42 actions. Notably, both datasets exhibit a strong preference for the **White** group, 43.34% in HPDv2 and 40.08% in Pick-a-Pic, followed by East Asian and Indian representations. Despite certain actions showing distinct preferences (*e.g.*, "bake" favoring Black individuals and "fish" favoring East Asians), the overall distributions reveal a pronounced imbalance skewed toward White representations. This suggests that the reward signals used to guide image generation may reflect and reinforce ethnic biases embedded in the datasets. This imbalance in collected preferences risks might propagate representational bias during reward model training, ultimately reinforcing societal inequities in downstream video generation. These findings underscore the urgent need for more inclusive and representative datasets that reflect global demographic diversity in both identity and activity contexts.

Datasets	White	Black	Latino	East			
Asian	Southeast						
Asian	India	Middle					
Eastern							
HPDv2	43.34	9.16	4.44	19.38	1.39	20.20	2.09
Pick-a-Pic	40.08	15.36	8.51	19.94	0.20	13.34	2.56

Table 9: Ethnicity distribution across reward datasets (in %).

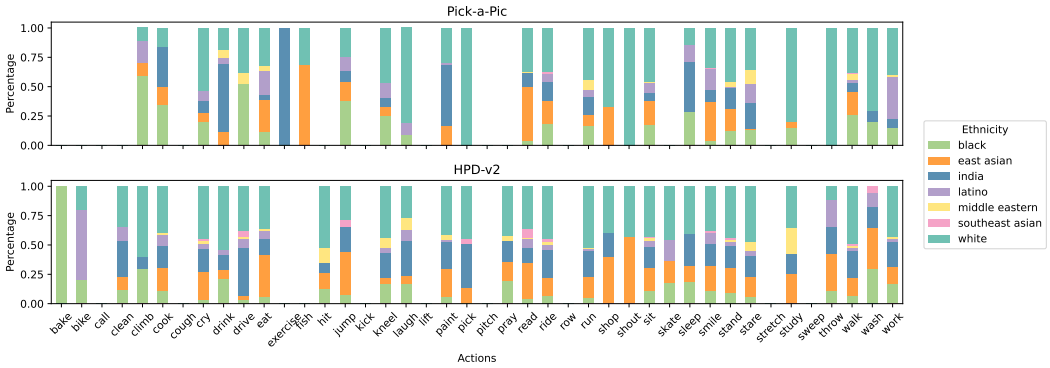


Figure 14: Ethnicity preference distribution across 42 actions.

E Social Biases in Image Reward Models

Preference Bias Evaluation. We evaluate four image reward models: (1) HPSv2.0 Wu et al. [2023], trained on the HPDv2 dataset; (2) HPSv2.1 Wu et al. [2023], trained on the unreleased HPDv2.1 dataset; (3) PickScore Kirstain et al. [2023b], developed using the Pick-a-Pic dataset; and (4) CLIP Radford et al. [2021], which serves as the base model for HPSv2.0, HPSv2.1, and PickScore prior to fine-tuning on their respective image reward datasets. Table 10 reports two complementary bias metrics, ethnicity-aware gender bias (PBS_G) and ethnic representation distribution (RDS_e and SDI).

Models	Average	White		Black		Latino		East Asian		Southeast Asian		India		Middle Eastern		Overall
	gray!7PBS_G	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	SDI
CLIP	gray!7-0.0726	gray!70.0343	**0.0182**	gray!7-0.1198	0.0002	gray!7-0.0934	-0.0013	gray!7-0.1315	0.0141	gray!7-0.0865	0.0094	gray!7-0.0508	-0.0299	gray!7-0.0607	-0.0108	**0.8495**
HPSv2.0	gray!70.6039	gray!70.6090	-0.0423	gray!70.7341	-0.0069	gray!70.6512	0.0237	gray!70.4752	-0.0031	gray!70.5192	-0.0100	gray!70.5922	0.0070	gray!70.6464	**0.0315**	0.8492
Δ	gray!7+0.6765	gray!7+0.5747	-0.0605	gray!7+0.8539	-0.0071	gray!7+0.7446	+0.0250	gray!7+0.6067	-0.0172	gray!7+0.6057	-0.0194	gray!7+0.6430	+0.0369	gray!7+0.7071	+0.0423	-0.0003
HPSv2.1	gray!7-0.0984	gray!7-0.0833	-0.0189	gray!70.0257	-0.0321	gray!7-0.0031	**0.0382**	gray!7-0.3044	0.0091	gray!7-0.2181	-0.0099	gray!7-0.0006	-0.0077	gray!7-0.1053	0.0214	0.8470
Δ	gray!7-0.0258	gray!7-0.1176	-0.0371	gray!7+0.1455	-0.0323	gray!7+0.0903	+0.0395	gray!7-0.1729	-0.0050	gray!7-0.1316	-0.0193	gray!7+0.0502	+0.0222	gray!7-0.0446	+0.0322	-0.0025
PickScore	gray!7-0.1157	gray!70.0321	0.0069	gray!7-0.0777	0.0279	gray!7-0.3479	-0.0118	gray!7-0.2257	**0.0316**	gray!7-0.2163	0.0115	gray!70.1531	-0.0391	gray!7-0.1277	-0.0271	0.8483
Δ	gray!7-0.0431	gray!7-0.0022	-0.0113	gray!7+0.0421	+0.0277	gray!7-0.2545	-0.0105	gray!7-0.0942	+0.0175	gray!7-0.1298	+0.0021	gray!7+0.2039	-0.0092	gray!7-0.0670	-0.0163	-0.0012

Table 10: Preference bias of reward models. All values represent *average* scores across 42 actions.

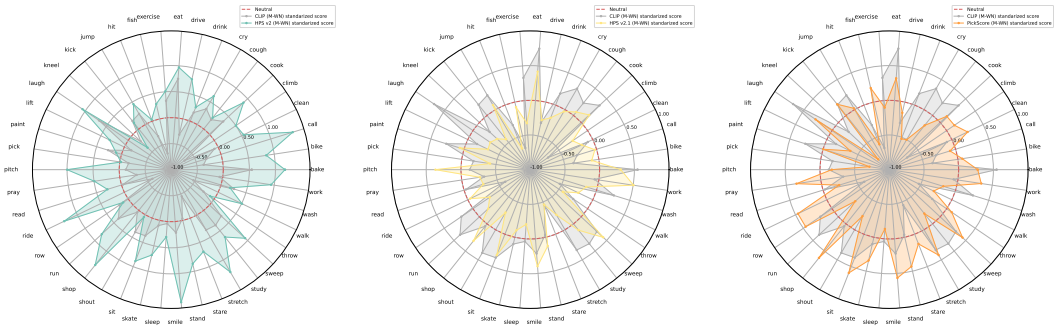
Ethnicity-Aware Gender Bias. We construct preference evaluation prompts in the format "A/An `[ethnicity]` person is `[action]`-ing `[context]`", covering all combinations of ethnicity and action, resulting in $|E| \times |A|$ evaluation prompts. For each preference prompt, we generate images using generation prompts in the format "A/An `[ethnicity]` `[gender]` is `[action]`-ing `[context]`", where gender, ethnicity, and action are explicitly specified. This yields a total of $|G| \times |E| \times |A| \times 100$ images. The reward scores assigned to these images by a reward model are standardized using their mean and standard deviation. We then compute the average standardized score across the 100 images for each generation prompt, resulting in $|G| \times |E| \times |A|$ mean scores. To compute the final PBS_G, we fix the ethnicity and action, and subtract the average standardized score for women from that for men, producing $|E| \times |A|$ PBS_G values.

A positive PBS_G score indicates a preference for men, while a negative score reflects a preference for women. The base reward model, CLIP, exhibits a mild woman-preference bias overall (PBS_G = –0.0726). After fine-tuning on HPDv2, HPSv2.0 reverses this trend and demonstrates a notable shift toward man-preference bias (+0.6039), consistent across most ethnic groups. In contrast, PickScore shows a stronger woman-preference bias (PBS_G = –0.1157), aligning with the characteristics of Pick-a-Pic. HPSv2.1 also exhibits a woman-preference trend (PBS_G = –0.0984), though its training data has not been publicly disclosed. These directional shifts are evident across all ethnic groups,

suggesting that model fine-tuning introduces consistent and dataset-aligned gender preferences. Figures 15 to 21 presents the PBS_G scores across 42 actions for each ethnicity group.

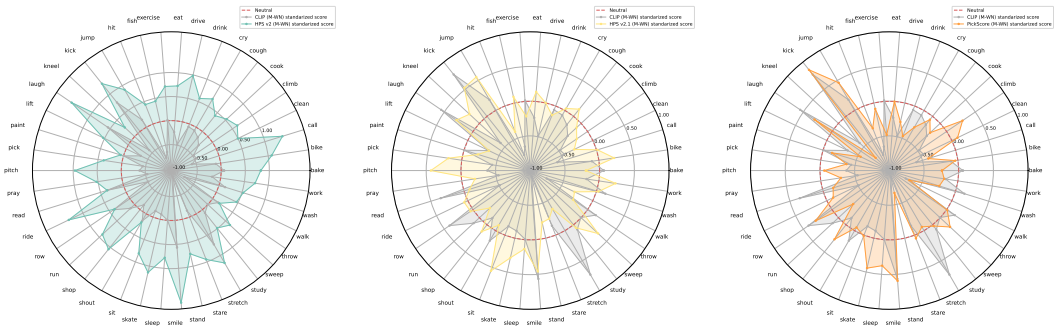
Ethnicity Bias. We use preference evaluation prompts in the form "A person is `[action]`-ing `[context]`", covering all actions and resulting in $|A|$ evaluation prompts. For each preference prompt, we have generated images using more specific generation prompts of the form "A/An `[ethnicity]` person is `[action]`-ing `[context]`", where the ethnicity and action are explicitly specified. For each such combination, we have a total of $|E| \times |A| \times 100$ images. The reward scores for these images provided by a reward model are standardized with mean and standard deviation. We then compute the average standardized score across the 100 images for each generation prompt, leading to $|E| \times |A|$ mean scores. To calculate RDS_e and SDI, we fix the action and apply softmax function Bridle [1990], Bishop [2006] to normalize the scores for each ethnicity. This results in $|E| \times |A|$ final RDS_e scores and $|A|$ SDI scores, indicating ethnicity preference within each action context.

A positive RDS_e score indicates overrepresentation of a specific ethnicity group, while a negative score reflects underrepresentation. A higher SDI score corresponds to more balanced and diverse outputs across all groups. The base reward model, CLIP, shows a mild overrepresentation of the White group (RDS = 0.0182) and achieves the highest SDI score (0.8495), indicating relatively balanced ethnic representation. After fine-tuning, HPSv2.0 shifts its preference toward Middle Eastern individuals (RDS = 0.0315), while HPSv2.1 displays a stronger bias toward the Latino group (RDS = 0.0382). PickScore, by contrast, favors East Asian individuals (RDS = 0.0352). Despite differences in the direction of bias, all fine-tuned reward models exhibit lower SDI scores compared to CLIP, suggesting a decline in ethnic diversity and balance following alignment. Figures 22 to 25 show the ethnicity bias across 42 actions.



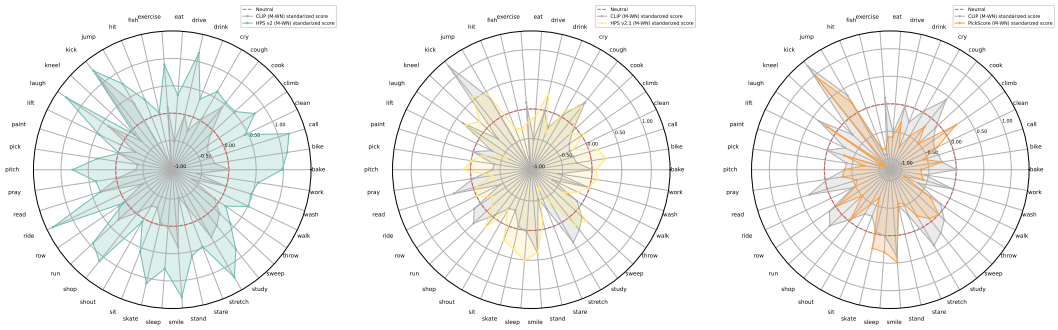
(a) HPSv2.0 vs. CLIP (b) HPSv2.1 vs. CLIP (c) PickScore vs. CLIP

Figure 15: Ethnicity-aware gender bias (White).



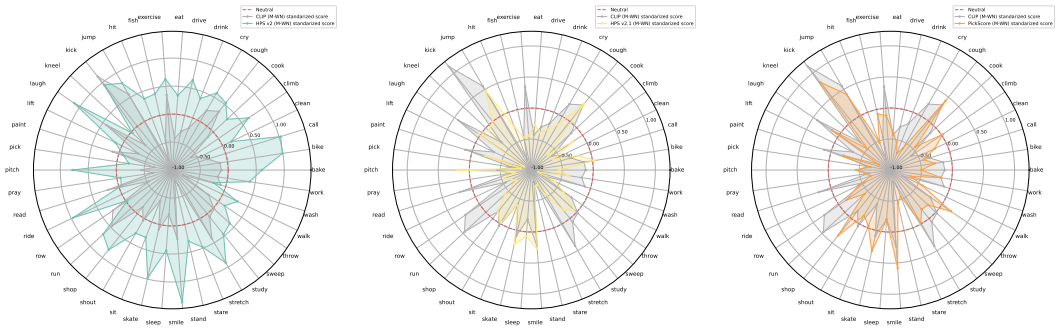
(a) HPSv2.0 vs. CLIP (b) HPSv2.1 vs. CLIP (c) PickScore vs. CLIP

Figure 16: Ethnicity-aware gender bias (Black).



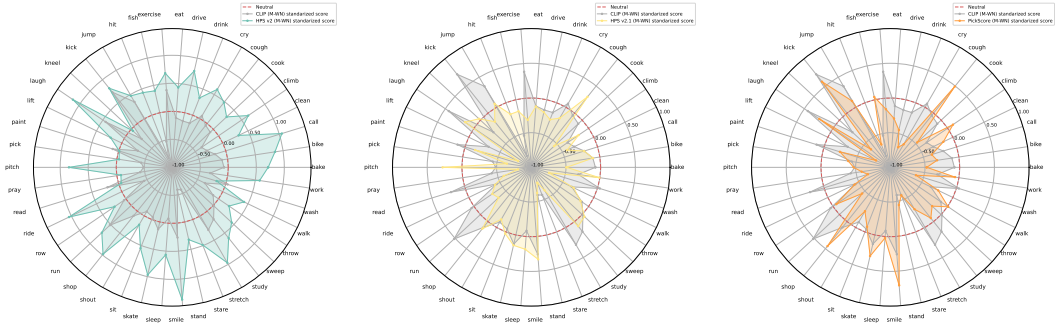
(a) HPSv2.0 vs. CLIP (b) HPSv2.1 vs. CLIP (c) PickScore vs. CLIP

Figure 17: Ethnicity-aware gender bias (Latino).



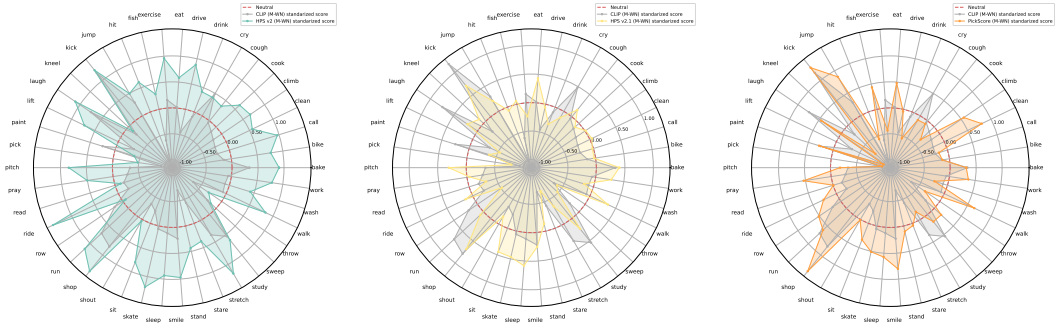
(a) HPSv2.0 vs. CLIP (b) HPSv2.1 vs. CLIP (c) PickScore vs. CLIP

Figure 18: Ethnicity-aware gender bias (East Asian).



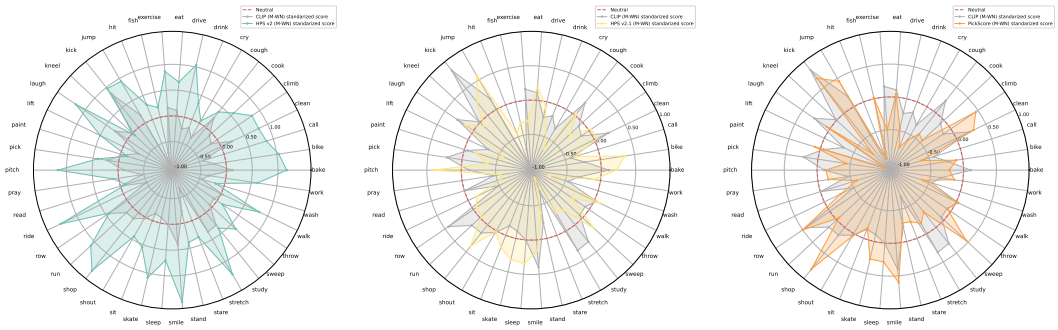
(a) HPSv2.0 vs. CLIP (b) HPSv2.1 vs. CLIP (c) PickScore vs. CLIP

Figure 19: Ethnicity-aware gender bias (Southeast Asian).



(a) HPSv2.0 vs. CLIP (b) HPSv2.1 vs. CLIP (c) PickScore vs. CLIP

Figure 20: Ethnicity-aware gender bias (Indian).



(a) HPSv2.0 vs. CLIP (b) HPSv2.1 vs. CLIP (c) PickScore vs. CLIP

Figure 21: Ethnicity-aware gender bias (Middle Eastern).

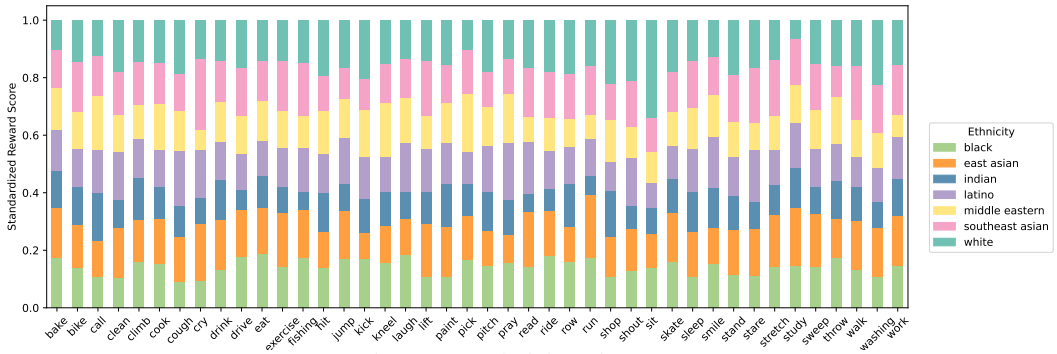


Figure 22: Ethnicity Bias - CLIP

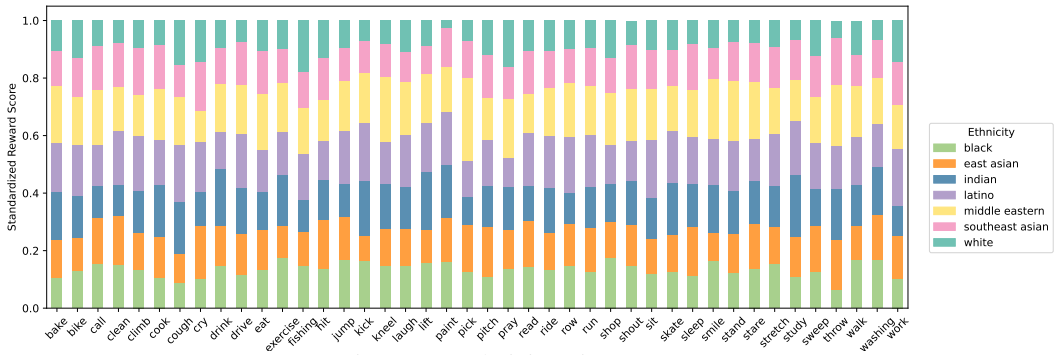


Figure 23: Ethnicity Bias - HPSv2.0

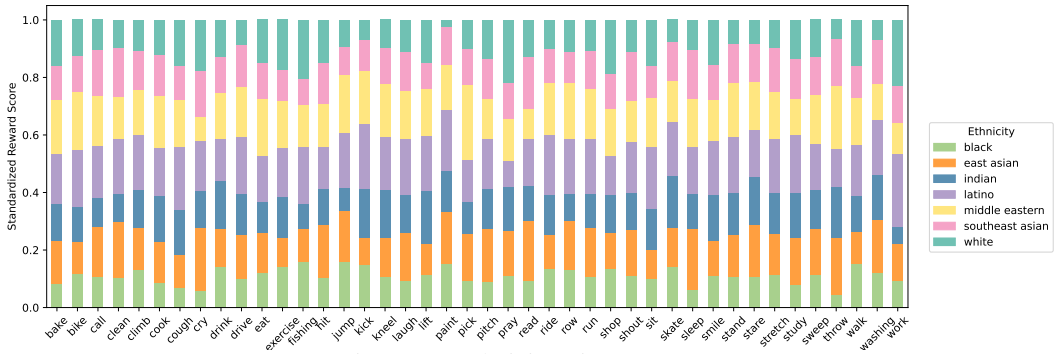


Figure 24: Ethnicity Bias - HPSv2.1

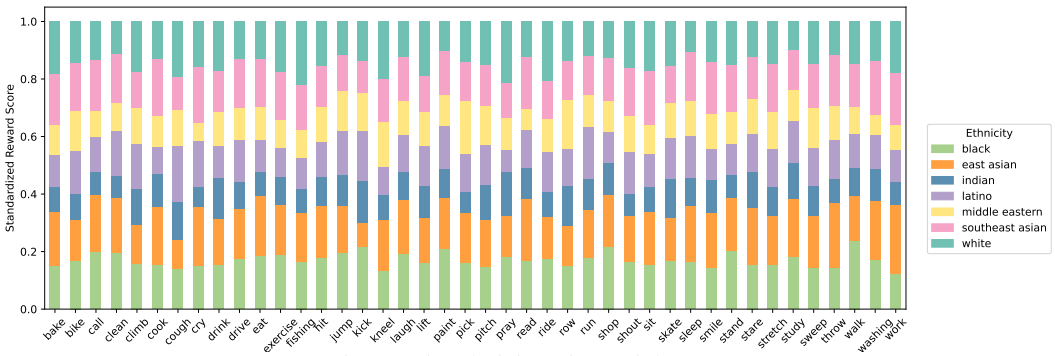


Figure 25: Ethnicity Bias - Pick Score

F Social Biases in Preference Alignment

Building on our analysis of gender and ethnicity biases in image reward models, we examine how preference alignment tuning affects bias in video generation. We fine-tune a Video Consistency Model distilled from VideoCrafter-V2 (VCM-VC2) Li et al. [2024] using three image-text reward models, HPSv2.0, HPSv2.1, and PickScore, and compare social bias distributions before and after tuning to assess how each reward model shapes identity representation. Following the T2V-Turbo-V1 training protocol Li et al. [2024], we incorporate reward feedback into the Latent Consistency Distillation process Luo et al. [2023] by using single step video generation. During student model distillation from a pretrained teacher text to video model, we directly optimize the decoded video frames to maximize reward scores from the image-text alignment models, guiding each frame toward representations more aligned with human preferences.

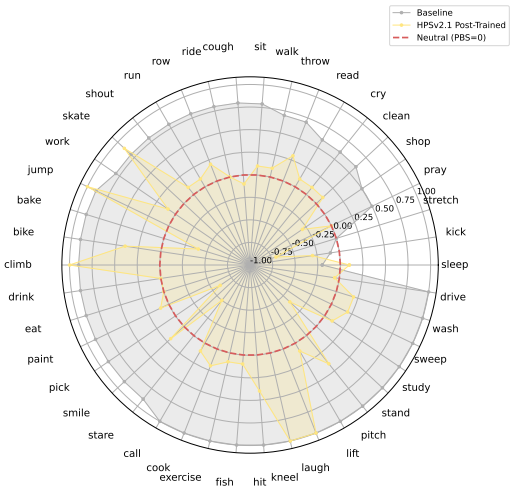
We evaluate aligned video diffusion models using our bias framework (§4). Table 11 reports two metrics: PBS_G for gender imbalance across ethnic groups, and RDS_e and SDI for ethnicity representation disparity and overall output diversity.

Models	Average	White		Black		Latino		East Asian		Southeast Asian		India		Middle Eastern		Overall
	gray!7PBS_G	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	gray!7PBS_G	RDS	SDI
VCM-VC2	gray!70.8034	gray!70.7925	**0.6405**	gray!70.7758	-0.2381	gray!70.8090	–	gray!70.7115	-0.2333	gray!70.7945	–	gray!70.8634	–	gray!70.8071	-0.1690	0.1433
+ HPSv2.0	gray!70.9116	gray!70.9667	0.3667	gray!70.9214	–	gray!70.9667	–	gray!70.8500	–	gray!70.9214	–	gray!70.9000	–	gray!70.8548	-0.3667	0.1257
Δ	gray!7+0.1082	gray!7+0.1742	-0.2738	gray!7+0.1456	–	gray!7+0.1577	–	gray!7+0.1385	–	gray!7+0.1269	–	gray!7+0.0366	–	gray!7+0.0477	-0.1977	-0.0176
+ HPSv2.1	gray!70.2267	gray!70.1321	**0.4286**	gray!70.2381	–	gray!70.3738	–	gray!70.1571	–	gray!70.3619	–	gray!70.1452	–	gray!70.1786	-0.4286	0.0976
Δ	gray!7-0.5767	gray!7-0.6604	-0.2119	gray!7-0.5377	–	gray!7-0.4352	–	gray!7-0.5544	–	gray!7-0.4326	–	gray!7-0.7182	–	gray!7-0.6285	-0.2596	-0.0457
+ PickScore	gray!70.3714	gray!70.3429	**0.6833**	gray!70.3357	-0.1810	gray!70.7190	-0.1929	gray!70.1450	-0.1952	gray!70.4548	–	gray!70.2500	–	gray!70.3515	-0.1143	0.1467
Δ	gray!7-0.4320	gray!7-0.4496	+0.0428	gray!7-0.4401	+0.0571	gray!7-0.0900	-0.1929	gray!7-0.5665	+0.0381	gray!7-0.3397	–	gray!7-0.6134	–	gray!7-0.4556	+0.0547	+0.0034

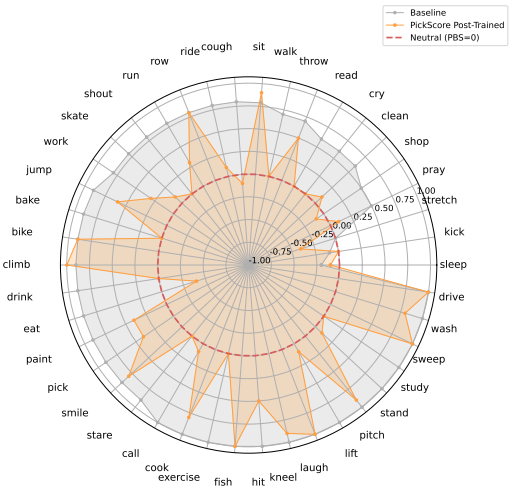
Table 11: Social biases of aligned models. All values represent *average* scores across 42 actions.

Ethnicity-Aware Gender Bias. To assess gender representation across ethnic groups, we prompt each model with "person" specified by an explicit ethnicity and compute PBS_G across 42 actions for each of the seven ethnic groups. A positive PBS_G score indicates a tendency to depict men more frequently, while a negative score suggests a preference for women. The base model, VCM-VC2, demonstrates a strong man bias across all ethnicities, which becomes more pronounced with alignment using HPSv2.0. In contrast, alignment with HPSv2.1 and PickScore significantly reduces PBS_G, indicating a shift toward more balanced or woman-preferred outputs. This change reflects the underlying woman bias present in the HPSv2.1 and PickScore reward models, which steer the model away from the man-dominant bias of the base model. Figures 26 to 33 presents the PBS_G scores across 42 actions for each ethnicity group.

Ethnicity Bias. Ethnic representation is evaluated by prompting each model with "person" while omitting explicit ethnicity, and calculating RDS_e for each group. Positive values indicate over-representation, and negative values indicate underrepresentation. Overall demographic balance is measured using SDI, where higher values reflect more equitable representation. The base model, VCM-VC2, strongly favors White individuals (RDS = 0.6405), while Black, East Asian, and Middle Eastern groups are underrepresented. Alignment with HPSv2.1 reduces some disparities by improving balance for White and Black groups, but significantly decreases Latino representation (RDS = –0.4352) and lowers SDI, indicating reduced diversity. In contrast, PickScore achieves the highest SDI and produces more balanced representation across most ethnic groups, resulting in the most demographically equitable outputs. Figure 34 shows the ethnicity bias across 42 actions.

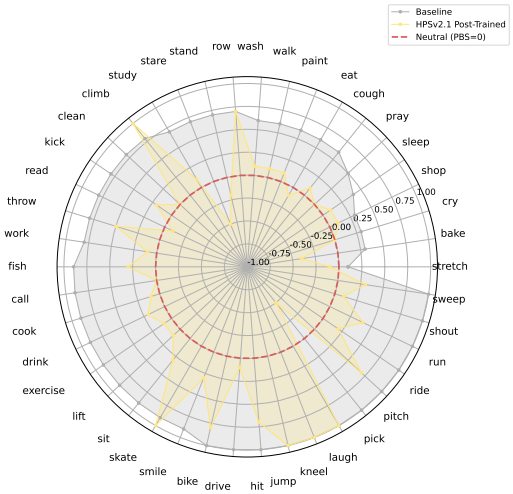


(a) HPSv2.1 vs. CLIP

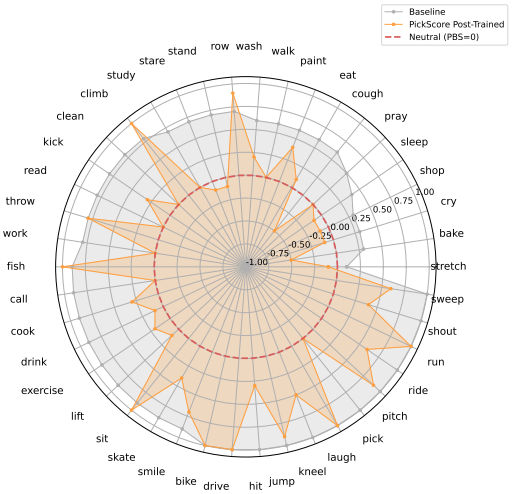


(b) Pick Score vs. CLIP

Figure 26: Ethnicity-aware gender bias (White).

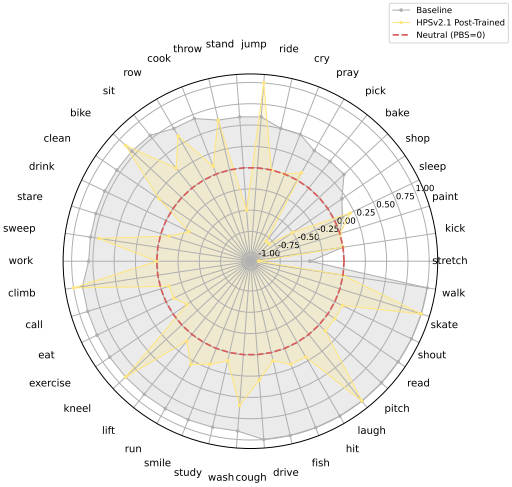


(a) HPSv2.1 vs. CLIP

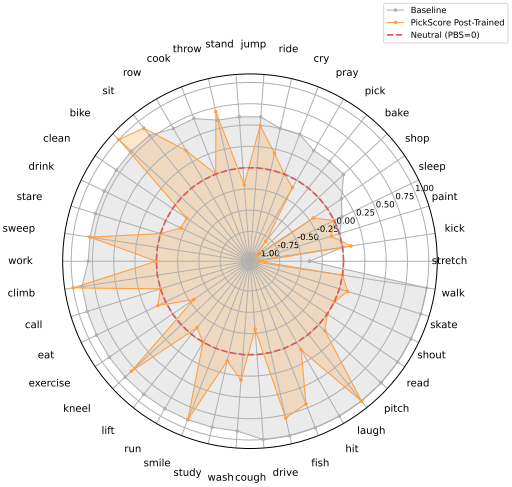


(b) Pick Score vs. CLIP

Figure 27: Ethnicity-aware gender bias (Black).

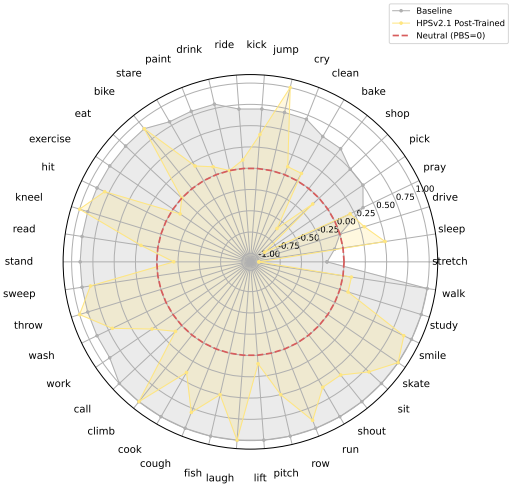


(a) HPSv2.1 vs. CLIP

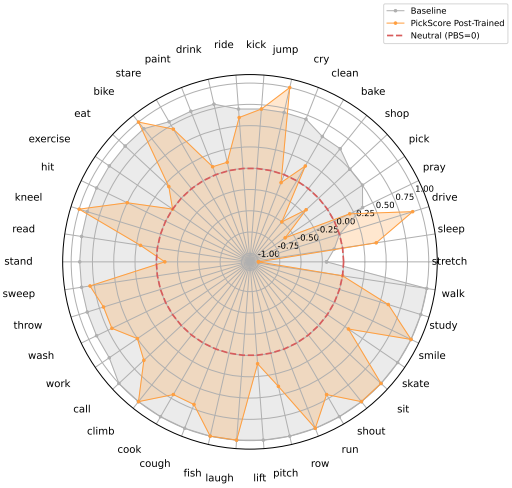


(b) Pick Score vs. CLIP

Figure 28: Ethnicity-aware gender bias (East Asian).

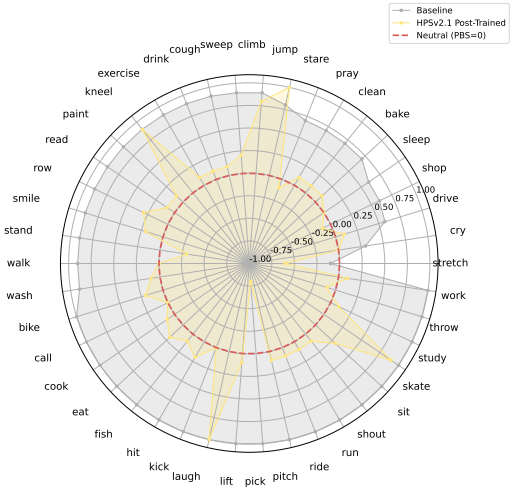


(a) HPSv2.1 vs. CLIP

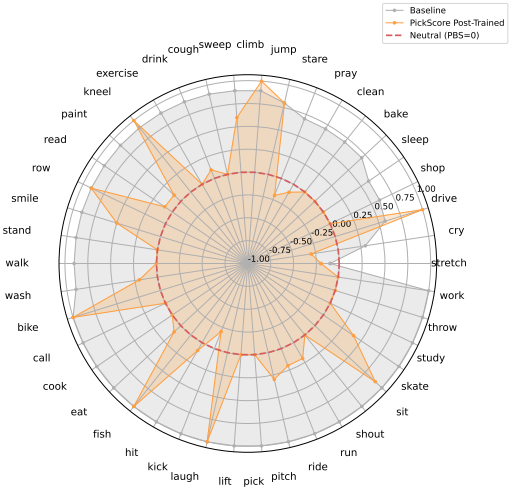


(b) Pick Score vs. CLIP

Figure 29: Ethnicity-aware gender bias (Southeast Asian).

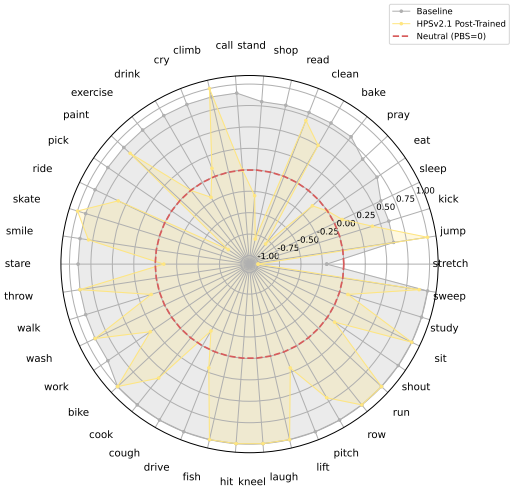


(a) HPSv2.1 vs. CLIP

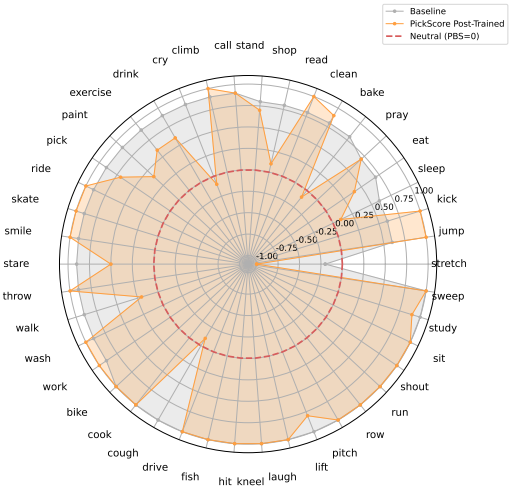


(b) Pick Score vs. CLIP

Figure 30: Ethnicity-aware gender bias (Indian).

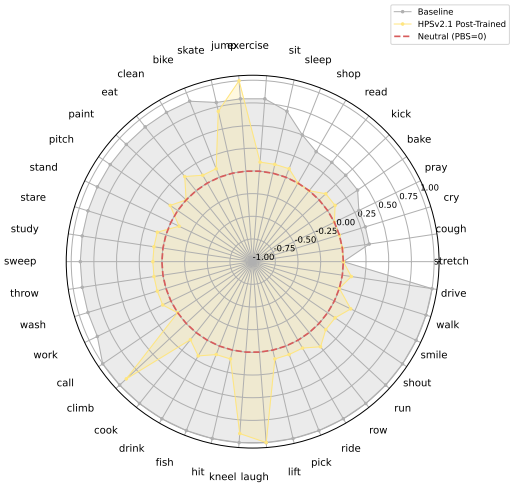


(a) HPSv2.1 vs. CLIP

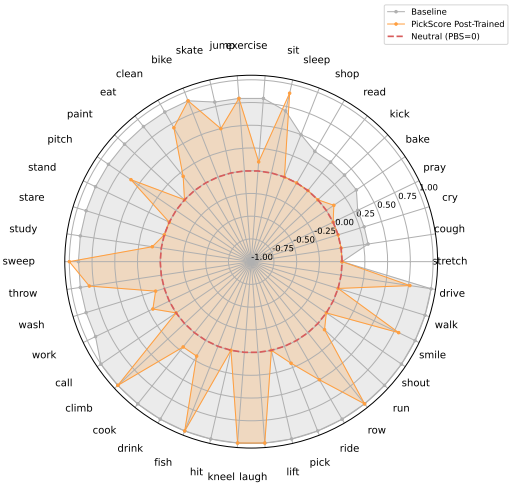


(b) Pick Score vs. CLIP

Figure 31: Ethnicity-aware gender bias (Latino).

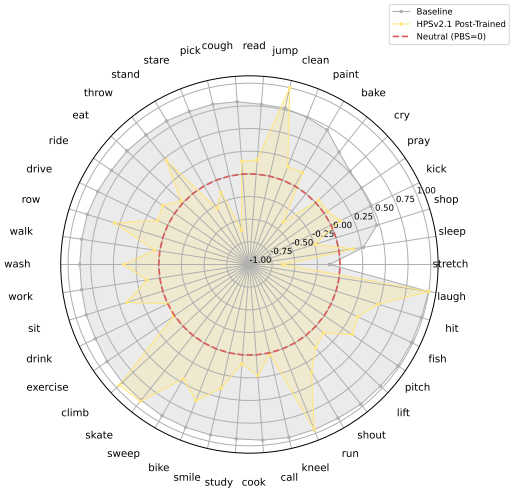


(a) HPSv2.1 vs. CLIP

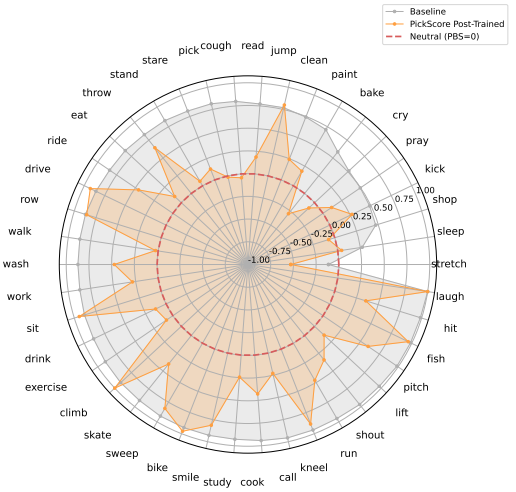


(b) Pick Score vs. CLIP

Figure 32: Ethnicity-aware gender bias (Middle Eastern).



(a) HPSv2.1 vs. CLIP



(b) Pick Score vs. CLIP

Figure 33: Ethnicity-aware gender bias (averaged).

Figure 34: Ethnicity bias distribution

G Controllable Preference Modeling for Video Diffusion Models

Building on prior findings, we observe that reward models trained on imbalanced image preference datasets inherit and amplify social biases. These biases are then reflected in video diffusion models fine-tuned with such reward signals, often leading to unbalanced outputs. In this section, we explore whether manipulating the distribution of social attributes in image datasets allows for controllable bias in reward models, enabling video models to produce more equitable outputs Sheng et al. [2020].

G.1 Image Reward Dataset Construction

Building on the generated images from §5.1, we construct two case-specific reward datasets: one with a man-preferred bias and the other with a woman-preferred bias. The man-preferred dataset is designed to steer both the reward model and the downstream diffusion model toward favoring man representations. Conversely, the woman-preferred dataset encourages a shift toward woman representations. Notably, when applied to a base video diffusion model that exhibits a man-preference bias, the woman-preferred dataset can serve as an effective counterbalance, enabling the training of models with more equitable gender representation.

More specifically, we construct preference pairs using images from the `Gender+Ethnicity` dataset by selecting two images that depict the same action and belong to the same ethnicity group, one featuring a man and the other a woman (for example, images M-1 and W-1 in Figure 3). These image pairs are used to train reward models with prompts of the form: "A/An `[ethnicity]` person is `[action]`-ing `[context]`." For the man-preference dataset, we assign a reward score of 1 to the image with a man character and 0 to the image with a woman character. In contrast, f or the woman-preference dataset, we assign a reward score of 0 to the image with a man character and 1 to the image with a woman character. This process results in 2.94 million preference pairs in each dataset, calculated as 42 actions multiplied by seven ethnicity groups, with 100 male and 100 female images per group. To improve the representation of no-face content, we additionally incorporate 537,660 face-free image pairs from HPDv2, which enhances balance in our proposed reward datasets.

G.2 Image Reward Model Development & Alignment Tuning

Leveraging the man-preferred and woman-preferred image datasets introduced in §7.1, we fine-tune two reward models on top of a pre-trained CLIP vision encoder: the Man-Preferred Reward Model (RM_M) and the Woman-Preferred Reward Model (RM_W). Each model is trained to reflect gender-specific preferences based on its respective dataset. As shown in Table 12, RM_M consistently assigns greater PBS_G scores across all demographic groups, indicating a strong alignment with man-preferred representations. In contrast, RM_W exhibits an opposite trend, systematically favoring woman-preferred content. The clear divergence between these models highlights the effectiveness of reward tuning in capturing and reinforcing gendered preferences.

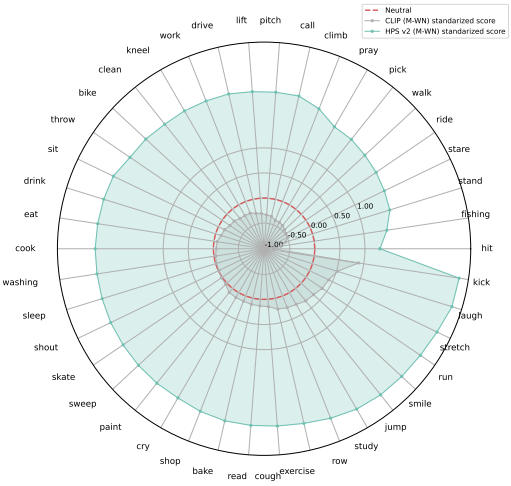
Models	Average	White	Black	Latino	East Asian	Southeast Asian	India	Middle Eastern
CLIP	-0.0726	0.0343	-0.1198	-0.0934	-0.1315	-0.0865	-0.0508	-0.0607
RM_M	$1.5280_{+1.60}$	$1.6300_{+1.60}$	$1.5752_{+1.70}$	$1.5524_{+1.65}$	$1.4323_{+1.56}$	$1.4525_{+1.54}$	$1.5619_{+1.61}$	$1.4914_{+1.55}$
RM_W	$-0.7448_{-2.27}$	$-0.6318_{-2.26}$	$-0.7943_{-2.37}$	$-0.8279_{-2.38}$	$-0.6282_{-2.06}$	$-0.6429_{-2.10}$	$-0.8846_{-2.45}$	$-0.8042_{-2.30}$

Table 12: Preference bias of reward models. All values represent *average* scores across 42 actions.

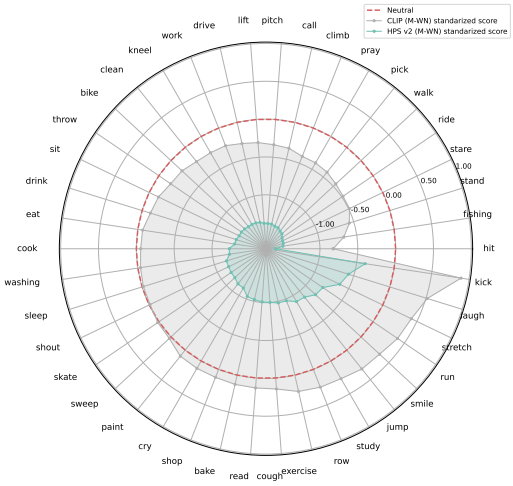
Building on our earlier reward model training, we applied RM_M and RM_W to guide alignment tuning of a base video diffusion model using the same preference-driven training strategy. These reward signals enabled the generation of two distinct variants: one aligned with man-preferred content and the other with woman-preferred content. As shown in Table 13, alignment with RM_M led to consistently greater PBS_G scores across all demographic groups, reinforcing man-preference bias. Conversely, alignment with RM_W resulted in substantially smaller scores, indicating a strong shift toward woman-preference bias. These results confirm that our controllable preference modeling approach can effectively modulate gender bias in video generation, offering a flexible mechanism to either amplify or reduce specific social tendencies in model outputs. Figures 35 to 39 presents the PBS_G scores across 42 actions for each ethnicity group.

Models	Average	White	Black	Latino	East Asian	Southeast Asian	India	Middle Eastern
VCM-VC2	0.8034	0.7925	0.7758	0.8690	0.7115	0.7945	0.8634	0.8071
+ RM_M	$0.9584_{+0.16}$	$0.9595_{+0.17}$	$0.9524_{+0.18}$	$0.9756_{+0.11}$	$0.9437_{+0.23}$	$0.9447_{+0.15}$	$0.9640_{+0.10}$	$0.9709_{+0.16}$
+ RM_W	$0.3082_{-0.50}$	$0.3341_{-0.46}$	$0.3913_{-0.38}$	$0.3314_{-0.54}$	$0.1008_{-0.61}$	$0.2639_{-0.53}$	$0.3446_{-0.52}$	$0.3894_{-0.42}$

Table 13: Social biases of aligned models. All values represent *average* scores across 42 actions.

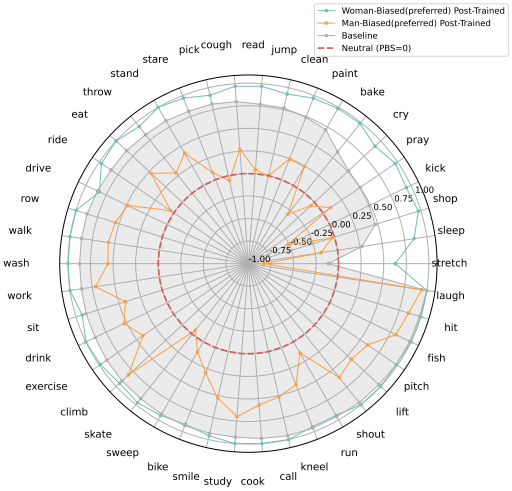


(a) Ethnicity-aware gender bias (averaged) of man-preferred reward model RM_M.

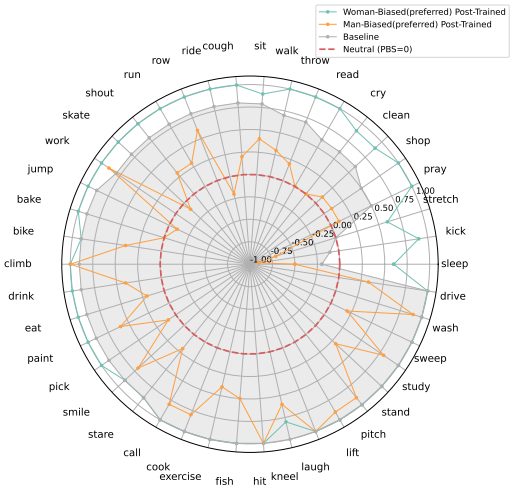


(b) Ethnicity-aware gender bias (averaged) of woman-preferred reward model RM_W.

Figure 35: Ethnicity-aware gender bias (averaged) of woman-preferred reward model RM_M and RM_W.



(a) Ethnicity-aware gender bias (averaged) of man-preferred and woman-preferred post-trained video generation model by reward model RM_M and RM_W.

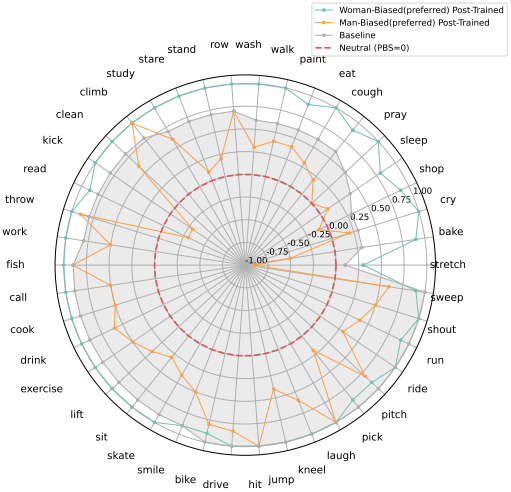


(b) Ethnicity-aware gender bias (White) of man-preferred and woman-preferred post-trained video generation model by reward model RM_M and RM_W.

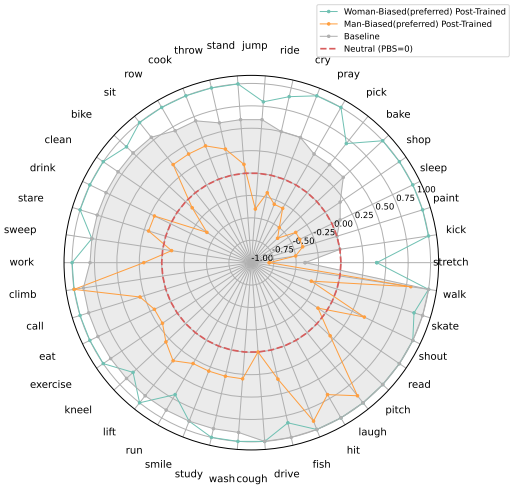
Figure 36: Ethnicity-aware gender bias (White) of man-preferred and woman-preferred post-trained video generation model by reward model RM_M and RM_W.

G.3 Actions Correlation Analysis

We analyze the changes in the reward model preference for 42 events and the bias of the video generation model before and after post-training, using the training results from §7. In Figure 40, the horizontal axis represents the reward model preference (PBS_G), and the vertical axis represents the change in the video generation model's bias before and after post-training ($\Delta\ \mathrm{PBS}_G$). In Figure 41 and Figure 42, the horizontal axis represents the event, and the vertical axis represents the change in the video generation model's bias before and after post-training ($\Delta\ \mathrm{PBS}_G$) divided by the reward model preference (PBS_G). This ratio indicates the sensitivity of a particular event to the bias during post-training. We have arranged the events in the figure from left to right in ascending order of the vertical axis values; events further to the right are more sensitive.

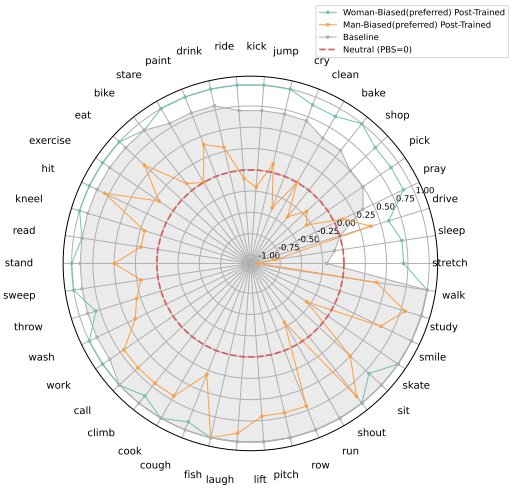


(a) Ethnicity-aware gender bias (Black) of man-preferred and woman-preferred post-trained video generation model by reward model RM_M and RM_W.

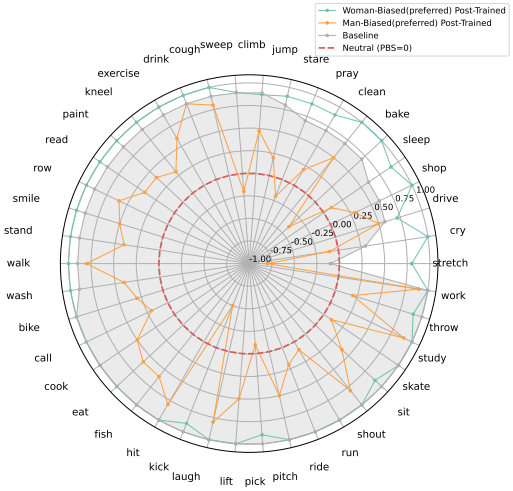


(b) Ethnicity-aware gender bias (East Asian) of man-preferred and woman-preferred post-trained video generation model by reward model RM_M and RM_W.

Figure 37: Ethnicity-aware gender bias of man-preferred and woman-preferred post-trained video generation model by reward model RM_M and RM_W.



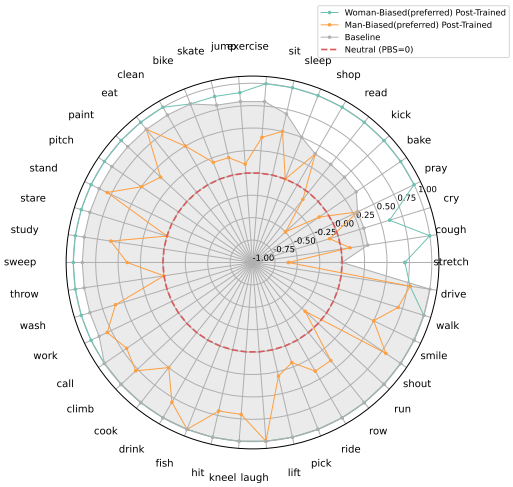
(a) Ethnicity-aware gender bias (Southeast Asian) of man-preferred and woman-preferred post-trained video generation model by reward model RM_M and RM_W.



(b) Ethnicity-aware gender bias (Indian) of man-preferred and woman-preferred post-trained video generation model by reward model RM_M and RM_W.

Figure 38: Ethnicity-aware gender bias of man-preferred and woman-preferred post-trained video generation model by reward model RM_M and RM_W.

(a) Ethnicity-aware gender bias (Latino) of man-preferred and woman-preferred post-trained video generation model by reward model RM_M and RM_W.



(b) Ethnicity-aware gender bias (Middle Eastern) of man-preferred and woman-preferred post-trained video generation model by reward model RM_M and RM_W.

Figure 39: Ethnicity-aware gender bias of man-preferred and woman-preferred post-trained video generation model by reward model RM_M and RM_W.

H Reward Model Training and Inference Details

For both the training and inference of the reward model (RM), we largely followed the settings outlined in Wu et al. [2023]. We also utilized the HPSv2 codebase available at `https://github.com/tgxs002/HPSv2` for these processes.

Training: We employed a batch size of 16 and the AdamW optimizer. The man-preferred and woman-preferred datasets that we constructed were adapted to the data loading format specified in the HPSv2 code (`https://github.com/tgxs002/HPSv2`). Ultimately, we trained the RMs for man-preferred and woman-preferred data for 1 epochs (equivalent to 23000 steps), with no data repetition within each step. The model training was initialized from a CLIP checkpoint.

Inference: We used the CLIP score as the inference score for the RM.

I Video Model Post-Training and Inference Details

For post-training during alignment tuning, we used the t2v-turbo-v1 codebase Li et al. [2024], available at `https://github.com/Ji4chenLi/t2v-turbo`. A reward model loss scale of 1 was applied. The video model was jointly trained with both the reward model loss and the diffusion loss over 200 steps, using data sampled from the WebVideo dataset.

For inference, we also utilized the same t2v-turbo-v1 codebase. Each inference setting was run 10 times with different random seed to ensure consistency and robustness of the results.

J Limitations

While our work presents a comprehensive evaluation of social biases introduced through alignment tuning in video diffusion models, several limitations warrant further consideration. *First*, our analysis focuses on two social dimensions, gender and ethnicity, using predefined categories based on U.S. Census conventions and prior literature. These categories, while practical for controlled evaluation, are inherently socially constructed and cannot fully capture the fluidity, intersectionality, or cultural nuances of identity. Future work should explore richer identity representations, including intersectional groups. *Second*, our VLM-based evaluators, though validated against human judgments, rely on image-level classification and may exhibit their own biases or inaccuracies, particularly when interpreting identity in stylized or ambiguous frames. While we ensemble multiple models to mitigate

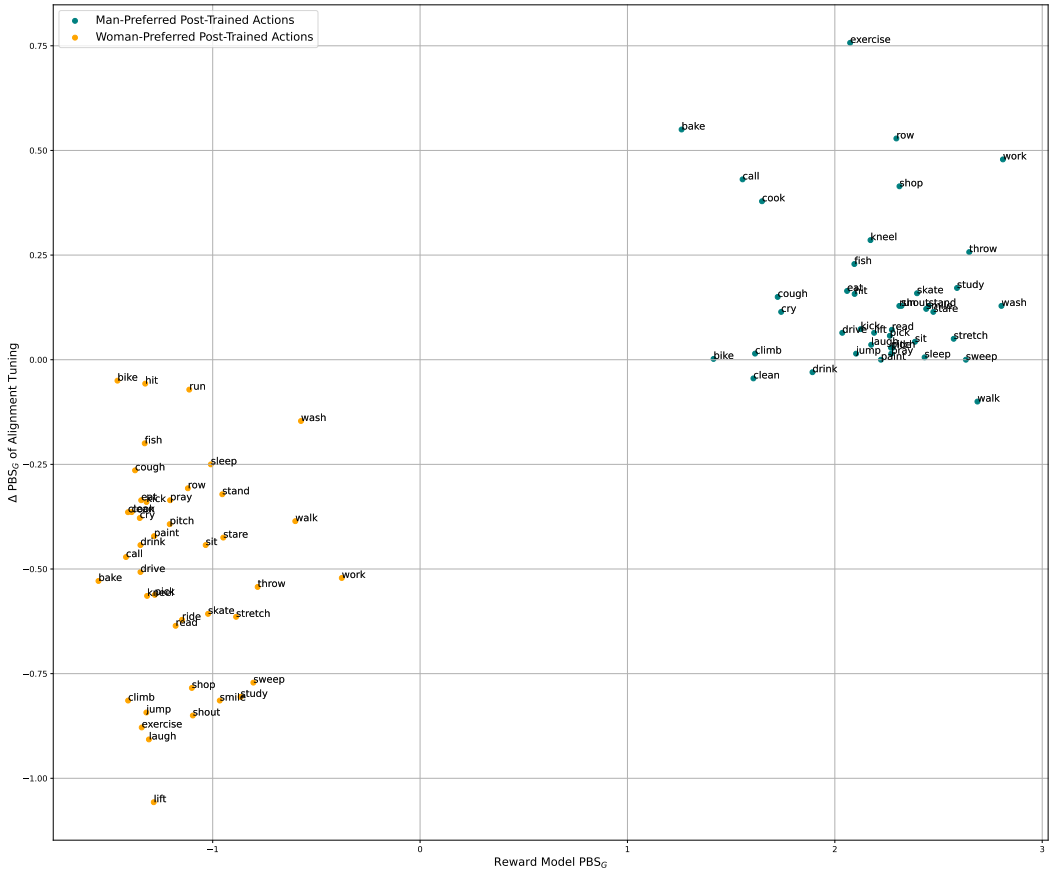


Figure 40: Δ PBS$_G$ of video generation model before and after alignment tuning by RM$_\text{M}$ and RM$_\text{W}$. Results are broken down into actions. Figure 41 and Figure 42 are based on this figure.

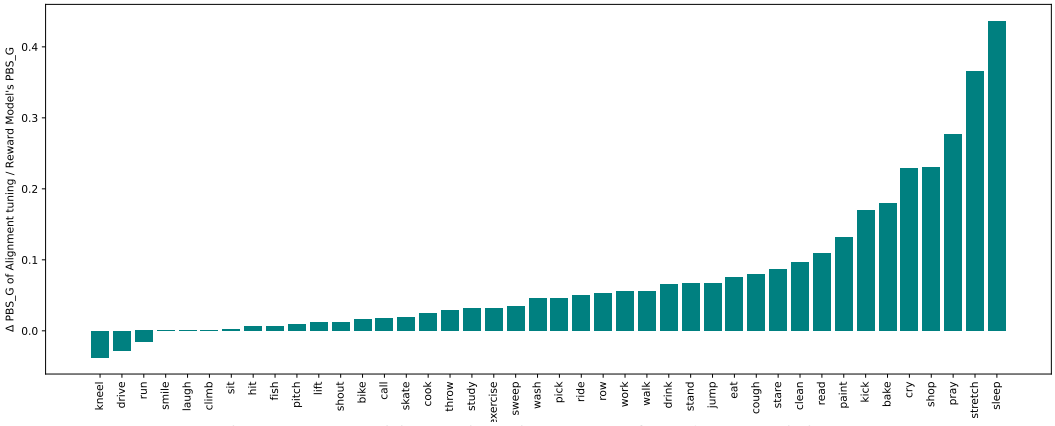


Figure 41: Sensitive actions in man-preferred post-training.

this, ground truth annotations for a larger and more diverse set of videos would further strengthen the reliability of our measurements. *Third*, we primarily assess alignment impacts under a specific training strategy (single-step latent consistency distillation) and a limited set of reward models. Other training protocols, such as RL-based tuning or multi-turn video instruction alignment, may exhibit different bias dynamics not captured in our study. *Fourth*, our controllable preference modeling experiments, while demonstrating the feasibility of targeted bias modulation, are constrained to synthetic manipulations of gender preference. These interventions do not address broader questions of value alignment, normative appropriateness, or long-term societal impact, which are crucial for the responsible deployment of generative video systems. *Lastly*, our evaluation framework, VIDEOBIASEVAL, is currently benchmarked on a fixed set of 42 socially associated actions. While this enables fine-grained control, it may limit generalizability to open-ended generation settings or novel actions

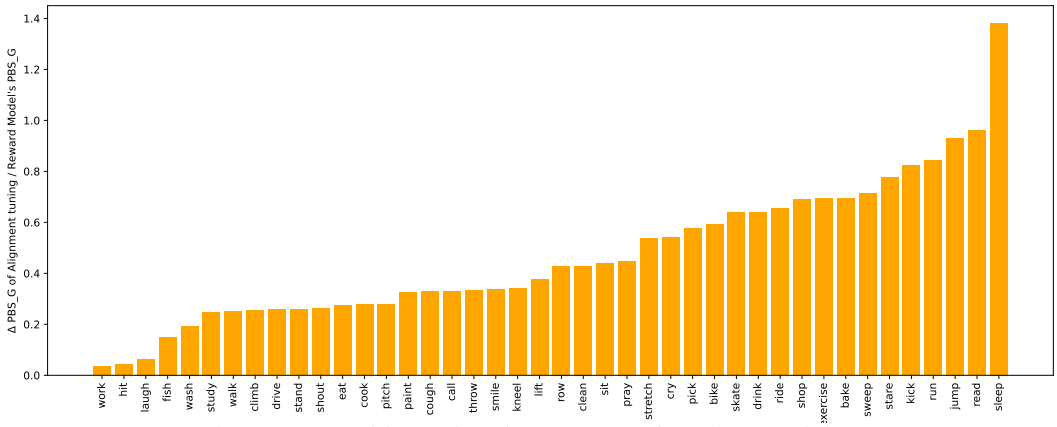


Figure 42: Sensitive actions in woman-preferred post-training.

not covered in our taxonomy. We hope that these limitations encourage further research into holistic, culturally grounded, and ethically aligned evaluation pipelines for video generative models.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state them in the introduction and abstract.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we have limitation section Appendix J.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: N/A

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, we provide the details in §3, §4, §5, §6, and §7.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, we provide the details in §3, §4, §5, §6, and §7.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, we provide the details in §3, §4, §5, §6, and §7.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: N/A

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, we provide the details in §4, §5, §6, and §7.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, we describe it in the introduction and limitation sections.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: N/A

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, we provide the details in §5.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: N/A

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: N/A

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: N/A

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: N/A

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.