What TikTok Claims, What Bold Glamour Does: A Filter's Paradox

Miriam Doh
Université de Mons (UMONS)
ISIA Lab
Mons, Belgium
Université Libre de Bruxelles (ULB)
IRIDIA Lab
Brussels, Belgium
miriam.doh@umons.ac.be

Corinna Canali
Weizenbaum Institute/Universität der Künste Berlin
Berlin, Germany
c.canali@udk-berlin.de

Nuria Oliver
ELLIS Alicante
Alicante, Spain
nuria@ellisalicante.org

Abstract

This paper critically examines the transformations applied by AI-driven augmented reality (AR) beauty filters, using TikTok's *Bold Glamour* filter as a case study. Through a multidisciplinary analysis, we explore how this hyper-realistic filter modifies user appearances, reinforces Eurocentric beauty standards, and perpetuates intersectional biases in gender and racial representation. The findings reveal discrepancies between TikTok's inclusivity claims and the filter's actual impact on faces. By analyzing the filter's impact across different demographic groups, this study highlights its alignment with market-driven objectives that commodify self-representation. We discuss the broader societal implications of such technologies, advocating for enhanced transparency and ethical governance to mitigate the biases embedded in digital self-representation tools.

CCS Concepts

• **Human-centered computing** → *Social media.*

Keywords

AR beauty filters, Social media, Transparency, Gender Bias, Self-representation.

ACM Reference Format:
Miriam Doh, Corinna Canali, and Nuria Oliver. 2025. What TikTok Claims, What Bold Glamour Does: A Filter's Paradox. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3715275.3732126

1 Introduction

Visually oriented platforms like TikTok [81], Instagram [57], and YouTube [37] are reshaping how individuals construct and present their bodies and identities in the digital world [8, 36, 47, 48]. This move toward digital self-construction has been closely tied to significant investments in technologies designed to shape, curate, and manage user interactions through self-image––efforts increasingly driven by monetization and market-oriented objectives. As noted in [43]: *"visually oriented platforms [...] are increasingly the site for identity-making online"* which has become a core component of platform economies and governance strategies. In this context, identity, tied to body appearance, is expressed and commodified [56]. Personalized advertisement and the deployment of AI-based content moderation and classification algorithms have entrenched sexual, gender, and racial biases within the tools used for digital interaction, self-expression, and self-presentation in social platforms [30, 62, 63, 70]. Governed by algorithmic processes, these platforms act as intermediaries, shaping perceptions of identity and self-representation and amplifying pressures on users to conform to idealized norms. Historically scrutinized for their appearance [24], women are particularly vulnerable to this phenomenon, with these platforms controlling self-expression under a veneer of autonomy, safety, and creativity.

The immensely popular AI-based augmented reality (AR) beauty filters exemplify these dynamics, serving as visual tools where the physical and virtual worlds converge [46]. Although marketed as empowering tools for creativity, beauty filters like TikTok's *Bold Glamour* filter—illustrated in Figure 1—have been found to reduce diversity and reinforce racial and gender biases rooted in Western beauty ideals, subtly normalizing and amplifying harmful norms in opaque ways [69, 71]. While a more substantial body of research has studied Instagram's AR beauty filters, the impact of these type of filters on TikTok—a platform centered on short-form video user-generated content (UGC) curated through human and algorithmic processes—remains unexplored. Existing frameworks for the analysis of AR beauty filters, such as OpenFilter [71] are also inapplicable to TikTok due to recent platform updates[1]. Furthermore, TikTok's advanced filter technologies, capable of performing hyperrealistic, real-time modifications to users' videos, have the potential to amplify biased influence on self-perception, body image, and user engagement. To shed light on this topic, we conduct a case study of TikTok's *Bold Glamour* beauty filter, an extremely popular filter designed and deployed by TikTok. We argue that these popular tools are not neutral; instead, they are imbued with cultural, social, and economic values that privilege certain more productive and normative identities over others. In this paper, we provide first an

FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732126

[1]During the writing of this paper, new policies and regulations were implemented on platforms such as Instagram and TikTok. Starting in December 2024, in response to growing criticism of AR beauty filters, TikTok introduced a policy prohibiting users under 18 from using beauty filters [4], while in January 2025 Instagram announced the banning of beauty filters entirely (see note 3). However, as of today, a quick test revealed that some available filters still apply beautifying features to the users' faces, a matter that requires further investigation. For further details on current updates, see A

overview of AI-based beauty filters from a sociotechnical perspective. Second, we draw inspiration from [20] to analyze TikTok's popular *Bold Glamour* beauty filter, focusing on the aesthetic transformations made by the filter, with particular attention to gender and racial biases and their alignment with Western beauty standards. Third, we study how the filter aligns or differs from the platform's declared inclusivity policies. Finally, we propose pathways for more equitable design and regulation of AI-driven identity technologies, emphasizing the need for transparency, diversity, and accountability in the development and deployment of these tools.

2 Related work

AI-based AR filters are automated photo-editing tools that leverage computer vision algorithms to apply customized effects to images or videos of faces in real time. Since their introduction on social media in 2015, these filters have evolved significantly in accuracy and popularity. Millions of users now engage with AR filters on visual-based platforms which also offer filter creation tools such as Meta Spark AR[2] and TikTok's Effect House[3]. AR filters serve various purposes: some are designed for playful photomontages, while others are brand-specific, allowing users to try on products like makeup or accessories virtually. A particular type of filters are *beauty filters*, explicitly designed to apply predefined beauty standards to the faces of their users, which typically include smoothing the skin and modifying facial features, such as the lips, nose, eyes, eyelids and cheeks, to conform with a socially constructed ideal of attractiveness.

In the past few years, although still limited by the platforms' non-disclosure practices, growing academic and public scrutiny has advanced the general understanding of the biases embodied and perpetuated by AR beauty filters. Previous work has addressed racial bias and colorism [73], and the filters' ability to reinforce idealized and stereotypical beauty canons that affect self-perception [13, 29, 45, 58], leading to heightened body self-surveillance based on Western-aligned standards. Recent work by [69] presents an in-depth examination of how beauty filters perpetuate racial biases, employing explainable AI techniques and publicly available datasets of beautified faces [71]. A complementary strand of research takes a more ethnographic and phenomenological approach and has explored how users experience and are impacted by these filters on a personal level through small ethnographic studies. For example, Rosalind Gill conducted in 2021 a survey for the City, University of London, where 175 UK-based young women (aged 18–30) were interviewed. The findings showed that around 90% of participants had used AR beauty filters, with 48% using them at least once a week [34, 39]. The study further highlighted that participants most commonly used these filters to even or alter skin tone, whiten teeth, enlarge eyes, plump lips, narrow noses, reshape jaws, and reduce weight (the so-called "skinny-filter"). Additionally, the survey reported that 94% of participants felt "pressure to look a certain way on social media" ([34]), with nearly 80% stating that social media negatively impacts their self-perception, and 60% reporting feelings of depression due to these aesthetic expectations. The psychological and social impact of beauty filters, especially on young females––a demographic group traditionally affected by prescriptive beauty standards [52, 60]– has been further studied by several authors [26, 45, 66].

Building on this body of work, we argue that further scrutiny of beauty filters is necessary. Drawing on Gerrard and Thornham's framework of social media "sexist assemblages" as a lens to understand how digital governance perpetuates normative gender roles through both human and mechanical elements [32], we similarly use Deleuzian theory to conceptualize beauty filters as collective assemblages that articulate territoriality and identity, matching "those social forms capable of generating them and using them" [16, 78, 79]. Hence, it is crucial to critically examine their opaque modes of creation, deployment, and governance to unpack further the nature and extent of the intersectional biases informing beauty filters as integral to their design. This approach also calls for studies that engage with the complex sociotechnical and political heritage that informs the technologies under study [21, 38]. To address these challenges, this paper adopts a multifaceted approach that integrates socio-cultural analyses with technical considerations, examining how platform governance and algorithmic design contribute to the reproduction and circulation of gender biases. As a case study, we analyze TikTok's *Bold Glamour* beauty filter, released by TikTok at the end of 2022 and widely adopted in the platform.

3 From Snapchat's AR filters to TikTok's Bold Glamour

Figure 1: Examples of the *Bold Glamour* filter applied to a female and a male samples (A = Asian, L = Latino/a) and genders (F = Female, M = Male). In each pair, the first image represents the original, unfiltered face (*e.g.*, *LF*), while the second image shows the face after the filter has been applied (*e.g.*, *LF), illustrating the aesthetic changes made by the filter. Original samples from Chicago Face Database (2015).**

The introduction of AR filters on Snapchat in 2015, following its acquisition of Looksery, revolutionized digital self-representation by incorporating 3D elements through what the platform branded as "Lenses" [12]. Snapchat's Dog Lens (2016) became iconic, blending playful overlays with subtle beautification aligned with Western beauty standards [14]. Promoted by influencers like Kim Kardashian [50, 55], the filter gained popularity but soon became a gendered tool, primarily associated with feminine users and subjected to sexist critiques, earning derogatory labels like the "hoe filter" [19, 42].

[2]https://spark.meta.com/; Note: Meta announced that it will discontinue AR beauty filters and shut down third-party AR effects and APIs for creating filters via Spark AR, starting January 2025.
[3]https://effecthouse.tiktok.com/

AR filter technology has since advanced, with platforms like Instagram[4] and TikTok enabling user-generated filters and utilizing sophisticated AI to create real-time, seamless transformations. Among these, *beauty filters* have become particularly popular, enhancing user appearances by reshaping and adjusting various facial features. An emblematic example is TikTok's *Bold Glamour*, developed by TikTok in 2023. Renowned for its unprecedented hyper-realism and seamlessness [72], it quickly went viral [75]. As of today, it has been featured in around 250 million videos. However, the hyper-realism of *Bold Glamour*, which not only achieves realistic accuracy but also superimposes idealized and vivid adaptations of reality, together with a lack of transparency in its design, deployment and adoption, has raised significant concerns about how this and other beauty filters shape perceptions of attractiveness, self-worth, and identity. Traditional media has long established beauty standards by presenting them on selectively chosen bodies, promoting methods to emulate these ideals. Beauty filters transform this practice by superimposing beauty standards in real time on users, thus allowing them to become the standard momentarily. This blending of physical and digital aesthetics embeds harmful biases into blurred digital realities, extending critiques of media's influence on gendered representations and psychological well-being [5, 6, 33, 36, 49]. The Westernized, heteronormative, and often racially biased beauty ideals embedded in these filters [69, 71], perpetuate exclusionary norms and deepen discriminatory standards. They amplify self-surveillance, particularly among women, pressuring conformity to predefined standards and deepening discriminatory impacts on body image and identity [35, 68, 77]. In addition, this blurring of the digital and physical selves is reinforced by social media platforms, which profit from and promote normative representations of femininity. Notably, TikTok exemplifies this dynamic, leading the beauty e-commerce sector with the platform achieving, in 2023 alone, a global Gross Merchandise Value (GMV) of roughly 2.5 billion USD from beauty sales with 370 million beauty and personal care products sold worldwide via TikTok Shops [88].

In this paper, we examine how beauty filters and particularly TikTok's *Bold Glamour* are shaped by and reinforce pre-existing heteronormative gender and market-driven content curation, which serves to profit the platforms from their users' aesthetic [24, 25] and glamour labor [85]. To perform our analysis, we have developed an improved version of the `Disclaimer Block` [20] tool that quantifies the transformations the filter applied to different facial features. We study the impact of the filter on a diverse set of faces to shed light on its effects on individuals of different genders and racial groups. Furthermore, we explore TikTok's "beauty paradox", highlighting the tension between fostering self-expression and pressuring users towards normative attractiveness and gendered productivity. By placing beauty filters within broader socio-cultural and economic systems, our analysis critiques their role in perpetuating intersectional biases and discriminatory structures, a pattern not unique to TikTok but prevalent across major Western-aligned social media platforms.

[4]Following criticism and legal actions [74], Meta announced that, starting in January 2025, it will discontinue AR beauty filters across its platforms. This includes the removal of all third-party face filters and AR effects on Facebook, Instagram, and Messenger, along with the shutdown of APIs enabling users to create filters via Meta Spark AR [15]; [59])

4 Analysis of the Bold Glamour Beauty Filter

This section presents the methodology and results of analyzing the impact of the *Bold Glamour* filter on a diverse set of faces. Our analyses are driven by the following research questions:

RQ1: Does *Bold Glamour* brighten the faces? Previous work has reported a potential brightening of the faces due to the application of Instagram's beauty filters [69], yet there is a lack of conclusive evidence in this regard.

RQ2: Are the filter transformations dependent on gender and race? While the *Bold Glamour* filter claims to personalize its effects based on the user's face shape and features, there is a lack of research on the role that gender and race play in the face modifications applied by the filter.

RQ3: Does *Bold Glamour* apply a facial feature morphological alignment? Recent research has suggested the existence of a *white* racial bias in Instagram's filters [69]. Yet, there is a lack of detailed, quantitative research on the transformations applied by the filter and its potential morphological alignment.

4.1 Dataset

Inspired by the data collection techniques described by [61], we selected 208 face images from the Chicago Face Database [54]: 26 images per race and gender category, across two genders [M = Male, F = Female] and four race categories [W = White, B = Black, A = Asian, L = Latino], which will be used as abbreviations throughout the paper. One-minute videos were generated from the images to adapt to TikTok's constraints, each featuring multiple face images. This process resulted in 16 videos capturing both unfiltered and filtered facial data. The videos were then displayed in front of a phone mounted on a selfie ring light to ensure consistent lighting and stable conditions. Each face was first recorded without filters and then with the *Bold Glamour* filter applied to them. This setup minimized variability in lighting and positioning across recordings, allowing accurate comparisons between unfiltered (F) and filtered (F^*) images. After video capture, frames corresponding to F and F^* were extracted and reassociated with their respective gender and racial labels per the Chicago Face Database metadata. The extracted frames were then processed and refined, preparing them for the analysis described next.

4.2 Methodology to Characterize Facial Features

To automatically characterize the facial features, we adapted and expanded the `Disclaimer Block` or `DB` framework introduced by [20], which is designed to bridge the gap between the technological opacity of beauty filters and user awareness. This revised approach, called `Disclaimer Block` V.2 (`DB` V.2) integrates enhanced imaging techniques to provide a more detailed analysis of the changes introduced by the beauty filter. Figure 2 illustrates both the original (V.1) and refined (V.2) `DB` frameworks applied to the original LF and beautified LF^* images shown on the left. For a detailed discussion of the improvements in `DB` V.2, refer to Appendix B.

Disclaimer Block V.1. The original implementation of the `Disclaimer Block` framework relied on a region-based analysis of the face to visualize and quantify changes introduced by the filter. This was

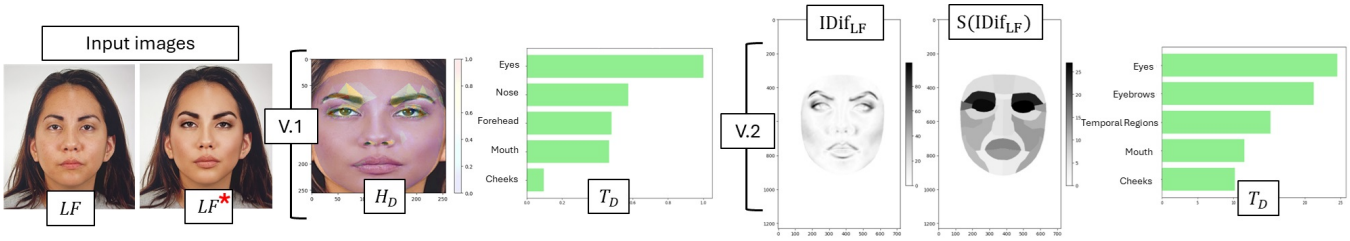

Figure 2: Left: Original (LF) and beautified (LF^*) faces. Middle: Output of the `Disclaimer Block` V.1 framework, showing a heatmap (H_D) and semantic analysis (T_D) of the facial changes. Right: Output of the `Disclaimer Block` V.2 framework proposed in this paper. It uses IDif_{LF} for more precise noise reduction and $\text{S}(\text{IDif}_{LF})$ for detailed regional analysis, enhancing the accuracy and granularity of the assessment of the filter on facial features (T_D). Example of a Latina = L Female = F.

achieved through (1) a heatmap H_D, which used Mediapipe tessellation to aggregate changes across predefined facial regions, offering a spatial overview of modifications, and (2) a semantic table T_D, which summarized the magnitude of changes across 13 macro areas of the face, based on Mediapipe landmarks and definitions from [22]. While effective in providing a general overview, this approach lacked the precision needed to capture pixel-level changes, as it smoothed over finer, localized modifications and was susceptible to noise introduced by variability in lighting and alignment.

Disclaimer Block V.2. To address these challenges, `DB` V.2 introduces several refinements. The semantic table (T_D) now leverages more granular facial regions, enabling detailed analysis of filter-induced changes in specific features. Additionally, H_D is replaced with IDif_A, a pixel-level difference map generated through precise alignment techniques (e.g., SIFT [53]) combined with a Gaussian blur to reduce noise. This approach provides a clear "footprint" of the filter's impact, mapping changes directly at the pixel level. Complementing IDif_A, `DB` V.2 introduced $\text{S}(\text{IDif}_A)$, a semantic map that visualizes pixel-level differences across all regions, bridging detailed spatial differences with interpretable semantic areas. These improvements enable a more accurate and comprehensive understanding of the filter's behavior.

From facial features to semantic feature vectors. Facial features are transformed into semantic vectors using FaceGen [28], which models facial morphology via PCA on 273 diverse 3D scans, yielding reproducible high-dimensional shape vectors that capture subtle structural variations. As a result, we transform the face into quantifiable vectors, $\overline{\mathbf{F}}^{(g,r)} \in \mathbb{R}^n$ and $\overline{\mathbf{F}}^{(g,r)*} \in \mathbb{R}^n$, where $\overline{\mathbf{F}}$ and $\overline{\mathbf{F}}^*$ denote the vectors corresponding to the original and beautified faces, respectively, and the superscript (g, r) denotes the gender and race group. These vectors enable a detailed comparative analysis, providing a structured way to quantify the aesthetic changes imposed by the filter, as illustrated in Figure 3. For further methodological details on how semantic feature vectors are derived from facial features with FaceGen, refer to Appendix C.

Average face per gender and race. The semantic feature vectors are computed from the average face within each racial and gender group to analyze general trends and patterns without relying on individual-level data, addressing practical and methodological limitations. Processing individual images is time-intensive and resource-demanding, particularly with manual facial analysis software like FaceGen, which requires manual processing of each image. Averaging provides a manageable and scalable method to analyze the broader impact of the filter while highlighting systemic biases and overarching trends that the filter may impose on distinct racial and gender groups. This approach is supported by [65], who demonstrated that face averaging effectively reveals systemic biases in automated systems. By revealing patterns such as biases in skin tone or facial shape, average faces offer insights beyond individual-level analysis, making them a robust tool for assessing group-level effects in line with the study's objectives. Furthermore, TikTok's architecture makes large-scale data extraction challenging. Working with average representations circumvents these technical barriers, enabling research while respecting the platform's limitations. By focusing on average faces, our study examines the broader sociotechnical implications of the filter while maintaining logistical efficiency.

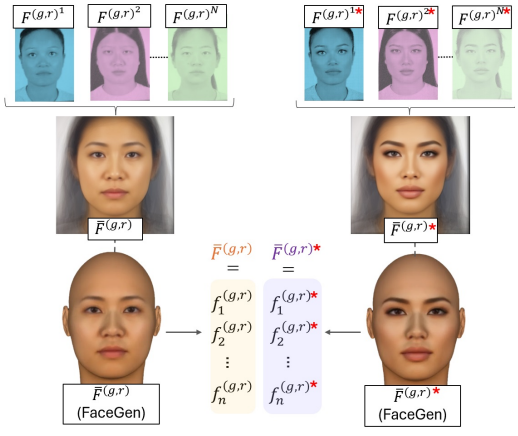

Figure 3: Creation of 2D average faces for groups categorized by race and gender, shown here for Asian females. Both original (F) and beautified (F*) averages were used to derive 3D reconstructions with FaceGen, yielding vector representations.

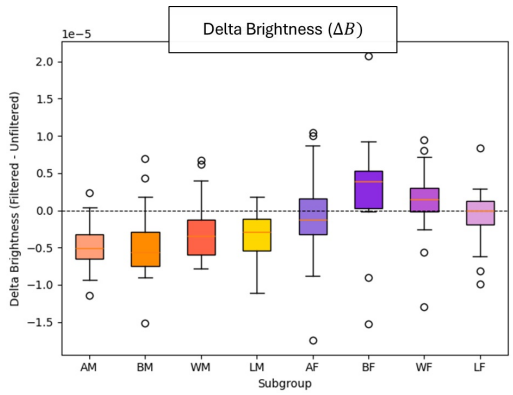

Figure 4: Boxplots illustrating variations in ΔB (Delta Brightness) across subgroups. Values below the zero line represent a decrease in brightness, while values above indicate an increase.

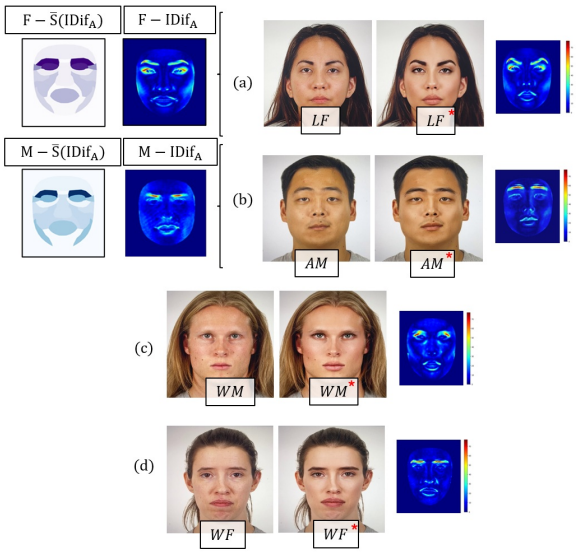

Figure 5: Gender differences in the transformations the filter applies based on an implicit gender classification of the input face. (a) Females (F): When the filter classifies the face as a female, its output enhances lips, cheeks, and eyes with makeup; (b) Males (M): When the filter classifies the face as male, it focuses on structural changes without makeup. Unfortunately, the filter might misclassify the input face, as illustrated in (c) and (d),where a man has applied the female transformation (c), and a woman has applied the male transformation (d).

Note that our approach does not provide an absolute measure of facial feature changes. Instead, it allows us to compare the relative impact of the filter across different groups by analyzing how each group's average facial features change after the filter's application. Previous studies have addressed the challenge of interpreting changes without an absolute reference by establishing reference baselines, such as using blurred images to quantify information loss [69]. Similarly, our analysis focuses on ranking facial features based on the degree of change observed in the average faces, providing a relative understanding of the filter's impact.

4.3 Results

4.4 RQ1: Does *Bold Glamour* brighten the faces?

We assess brightness modifications from TikTok's Bold Glamour filter to detect potential skin tone bias. Adapting the methodology of Riccio et al. [69], we convert the RGB values of 208 face images to the HSV space and extract the Value component. For each image, ΔB is defined as the mean brightness difference between the filtered and original versions. Figure 4 presents the distributions of ΔB by gender and race.

Because ΔB deviates from normality, we apply Wilcoxon signed-rank tests [87] to each gender–race subgroup. All male subgroups exhibit significant brightness decreases: Asian (AM, $p = 7.45\times10^{-7}$), Black (BM,$p = 6.032\times10^{-5}$), Latino (LM, $p = 1.11\times10^{-5}$) and White (WM, $p = 6.69 \times 10^{-3}$). Among females, Black (BF, $p = 0.0304$) and White (WF, $p = 0.0516$) faces show significant brightness increases, while Asian (AF) and Latina (LF) faces do not.

Given the small samples in each gender–race subgroup, we also aggregate all 208 images by gender and then compare mean and median ΔB using Welch's t-test (t = -6.409, $p = 1.50 \times 10^{-9}$) and the Mann–Whitney U test (U = 2168, $p = 1.436 \times 10^{-11}$). Male faces consistently darken (mean $\Delta B = -3.88 \times 10^{-6}$), and female faces consistently lighten (mean $\Delta B = 4.82 \times 10^{-7}$), confirming a robust gender-dependent brightness bias.

These results suggest that *Bold Glamour* systematically darkens male faces and lightens female faces across all races. Such a gender-dependent pattern suggests underlying algorithmic design choices that differentially impact users by gender, warranting further investigation into the filter's mechanisms and their implications for user representation and identity.

4.5 RQ2: Are the filter transformations dependent on gender and race?

Male vs Female Modifications. A close examination of IDif_A and $\text{S}(\text{IDif}_A)$ reveals that the *Bold Glamour* filter adapts noticeably based on the inferred gender of the subject. Because TikTok's API does not expose any explicit gender labels, we reconstructed this implicit classification by identifying two consistent transformation patterns—"Feminine Output" ($\text{F}-(\text{IDif}_A)$) and "Masculine Output" ($\text{M}-(\text{IDif}_A)$) —applied systematically to each face (see Figure 5)

In the case of faces that are implicitly classified as females by the filter, it transforms the lips, cheeks, and eyes, adding makeup to the eyelids and enhancing cheekbones with blush (Figure 5 (a)). Conversely, faces implicitly classified as males experience a vertical blush or shadow effect down the face, in contrast with the diagonal blush along the cheekbones in the feminine output. This vertical shading highlights a more angular, square facial structure, particularly by enhancing the jawline, reinforcing a masculine aesthetic

with a firmer jawline contour. In both genders, the filter increases the eyebrows' volume and intensity (Figure 5 (b)).

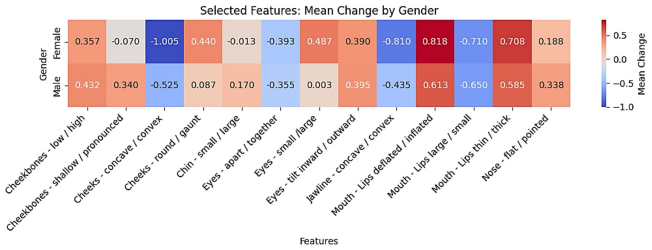

Figure 6: Average gender-specific modifications of facial features. The color represents the direction (red for increases and blue for reductions) and the intensity and magnitude of the changes. Observe how the filter's impact on different facial features depends on the inferred gender of the input image.

To quantify the observed gender differences, we compute the difference vector $D^{(g)}$ for each gender, where each component $d_i^{(g)}$ represents the change in the $i-th$ facial feature and is defined as $d_i^{(g)} = f_i^{(g)^*} - f_i^{(g)}$. Here, $f_i^{(g)}$ and $f_i^{(g)^*}$ are the components of the previously described feature vectors $\overline{F}$ and $\overline{F}^*$, representing the pre- and post-filter values, respectively. We select the 13 most commonly modified features, namely areas related to the size and shape of the eyes, cheekbones, chin, jawline, lips, and nose, as shown in Figure 6. The quantitative analysis reveals distinct gender-specific transformations, with women experiencing overall more pronounced changes, as reflected by an average absolute change of 0.491 compared to 0.379 for men. Lips are the most transformed feature across genders, with significant volumizing effects (+0.818 for women and +0.613 for men). Additionally, women exhibit dramatic softening and contouring of the cheeks (−1.005), along with slight enlargement and inward tilting of the eyes (+0.39) and a softened jawline (−0.810), creating a more delicate appearance. In contrast, men's transformations emphasize angularity and structure, with enhancements in the jawline (−0.435) and cheekbones (+0.433), reinforcing traditionally masculine features. These findings highlight the filter's alignment with conventional beauty standards: emphasizing softness and symmetry for women versus structure and definition for men. Furthermore, the gender classification is performed automatically by the filter, without the knowledge of the user, which also yields undesirable results in some failure cases, as illustrated in Figure 5 (c) and (d).

Sensitivity to facial features. As previously described, gender misclassifications by the filter can occur leading to the application of different transformations by the filter. Our analysis reveals that such misclassifications disproportionately affect female subjects. Specifically, 8 out of 26 Black Female images were misclassified (30.76%), compared to 3 out of 26 White Females (11.53%) and 2 out of 26 Latina Females (7.69%). Among male subjects, only 1 out of 26 White Men (3.84%) was misclassified.

To investigate the factors contributing to these gender misclassifications, we employed an input perturbation approach using [27],

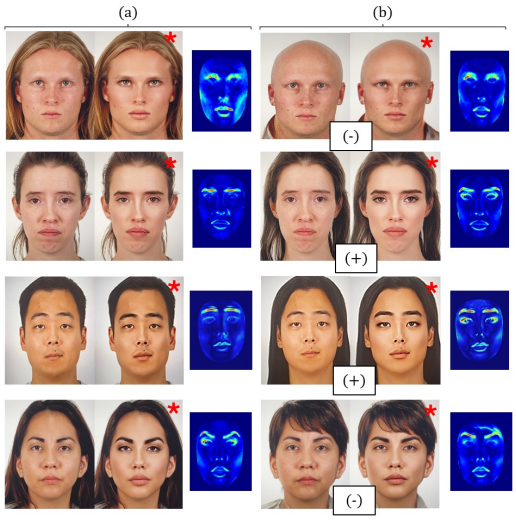

Figure 7: The effect of hair length changes on filter gender classification. In (a), original images show filter outputs. In (b), modifications to hair length are shown: The "(+)" and "(-)" symbols indicate whether extending or reducing hair length. The * represents the beautified version of the image

which allows for modifications in facial features such as hairstyles or facial expressions. While computationally expensive, this approach provided key insights into how the filter perceives faces. Hair length emerged as a significant factor, as many misclassified women had their hair tied back, while the only misclassified man had longer hair. Figure 7 illustrates the impact of hair length on the filter's implicit gender classification.

As shown in Figure 7.a and 7.b, lengthening (+) or shortening (−) the hair of misclassified women or men corrected the classification in 64.29% of cases. Specifically, 5 Black Female images, 1 White Female image, and 1 Latina Female image remained misclassified after adjustments, while the only misclassified male was correctly classified. Conversely, altering hair length in correctly classified cases had the opposite effect: lengthening the hair of correctly classified men caused the filter to apply transformations typically associated with women, while shortening the hair of correctly classified women resulted in transformations corresponding to men. These findings highlight the filter's sensitivity to hairstyle in its implicit gender classification and align with the findings of [1], which emphasized the role of hairstyle in gender perception and its impact on facial recognition systems.

4.6 RQ3: Does *Bold Glamour* apply a facial feature morphological alignment?

Next, we turn our attention to RQ3 and measure the impact of the filter on aligning facial features across different race category groups. We address this question by studying how the filter affects the distances between facial characteristics of various subgroups, assessing whether there is a trend towards aligning the facial features with those of a specific pre-filter race category group. For each gender, we compute $\Delta d_{(i,j)} = \left\|\overline{F}^{(i)^*} - \overline{F}^{(j)}\right\| - \left\|\overline{F}^{(i)} - \overline{F}^{(j)}\right\|$,

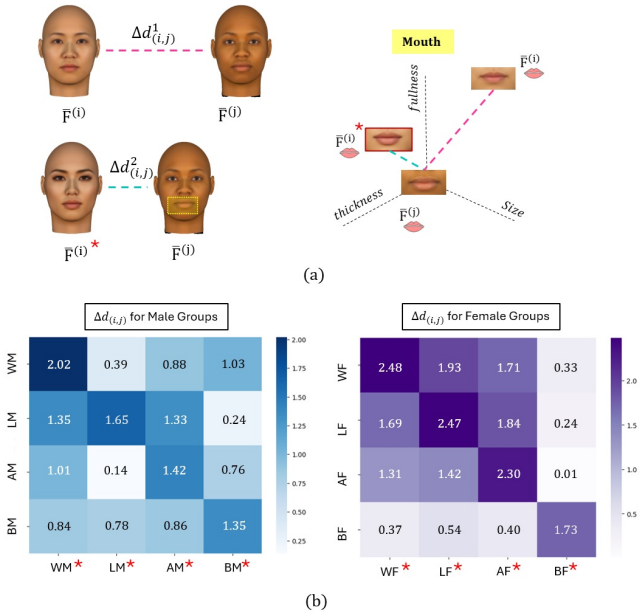

Figure 8: (a) Illustration of the metric $\Delta d_{(i,j)}$, which quantifies changes in Euclidean distance between post-filter features of group i and pre-filter features of group j, relative to the pre-filter distance. The mouth diagram provides an example of how alignment is measured through changes in specific features. (b) Heatmaps visualizing $\Delta d_{(i,j)}$ for females (purple) and males (blue). Lighter colors represent smaller changes, while darker colors indicate larger changes, highlighting patterns of facial feature alignment induced by the filter.

where $\overline{F}^{(i)}$ and $\overline{F}^{(i)^*}$ correspond to the facial features of a specific racial subgroup (denoted by the i and j subscripts) before and after (*) the application of the filter, as illustrated in Figure 8.a). This measure quantifies the change in Euclidean distance between the facial feature vectors of a racial group i post-filter application and a pre-filter racial group j.

The heatmaps in Figure 8.b depict the $\Delta d_{(i,j)}$ for both female (purple) and male (blue) faces per race subgroup. In the case of females, there is a noticeable trend where the features of White (WF), Latina (LF), and Asian (AF) Females shift towards those of Black Females (BF), *i.e.* the $\Delta d_{(i,j)}$ are the smallest when j = Black, for i = White, Latino, Asian. This indicates a convergence towards the Black Female facial features. In the case of male images, the features of White (WM) and Asian (AM) Males are closer to those of Latino Males (LM). Meanwhile, Black (BM) and Latino (LM) Males demonstrate mutual proximity, indicating a reciprocal alignment of their facial features post-filter. These insights could serve as a foundation for further studies. Note that the feature vectors only consider facial features and do not have any information regarding skin color.

5 Platform Policies

Following the analysis of the *Bold Glamour* filter, we briefly present next the platform policies and reported guidelines for filter creation to shed light on their alignment with the actual behavior of the filter.

5.1 Guidelines vs. Actual Practice

TikTok's community guidelines, which apply to everyone and everything on their platform, emphasize its focus on creating a welcoming, safe, and entertaining experience. [81]. Furthermore, TikTok's website provides best practices for creating TikTok effects, including a focus on **Diversity and Inclusivity**: *'When making effects, ensure it is inclusive of a variety of skin tones, hairstyles, facial features, body shapes, accessibility levels, and other differences. Avoid effects that reinforce negative or discriminatory stereotypes relating to gender, sexual orientation, age, ethnicity or disability"*; and **Positivity**: *'TikTok is a place for authentic, joyful, and uplifting content. Think about creating effects that empower creators to express themselves, explore their self-identity, and share their creativity in uniquely TikTok ways. Effects should promote a positive self-image and avoid reinforcing narrow and unattainable beauty ideals. For example, don't create effects that make users look thinner or which imply women must wear makeup – or that men can't"* [80]

These guidelines starkly contrast with the actual transformations applied by the *Bold Glamour* filter, as revealed in our analyses, which create effects where women are applied makeup and men are not, and where there are clear gender and racial dependencies in its effects. Such disparities reflect broader systemic patterns in the social media ecosystem, where filters have become pervasive tools that actively shape user experiences and perceptions. In fact, platforms like TikTok and Instagram not only host filters but also incentivize their creation through specialized tools—such as Meta's Spark AR and TikTok's Effect House—and financial rewards. ByteDance, for instance, launched the TikTok Effect Creator Rewards program in 2023, dedicating $6 million to reward creators for viral effects and filters [75]. This investment underscores the market value of AR filters and highlights the pressures these technologies impose on users to conform to prescriptive and lucrative beauty standards.

In this context, the users' faces and bodies –particularly of female users– are transformed into malleable assets, acting as sites of aesthetic labor that bolster platforms' profitability. This labor is deeply gendered, as noted by [24], who highlight the emphasis on "female attractiveness" and the pressures to adhere to specific appearance norms, often at significant personal cost. Such dynamics illustrate how visual appearance, mediated by rapidly advancing digital technologies, is increasingly entangled with social media's commodification of identity. As described by [84], "managing the body is the means by which women acquire and display their cultural capita." This process reinforces surveillance and normative commodification, subjecting feminine users to persistent pressures to align with idealized standards. The *Bold Glamour* filter exemplifies this trend, showcasing the intersection of technology, market-driven aesthetics, and gendered expectations in the digital age. These transformations reflect biases embedded in algorithmic design and the broader socio-economic structures perpetuating them.

5.2 Filtering Out The "Ugly"

TikTok is not unfamiliar with ambiguous techniques for aesthetic curation. Despite what the aforementioned policies claim, and as evidenced by both TikTok's investments into viral AR filters and our technical investigation into the *Bold Glamour* filter, TikTok actively produces and circulates beauty filters that reinforce "narrow and unattainable beauty standards" in direct contradiction with its own Best Practices for creating AR filters [80]. Moreover, our technical analysis of *Bold Glamour* reveals intersectional biases that appear systemic, emerging from the platform's infrastructure. Beauty filters like *Bold Glamour* are shaped and circulated within the confines of the platform's governance, a complex structure of moderation of the content that users produce and consume. Within this framework, aesthetic curation—a practice often imbued with intersectional bias—plays a critical role. Thus, a broader overview of the platform's (not always transparent) governance techniques is essential to fully understanding the phenomenon of beauty filters.

The results of our analysis on the *Bold Glamour* filter align closely with broader patterns of aesthetic curation and bias on social platforms. TikTok's emphasis on beauty standards is evident in its filters and algorithmic moderation practices, which actively curate content to favor certain aesthetic ideals. A notable example of this occurred in March 2020, when The Intercept reported that TikTok had internal pressures on content moderation teams to suppress posts from users deemed "too ugly, poor, or disabled" for the platform [3]. Moderators were instructed to filter content for TikTok's influential For You feed, which most users encounter when opening the app. The criteria for content removal were explicitly biased, listing reasons such as "abnormal body shape, "ugly facial looks, "too many wrinkles," and other "low quality" traits [3]. Moreover, videos filmed in "shabby and dilapidated" environments, such as slums or rural fields, were systematically hidden, while videos showcasing "rural beautiful natural scenery" were exempt from these restrictions.

These moderation practices are deeply intertwined with the effects of AR beauty filters, like *Bold Glamour*, which emphasize specific idealized beauty standards. For instance, the increase in brightness observed for Black and White Female faces in our analysis reflects a trend toward lighter, more "polished" appearances—aligning with the platform's historical suppression of certain content. As these filters and moderation policies show, platforms are shaping beauty ideals not only through the tools they provide users but also through the algorithmic curation of content that actively excludes diverse or non-conforming appearances, bodies, and identities ([2, 18, 40, 64, 70, 86]). Tik Tok's response to these revelations—acknowledging the primary goal of preventing bullying but dismissing the policies as outdated—suggests that these biases are deeply ingrained and may persist subtly. Our findings, particularly in the context of racial and gender differences, further shed light on how these platforms' design choices perpetuate exclusionary beauty standards and contribute to the systemic marginalization of certain appearances, reinforcing the pressures placed on users to conform to a specific, profitable image. Thus, beauty filters are not only tools of self-expression but also instruments of broader market-driven agendas that align with established norms in platform economies.

6 Discussion and Implications

Our findings underscore critical implications regarding the intersection of beauty filters, social media platforms, and systemic biases. These implications extend across individual, societal, and platform-level dimensions, highlighting the broader consequences of these technologies.

`Disclaimer Block`, *a Tool for Transparency.* We present an improved version of the `Disclaimer Block` introduced by [20] which enables us to analyze the transformations performed by beauty filters. This tool provides a detailed breakdown of the modifications applied, offering insights beyond the available minimal information, such as the filter's name and creator. If widely adopted, the `Disclaimer Block` could become an integral feature of social media platforms, either as a real-time tool accessible to users or as a mandatory documentation process for filter designers working with platforms like TikTok's Effect House. Alternatively, the `Disclaimer Block` could be part of an external platform where users would upload pre- and post-filter images to generate a visual "footprint" of the filter's effects. Such an interactive analysis could include quantitative data on the magnitude of changes, a heatmap illustrating affected areas, and text-based explanations generated by advanced language models [7, 17, 83]. By clearly communicating semantic and aesthetic alterations, this system would make the transformations more accessible and understandable, empowering users to evaluate how these digital tools shape their self-representation critically.

Implementing tools like the `Disclaimer Block` represents a crucial step toward promoting transparency and accountability in the design and application of beauty filters. By providing users with detailed insights into the specific changes performed by these filters, platforms could help mitigate the psychological and social pressures tied to unrealistic beauty standards. This transparency can shift the focus from passive consumption of idealized aesthetics to active, informed engagement with digital self-representation. Furthermore, such tools could encourage filter creators to reflect on the societal impacts of their designs, fostering a more ethical approach to digital beauty. Ultimately, widespread adoption of these strategies could help counteract the reinforcement of exclusionary norms and encourage the development of more inclusive and empowering digital environments.

Automatic Gender Inference. The *Bold Glamour* filter adjusts facial transformations based on inferred gender, aligning with traditional gender norms. For females, the filter enhances the lips, cheeks, and eyes, often adding makeup-like effects to emphasize softness and symmetry. In contrast, it performs structural changes for males, such as sharpening the jawline and emphasizing angularity, without applying makeup. Females generally experienced more pronounced changes than males, whose transformations focused on structural definition. A critical issue identified in this research is the filter's gender misclassification, particularly in the case of female faces. Black Females experienced the highest gender misclassification rates. These findings echo concerns raised by studies like Boulamwini and Gebru's *Gender Shades* [9], highlighting biases in models and questioning the diversity of datasets used for training. The lack of declared or user-controlled gender classification further

exacerbates the issue, as implicit and potentially flawed classifications enforce reductive norms and fail to respect diverse identities. This opacity, combined with reliance on simplistic visual cues such as hair length, raises serious ethical questions about the fairness and inclusivity of such systems.

Racial and gender dependencies. Our analyses reveal significant racial and gender biases in how beauty filters apply enhancements. The filter increases brightness for Black and White females, suggesting a skin-lightening effect that aligns with historical biases favoring lighter skin tones [69]. These transformations reinforce exclusionary beauty standards rooted in Westernized and heteronormative ideals, marginalizing those who do not conform, with potential negative impacts on self-perception and mental health, especially among underrepresented groups. Our study also reveals racial disparities in facial modifications. For women, the filter adjusted features of White, Latina, and Asian individuals to resemble traits associated with Black females, such as fuller lips. However, Black females experienced minimal changes, indicating selective aesthetic idealization. Among men, the filter made features of White and Asian males resemble those of Latino males, while Black and Latino males showed the least alteration post-filter. These findings suggest a dual process: the filter selectively emphasizes traits associated with Black individuals, such as fuller lips while upholding lighter skin tones as the dominant ideal. This mirrors broader beauty trends, including cosmetic surgery, where traits from non-white groups are selectively embraced, but Eurocentric standards, especially lighter skin, remain prioritized [44].

Commodification of Identity and Aesthetic Labor. AI-based augmented reality (AR) beauty filters blur the boundary between organic and digital self-representation, functioning as algorithmically codified simulations of identity [41]. They render gender performances instantaneous and computationally structured, aligning with Butler's conceptualization of gender performativity [5] [10], while serving as "normative discursive strategies" that influence user agency [82]. The commodification of identity, driven by the platform's economic priorities and the incentivization of viral content, positions beauty filters as tools for aesthetic and glamour labor, disproportionately pressuring feminine users to conform to appearance-based norms [24, 25, 85]. This dynamic intertwines visual appearance with social and economic capital, intensifying beauty surveillance [24] and encouraging bodily transformations that enhance platforms' market appeal. Despite claims that filter use is voluntary, normative governance subtly shapes user decisions through unacknowledged mechanisms, embedding harmful beauty standards that are neither transparent nor easily recognized [76]. These filters thus perpetuate exclusionary norms while serving the platforms' economic interests.

[5]Gender, according to Judith Butler, is understood as a socially constructed phenomenon that cannot be separated from the "cultural intersection" that both "produce and maintain" it ([10]). Butler argues that gender is not something one is, but repeated and ever-changing performances that align with societal norms and expectations. These performances, embedded in cultural, social, and historical contexts, collectively constitute gender identity, highlighting its fluid and performative nature rather than a fixed or inherent characteristic.

Beauty Filters as Technologies of Gender. Our analyses reveal that the beautification parameters of *Bold Glamour* align with and actively contribute to discriminatory gender constructs, echoing [11] and resonating with the concept of "technologies of gender" [51]. Drawing on Judith Butler's framework, our study emphasizes that gender norms are performatively enacted and inscribed on the body through external technologies and narratives, shaping and enforcing idealized body standards [10]. Beauty filters like *Bold Glamour* amplify these dynamics by providing immediacy and hyper-realism into the construction of gendered performance, fostering a personalized yet standardized beauty ideal that imposes biased aesthetic norms on users in opaque and potentially harmful ways to their mental and physical well-being. While promoting creativity, they also reveal an ambiguity in the governance of feminized bodies and identities on Western platforms, where these identities are treated both as valuable commodities and as subjects of normative control. The ambiguous and opaque governance of beauty filters, combined with their biases, limit their users' ability to critically engage with these technologies and understand their implications. Platforms like TikTok curate and monetize feminine representations, reinforcing heteronormative beauty standards. Addressing these issues requires a multifaceted approach that considers the platforms' context and governing models. Our analyses reveal and help mitigate the hidden biases embedded in these technologies, debunking corporate self-affirming narratives and critically engaging with their products.

Ethical and Regulatory Considerations. Beauty filters reflect broader cultural and technological shifts, where the boundary between physical and digital selves is increasingly blurred. This phenomenon compels users to navigate their identities through the lens of algorithmically mediated aesthetics. The privileging of certain appearances over others exacerbates social inequalities and shapes cultural norms in deeply inequitable and often invisible ways. Our research raises ethical concerns about the responsibilities of platforms in mitigating harm caused by biased filters. While these tools are marketed as empowering and creative, they can contribute to adverse psychological outcomes, such as body dissatisfaction and diminished self-worth. This calls for stricter regulatory oversight of platform practices, including transparency in algorithmic processes, explicit ethical guidelines for filter development, and mechanisms to counteract racial and gender bias.

7 Conclusion and Limitations

In this paper, we have analyzed TikTok's *Bold Glamour* filter within the broader context of platform governance, highlighting how gender and racial biases embedded in its AI-driven modifications reflect socially informed systemic design choices aimed at enhancing corporate profitability and desirability, often at the expense of user well-being. We find that the filter reinforces Eurocentric beauty standards, particularly affecting female users by promoting traits like fuller lips and pronounced contours associated with Black individuals while favoring lighter skin tones. These modifications perpetuate exclusionary beauty norms. We have developed and used an improved version of the `Disclaimer Block` tool, demonstrating how transparency tools can shed light on the transformations enacted by AR filters, empowering users with greater control

over their digital representation. Despite TikTok's public commitments to diversity and inclusion, the filter's design contradicts these claims by reinforcing narrow, biased aesthetics, exposing a need for transparency and accountability in filter development. Our study calls for future research to explore the impact of various filters on diverse identities, including non-binary and multiracial users, and to assess the effectiveness of transparency tools in fostering critical user engagement. These efforts aim to dismantle exclusionary norms and promote equitable, well-being-oriented digital spaces.

Our work, however, is not exempt from limitations. First, it has studied only one beauty filter as a case study, chosen for its popularity and because it was designed and deployed by a social platform, reflecting its aesthetic values, which contrast with their stated policies. While this filter choice provides valuable insights, it limits the generalizability of findings to other filters or platforms. Second, part of our analyses rely on average faces per race and gender due to ethical, technical, and logistical constraints in data collection, reducing granularity and individual variation. Additionally, we have used binary gender labels and four racial categories based on the Chicago Face Database, which do not account for non-binary, multiracial, and other underrepresented identities. While we use these labels for methodological consistency, our study does not engage in the reification of race and gender through facial images, recognizing these as socially constructed categories. Despite these limitations, our work provides a foundational contribution to understanding beauty filter biases and emphasizes the need for future research to adopt more inclusive and comprehensive approaches.

Acknowledgments

M.D. acknowledges support from the ARIAC project (No. 2010235), funded by the Service Public de Wallonie (SPW Recherche), and funding from the FNRS (National Fund for Scientific Research) for her visiting research at the ELLIS Alicante Foundation. C.C. acknowledges funding from the Berlin University of the Arts and the Weizenbaum Institute for the Networked Society in Berlin, supported by the German Federal Ministry of Education and Research (BMBF). N.O. acknowledges partial funding from a nominal grant provided by the ELLIS Unit Alicante Foundation, awarded by the Regional Government of Valencia in Spain (Convenio Singular signed with Generalitat Valenciana, Conselleria de Innovación, Industria, Comercio y Turismo, Dirección General de Innovación), as well as funding from the European Union under the HE ELIAS Grant Agreement 101120237. The views and opinions expressed herein are solely those of the authors and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA).

References

[1] V. Albiero, K. Zhang, M. C. King, and K. W. Bowyer. 2022. Gendered Differences in Face Recognition Accuracy Explained by Hairstyles, Makeup, and Facial Morphology. *IEEE Transactions on Information Forensics and Security* 17 (2022), 127–137. doi:10.1109/TIFS.2021.3135750

[2] Julia Alexander. 2019. YouTube moderation bots punish videos tagged as 'gay' or 'lesbian,' study finds. *The Verge* (Sept. 2019). https://www.theverge.com/2019/9/30/20887614/youtube-moderation-lgbtq-demonetization-terms-words-nerd-city-investigation

[3] S. Biddle, P. V. Ribeiro, and T. Dias. 2020. *Invisible Censorship.* The Intercept. https://theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination/ [Accessed: Nov 2024].

[4] Robert Booth. 2024. TikTok to block teenagers from beauty filters over mental health concerns. *The Guardian* (Nov. 2024). https://www.theguardian.com/technology/2024/nov/26/tiktok-to-block-teenagers-from-beauty-filters-over-mental-health-concerns

[5] Susan Bordo. 1989. *Reading the slender body.* University of California Press, Berkeley, Los Angeles, London, 83–112. doi:10.4324/9780429500527-33

[6] SUSAN BORDO. 2003. *Unbearable Weight: Feminism, Western Culture, and the Body* (1 ed.). University of California Press. http://www.jstor.org/stable/jj.8441705

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[8] L. Bullingham and A. C. Vasconcelos. 2013. 'The presentation of self in the online world': Goffman and the study of online identities. *Journal of Information Science* 39, 1 (2013), 101–112.

[9] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[10] Judith Butler and Gender Trouble. 1990. Feminism and the Subversion of Identity. *Gender trouble* 3, 1 (1990), 3–17.

[11] Sofia P Caldeira, Sander De Ridder, and Sofie Van Bauwel. 2018. Exploring the politics of gender representation on Instagram: Self-representations of femininity. *DiGeSt. Journal of Diversity and Gender Studies* 5, 1 (2018), 23–42.

[12] J. Constine. 2015. *Snapchat Acquires Looksery To Power Its Animated Lenses.* TechCrunch. https://techcrunch.com/2015/09/15/snapchat-looksery/ [Accessed: Nov 2024].

[13] Juraj Cug, Alina Tănase, Cristian Ionuț Stan, and Tanța Camelia Chitcă. 2022. Beauty filters for physical attractiveness: Idealized appearance and imagery, visual content and representations, and negative behaviors and sentiments. *Journal of Research in Gender Studies* 12, 2 (2022), 33–47.

[14] K. Dash. 2016. *There's a Reason Why You Love the Dog Filter on Snapchat.* Allure. https://www.allure.com/story/snapchat-dog-filter [Accessed: Nov 2024].

[15] Wes Davis. 2024. Meta is ending support for custom face filters in its apps. *The Verge* (Aug. 2024). https://www.theverge.com/2024/8/27/24229643/meta-spark-ar-effects-face-filters-shutdown-tiktok-snapchat

[16] G. Deleuze. 1992. Postscript on the Societies of Control. *October* 59 (1992), 3–7. https://www.jstor.org/stable/778828 [Accessed: Nov 2024].

[17] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*. Minneapolis, Minnesota, 1–2.

[18] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & Culture* 25 (2021), 700–732.

[19] Urban Dictionary. 2016. *The Hoe Filter.* Urban Dictionary. https://www.urbandictionary.com/define.php?term=the%20hoe%20filter Accessed: Nov 2024.

[20] Miriam Doh, Corinna Canali, and Anastasia Karagianni. 2024. Pixels of Perfection and Self-Perception: Deconstructing AR Beauty Filters and Their Challenge to Unbiased Body Image. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences*. ACM, Stockholm, Sweden, 349–353.

[21] Miriam Doh and Anastasia Karagianni. 2024. "My Kind of Woman": Analysing Gender Stereotypes in AI through The Averageness Theory and EU Law. arXiv:2407.17474 [cs.CY] https://arxiv.org/abs/2407.17474

[22] Miriam Doh, Caroline Mazini Rodrigues, Nicolas Boutry, Laurent Najman, Matei Mancas, and Hugues Bersini. 2024. Bridging Human Concepts and Computer Vision for Explainable Face Verification. arXiv:2403.08789 [cs.CV] https://arxiv.org/abs/2403.08789

[23] Miriam Doh, Caroline Mazini Rodrigues, N. Boutry, L. Najman, Matei Mancas, and Bernard Gosselin. 2025. Found in Translation: semantic approaches for enhancing AI interpretability in face verification. arXiv:2501.05471 [cs.CV] https://arxiv.org/abs/2501.05471

[24] Ana Elias, Rosalind Gill, and Christina Scharff. 2017. *Aesthetic labour: Beauty politics in neoliberalism.* Springer.

[25] A. S. Elias and R. Gill. 2018. Beauty surveillance: The digital self-monitoring cultures of neoliberalism. *European Journal of Cultural Studies* 21, 1 (2018), 59–77.

[26] J. Eshiet. 2020. *Real Me Versus Social Media Me: Filters, Snapchat Dysmorphia, and Beauty Perceptions among Young Women.* Doctoral Dissertation. California State University, San Bernardino. https://scholarworks.lib.csusb.edu/etd/1101 [Accessed: Nov 2024].

[27] FaceApp. 2024. *FaceApp: AI Face Editor.* Singular Inversions Inc. https://www.faceapp.com/ [Accessed: Oct 2024].

[28] FaceGen. 2024. *FaceGen 3D face modeling software.* FaceGen. https://facegen.com/index.htm [Accessed: Nov 2024].

[29] Rebecca Fribourg, Etienne Peillard, and Rachel Mcdonnell. 2021. Mirror, mirror on my phone: Investigating dimensions of self-face perception induced by augmented reality filters. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 470–478.

[30] Timnit Gebru. 2020. Race and gender. In *The Oxford Handbook of Ethics of AI*. Oxford University Press, Oxford, UK, 251–269.

[31] Tom Gerken, Liv McMahon, and Imram Rahman-Jones. 2025. What does Trump's executive order mean for TikTok? *BBC* (Jan. 2025). https://www.bbc.com/news/articles/clyng762q4eo

[32] Ysabel Gerrard and Helen Thornham. 2020. Content moderation: Social media's sexist assemblages. *New Media & Society* 22, 7 (2020), 1266–1286.

[33] R. Gill. 2007. *Gender and the Media*. Polity.

[34] Rosalind Gill. 2021. Changing the perfect picture: Smartphones, social media and appearance pressures. *City, University of London* (2021).

[35] D. Gimlin. 2002. *Body work: Beauty and self-image in American culture*. University of California Press.

[36] E. Goffman. 1979. *Gender advertisements*. Harper & Row.

[37] Google. 2024. *How YouTube Works*. Google. https://www.youtube.com/intl/en_us/howyoutubeworks/policies/community-guidelines/ [Accessed: Nov 2024].

[38] A. Gulati, M. Martinez-Garcia, D. Fernandez, M. A. Lozano, B. Lepri, and N. Oliver. 2024. What is Beautiful is Still Good: The Attractiveness Halo Effect in the era of Beauty Filters. *arXiv preprint arXiv:2407.11981* (2024).

[39] A Habib, T Ali, Z Nazir, and A Mahfooz. 2022. Snapchat filters changing young women's attitudes. *Annals of Medicine and Surgery (Lond)* 82 (2022), 104668. doi:10.1016/j.amsu.2022.104668

[40] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–35.

[41] D. J. Haraway. 1985. *A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century*. Routledge, 6.

[42] J. Hathaway. 2017. How did the dog face become Snapchat's "hoe filter"? *The Daily Dot* (2017). https://www.dailydot.com/unclick/snapchat-dog-filter-hoe-thot/ [Accessed: Nov 2024].

[43] Alison Hearn and Sarah Banet-Weiser. 2020. The beguiling: Glamour in/as platformed cultural production. *Social Media+ Society* 6, 1 (2020), 2056305119898779.

[44] C. J. Heyes. 2009. *"All Cosmetic Surgery is Ethnic": Feminism, Whiteness, and the Politics of Indignation*. Ashgate.

[45] Clara Isakowitsch. 2022. How augmented reality beauty filters can affect self-perception. In *Irish Conference on Artificial Intelligence and Cognitive Science*. Springer, 239–250.

[46] Ana Javornik, Ben Marder, Jennifer Brannon Barhorst, Graeme McLean, Yvonne Rogers, Paul Marshall, and Luk Warlop. 2022. 'What lies behind the filter?'Uncovering the motivations for using augmented reality (AR) face filters on social media and their effect on well-being. *Computers in Human Behavior* 128 (2022), 107126.

[47] R. Jenkins. 2010. *The 21st-century interaction order*. Routledge, 271–288.

[48] H. Jensen Schau and M. C. Gilly. 2003. We are what we post? Self-presentation in personal web space. *Journal of Consumer Research* 30, 3 (2003), 385–404.

[49] M. E. Kang. 1997. The portrayal of women's images in magazine advertisements: Goffman's gender analysis revisited. *Sex roles* 37 (1997), 979–996.

[50] Kevinvq2. 2017. *Dog Filter*. knowyourmeme. https://knowyourmeme.com/memes/dog-filter [Accessed: Nov 2024].

[51] TERESA DE LAURETIS. 1987. *Technologies of Gender: Essays on Theory, Film, and Fiction*. Indiana University Press. http://www.jstor.org/stable/j.ctt16gzmbr

[52] R. Z. Leeat, N. Shnabel, and P. Glick. 2019. The "prescriptive beauty norm" reflects a desire to enhance gender hierarchy and contributes to social policing of women and employment discrimination practices known as the "beauty tax.". *Journal of Personality and Social Psychology* (2019). Available at: Harvard Kennedy School | Gender Actional Portal.

[53] D. G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 2 (2004), 91–110.

[54] D. S. Ma, J. Correll, and B. Wittenbrink. 2015. The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data. *Behavior Research Methods* 47 (2015), 1122–1135. doi:10.3758/s13428-014-0532-5

[55] M. Malone Kircher. 2016. How to Get the Secret Dalmatian Snapchat Filter Maybe Inspired by Kim Kardashian. *New York Magazine* (2016). https://tinyl.io/BjRQ [Accessed: Nov 2024].

[56] Alice E Marwick. 2015. Instafame: Luxury selfies in the attention economy. *Public culture* 27, 1 (75) (2015), 137–160.

[57] Meta. 2024. *Terms and Policies | Community Guidelines | Instagram Help Center*. Meta. https://help.instagram.com/477434105621119 [Accessed: Nov 2024].

[58] Ramona Mihăilă and Ludmila Branişte. 2021. Digital semantics of beauty apps and filters: big data-driven facial retouching, aesthetic self-monitoring devices, and augmented reality-based body-enhancing technologies. *Journal of Research in Gender Studies* 11, 2 (2021), 100–112.

[59] Lauren A. Miller. 2024. Instagram has announced it will be removing beauty filters – but the damage is done. *The Conversation* (2024). https://theconversation.com/instagram-has-announced-it-will-be-removing-beauty-filters-but-the-damage-is-done-238582

[60] J. S. Mills, A. Shannon, and J. Hogue. 2017. *Beauty, Body Image, and the Media*. IntechOpen. doi:10.5772/intechopen.68944

[61] N. Mirabet-Herranz, C. Galdi, and J.-L. Dugelay. 2024. Facial Biometrics in the Social Media Era: An in-Depth Analysis of the Challenge Posed by Beautification Filters. *IEEE Transactions on Biometrics, Behavior, and Identity Science* (2024).

[62] A. Monea. 2023. *The digital closet: How the internet became straight*. MIT Press.

[63] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press. http://www.jstor.org/stable/j.ctt1pwt9w5

[64] Lauren Olson, Emitzá Guzmán, and Florian Kunneman. 2023. Along the margins: Marginalized communities' ethical concerns about social platforms. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*. IEEE, 71–82.

[65] Kentrell Owens, Erin Freiburger, Ryan Hutchings, Mattea Sim, Kurt Hugenberg, Franziska Roesner, and Tadayoshi Kohno. 2024. Face the Facts: Using Face Averaging to Visualize Gender-by-Race Bias in Facial Analysis Algorithms. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 1101–1111.

[66] Phillip Ozimek, Semina Lainas, Hans-Werner Bierhoff, and Elke Rohmann. 2023. How photo editing in social media shapes self-perceived attractiveness and self-esteem via self-objectification and physical appearance comparisons. *BMC psychology* 11, 1 (2023), 99.

[67] Aaron Pellish and Brian Stelter. 2025. TikTok shuts down in the United States hours ahead of a ban. *CNN* (Jan. 2025). https://edition.cnn.com/2025/01/18/business/trump-tiktok-ban/index.html

[68] R. M. Perloff. 2014. Social media effects on young women's body image concerns: Theoretical perspectives and an agenda for research. *Sex roles* 71 (2014), 363–377.

[69] Piera Riccio, Julien Colin, Shirley Ogolla, and Nuria Oliver. 2024. Mirror, Mirror on the Wall, Who Is the Whitest of All? Racial Biases in Social Media Beauty Filters. *Social Media+ Society* 10, 2 (2024), 20563051241239295.

[70] Piera Riccio, Thomas Hofmann, and Nuria Oliver. 2024. Exposed or Erased: Algorithmic Censorship of Nudity in Art. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[71] Piera Riccio, Bill Psomas, Francesco Galati, Francisco Escolano, Thomas Hofmann, and Nuria Oliver. 2022. OpenFilter: a framework to democratize research access to social media AR filters. *Advances in Neural Information Processing Systems* 35 (2022), 12491–12503.

[72] A. Ruggeri. 2023. *The problems with TikTok's controversial "beauty filters"*. BBC Online. https://tinyl.io/BjRX [Accessed: Nov 2024].

[73] Tate Ryan-Mosley. 2021. *How digital beauty filters perpetuate colorism*. MIT Technology Review. https://www.technologyreview.com/2021/08/15/1031804/digital-beauty-filters-photoshop-photo-editing-colorism-racism/ [Accessed: Nov 2024].

[74] Tate Ryan-Mosley. 2023. Why Meta is getting sued over its beauty filters. *MIT Technology Review* (Oct. 2023). https://www.technologyreview.com/2023/10/30/1082628/why-meta-is-getting-sued-over-its-beauty-filters/

[75] M. Sato. 2023. *TikTok will pay creators of viral filters and effects*. The Verge. https://tinyl.io/BjRV [Accessed: Nov 2024].

[76] C. Shane. 2023. *Augmented Reality Beauty Filters Are Changing the Face of Social Media*. Wired. https://tinyl.io/BjRd [Accessed: Nov 2024].

[77] R. G. Simmons and F. Rosenberg. 1975. Sex, sex roles, and self-image. *Journal of youth and adolescence* 4, 3 (1975), 229–258.

[78] D. Smith, J. Protevi, and D. Voss. 2023. *Gilles Deleuze*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2023/entries/deleuze/ The Stanford Encyclopedia of Philosophy (Summer 2023 Edition) [Accessed: Nov 2024].

[79] Charles J Stivale. 2005. *Gilles Deleuze: Key Concepts*. McGill-Queen's University Press. http://www.jstor.org/stable/j.cttq486p

[80] TikTok. 2024. *Best Practices for Creating TikTok Effects | Effect Guidelines*. TikTok. https://effecthouse.tiktok.com/learn/guides/general/best-practices-for-creating-tiktok-effects [Accessed: Nov 2024].

[81] TikTok. 2024. *Community Guidelines | TikTok*. TikTok. https://www.tiktok.com/community-guidelines/en [Accessed: Nov 2024].

[82] J. Van Dijck. 2007. *Mediated memories in the digital age*. Stanford University Press, Stanford, CA, USA.

[83] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (2024), 186345.

[84] A. Winch. 2015. Brand intimacy, female friendship and digital surveillance networks. *New Formations* 84, 84-85 (2015), 228–245.

[85] Elizabeth Wissinger. 2015. *This year's model: Fashion, media, and the making of glamour*. NYU Press, New York, USA.

[86] Alice Witt, Nicolas Suzor, and Anna Huggins. 2019. The rule of law on Instagram: An evaluation of the moderation of images depicting women's bodies. *University of New South Wales Law Journal, The* 42, 2 (2019), 557–596.

[87] Robert F Woolson. 2005. Wilcoxon signed-rank test. *Encyclopedia of Biostatistics* 8 (2005).

[88] L. Yltävä. 2024. *Beauty shopping on social media - statistics & facts*. Statista. https://www.statista.com/topics/12547/beauty-shopping-on-social-

media/#topicOverview [Accessed: Jul 2024].

A Updates on Governmental Acts

The future of TikTok in the U.S. remains uncertain. In January 2025, the U.S. Supreme Court upheld a bipartisan law, signed by President Joe Biden in April 2024, banning the app unless ByteDance, its Chinese parent company, sold it to a U.S. or allied buyer [67]. TikTok briefly went offline after the Court denied ByteDance's appeal to overturn the ban but was reinstated a day later, with TikTok thanking newly inaugurated President Donald Trump for his efforts. One of Trump's first acts in office was an executive order granting TikTok a 75-day reprieve, instructing the attorney general not to enforce the ban while exploring solutions. Trump has proposed a potential joint venture, suggesting a 50-50 ownership split between ByteDance and the U.S., though details remain unclear [31]. What is certain, however, is that the current debate surrounding TikTok predominantly revolves around data and market concerns rather than issues related to users' safety and health.

B Improvements from `Disclaimer Block` V1 to V2

Background: The first version of the `Disclaimer Block` was introduced in [20] to improve transparency in the use of beauty filters by providing clearer information about the modifications applied to users' faces (Figure 9). The main concerns analyzed in the work were: (1) While beauty filters are often marketed as tools for minor modifications, they can significantly reshape facial features (*"Beyond Simple Aesthetic Adjustments"*). (2) The naming conventions of beauty filters often use vague or overly positive descriptors, such as "Prettiest" or "Pure Eyes," which do not accurately reflect the specific changes they introduce. The disconnect between a filter's name and its actual effects can lead to unintentional misrepresentation and limit users' ability to make informed choices about their digital self-representation (*"Nominal Ambiguity and Presentation Autonomy"*).

To address these issues, the Disclaimer Block aims to provide users with a detailed understanding of the specific changes applied by beauty filters.

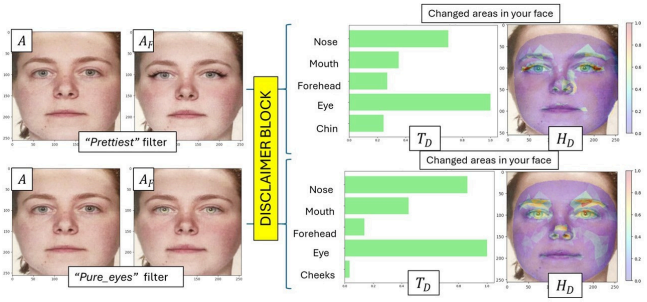

Figure 9: The `Disclaimer Block` V1. Examples of the "Prettiest" filter and "Pure Eyes" filter from TikTok. Image from [20].

In refining the `Disclaimer Block` from V1 to V2, we focused on improving data acquisition and image processing techniques to achieve higher accuracy, greater precision, and deeper insights into

the effects of beauty filters. Below, we outline the enhancements introduced in V2 and how they address the limitations of V1.

Data Acquisition Enhancements: In `Disclaimer Block` V1, the data acquisition process relied on videos recorded for each subject, capturing their face first without any filter and then applying the beauty filter. These videos were used to generate multiple frames for each subject, and the analysis was performed by calculating an average difference image across corresponding frames (e.g., frame 1 pre-filter vs. frame 1 post-filter). While this approach reduced noise caused by variations in lighting, positioning, and subject movement, it also smoothed out finer, localized changes introduced by the filter and introduced potential artifacts.

In V2, the acquisition process shifted to directly comparing single high-quality frames of unfiltered and filtered images for each individual, eliminating the need for intra-subject averaging. This change avoided the loss of detail and artifacts seen in V1, resulting in a cleaner and more accurate dataset.

To ensure consistency across subjects, V2 employed a controlled setup with a phone mounted on a selfie ring light, providing stable lighting and positioning during data collection. This addressed the variability issues present in V1, where differences in lighting and positioning between frames often required additional corrections.

By focusing on static images rather than averaged frames, V2 achieved a cleaner and more precise dataset.

From H_D to IDif_A. In `Disclaimer Block` V1, the image processing pipeline had to address the inherent noise in the dataset caused by variability in lighting, positioning, and alignment across frames. To manage these inconsistencies, V1 relied on the Mediapipe tessellation to divide the face into predefined regions. Each tessellation segment represented an area of the face, and changes were aggregated as the average magnitude of differences within these regions rather than at the pixel level. This approach resulted in a heatmap (H_D) that was visually intuitive but lacked precision, as it smoothed over finer, localized changes introduced by the filter.

In V2, the improved data acquisition process provided a much cleaner and more consistent dataset with aligned images and stable lighting conditions. Building on this, we introduced a more precise alignment technique using the Scale-Invariant Feature Transform (SIFT) algorithm [53]. SIFT accurately identifies and matches key points between filtered and unfiltered images, enabling us to apply a homography transformation to correct any residual shifts, rotations, or distortions. We applied a Gaussian blur to the different images to further refine the results, reducing noise and removing minor artifacts. This improvement eliminated the need for region-based aggregation, allowing us to move from area-level analysis to direct pixel-by-pixel comparisons.

This step, combined with the improved alignment process, enabled us to create IDif_A. Unlike the regional aggregation used in V1, IDif_A captures the precise "footprint" of the filter at the pixel level, providing a much more detailed and accurate representation of changes. This approach ensures that the heatmap is no longer an approximation of changes across larger regions but a direct mapping of the filter's impact on every face pixel

Semantic Analysis Improvements: T_D and $\text{S}(\text{IDif}_A)$. In `Disclaimer Block` V1, semantic analysis was performed using T_D, a table summarizing the average magnitude of changes introduced by the filter across 13 predefined macro areas of the face. These areas were segmented based on Mediapipe landmarks and defined by [22]. The purpose of T_D was to complement the spatial heatmap (H_D) by providing a semantically interpretable summary of changes in broader facial regions, such as the eyes, cheeks, and mouth. While H_D visualized changes spatially usingMediapipe'ss tessellation, T_D offered a high-level overview of how different regions were affected by the filter, aiding interpretability.

In V2, we retained T_D but refined the underlying segmentation process. Instead of relying on the original 13 macro areas, we introduced a more granular set of 30 semantic masks based on Mediapipe landmarks, previously introduced by [23]. This increased granularity allowed for a more detailed analysis of the filter's effects on specific facial features. For example, the cheeks, previously treated as a single region, were divided into smaller subregions, enabling a more precise understanding of how the filter modifies distinct areas.

While T_D remains a key component for summarizing changes semantically, in V2, we extended the use of these semantic masks beyond the table. The masks are now directly employed to generate IDif_A, a semantic map that visualizes changes across the 30 regions. This map bridges the gap between detailed spatial differences and semantically interpretable areas, offering a more intuitive understanding of the filter's behavior. This level of granularity and structure, absent in V1, significantly enhances the explanatory power of our analysis, enabling deeper insights into the filter's behavior.

C From Faces to Semantic Feature Vectors with FaceGen

To analyze the impact of the *Bold Glamour* filter, we used the FaceGen software [28], which offers a comprehensive set of facial features for modeling and analysis. FaceGen allows for the creation of 3D reconstructions from 2D images and provides calibrated numerical values for a wide range of facial features. Among the extensive set of features available in FaceGen, we selected those most commonly targeted by beauty filters, focusing on attributes that directly influence symmetry, balance, and the perceived attractiveness of the face. These features are categorized in Table 1, organized by primary facial attributes and specific calibration adjustments:

The analysis began by creating average 2D images of faces pre- and post-filter, categorized by race and gender groups (Figure 3). These average images were reconstructed into 3D models using FaceGen, which allowed us to extract and represent the selected features numerically. Each feature was assigned a calibrated value, enabling direct comparisons between the pre and post-filter conditions.

To understand how the filter modifies facial features across different ethnic groups, we modeled each face as a vector of features before and after applying the filter. Let $\overline{F}^{(g,r)} \in \mathbb{R}^n$ and $\overline{F}^{(g,r)*} \in \mathbb{R}^n$ represent the pre-filter and post-filter feature vectors, respectively. Each vector $\overline{F}$ contains n components, each representing a specific facial feature. For instance, the vector for (g, r) group, where g and r stand for gender-race, before and after filtering can

Table 1: Facial features selected for quantitative analysis in FaceGen software to assess the impact of the *Bold Glamour* filter. Features are categorized by primary facial attributes and specific calibration adjustments, allowing for a detailed examination of how the filter alters individual aspects of facial structure across groups.

Face Feature	Feature Specific
Eye	• small/large • apart/together • tilt inward/outward
Cheekbones	• low/high • shallow/pronounced
Cheeks	• concave/convex • round/gaunt
Chin	• small/large
Jawline	• concave/convex
Mouth	• Lips deflated/inflated • Lips large/small • Lips thin/thick
Nose	• flat/pointed

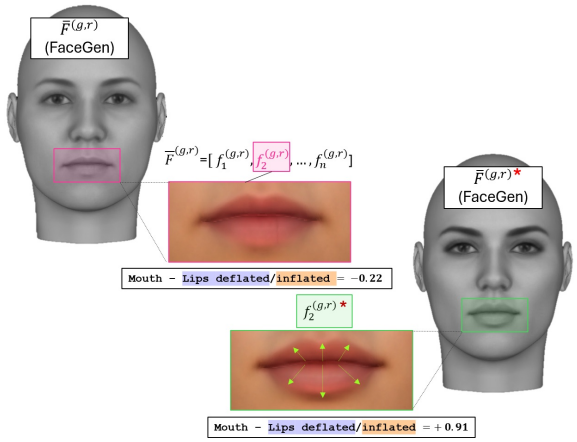

Figure 10: Semantic calibration of the "Lips deflated/inflate" feature for the average faces $\overline{F}$. The unfiltered face has a score of -0.22, indicating lips closer to the "deflate" pole. After applying the filter, the score shifts to +0.91, showing movement toward the""inflate"" pole. In the example, the average face for the White Female group is taken into account

be written as:

$$\overline{F}^{(g,r)} = \left[f_1^{(g,r)}, f_2^{(g,r)}, \ldots, f_n^{(g,r)}\right]$$
$$\overline{F}^{(g,r)*} = \left[f_1^{(g,r)*}, f_2^{(g,r)*}, \ldots, f_n^{(g,r)*}\right]$$

Each element in the vector $\overline{F}$ corresponds to a specific facial characteristic, such as eye size or lip fullness.

The calibrated values extracted by FaceGen allow us to quantify changes introduced by the filter along a semantic continuum defined by the poles of each selected feature. For example, the feature "Lips deflated/inflated" positions lips along a scale where negative values correspond to the "deflated" pole (indicating less fullness), and positive values correspond to the "inflated" pole (indicating increased fullness) (Figure 10). This calibration system provides a structured framework to measure the direction and magnitude of changes the filter applies.