Sensations into Stereotypes: Large-Scale Measurement of Cross-Sensory Bias in Text-to-Image Generation

Fangjian Shen Zifeng Liang Yifan Zeng Wushao Wen[‡]
School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, Guangdong, China
{shenfj3, liangzf5, zengyf53}@mail2.sysu.edu.cn wenwsh@mail.sysu.edu.cn

Abstract

Human perception relies on crossmodal correspondences, systematic links between different sensory modalities. As text-to-image (T2I) models increasingly serve as aesthetic infrastructure, they risk codifying cultural biases embedded in sensory language. We analyze the mapping of cross-sensory bias in mainstream T2I models, examining how gustatory, tactile, auditory, and olfactory adjectives map onto visual attributes. Using a modality × carrier design, we evaluate five bias dimensions in 6,000 images via automated VLM-based assessment: color, demographic, entity, situational, and layout. Our results reveal widespread cross-sensory homogenization, with models projecting abstract sensations onto a narrow set of cultural prototypes. This study presents a framework for quantifying cross-sensory bias in T2I models and offers tools for auditing and mitigating their broader cultural and social impact.

Keywords: Crossmodal correspondences; Generative AI; Text-to-Image Models; Cross-sensory Bias; Human-AI alignment

Introduction

Human perception is richly cross-sensory, as we intuitively associate "sweet" with pink or "prickly" with angularity through systematic links known as crossmodal correspondences (Di Stefano & Spence, 2024; Spence, 2011). These regularities provide a powerful entry point into the organization of conceptual representation in human cognition (Barsalou, 2008). In parallel, recent advances in text-to-image (T2I) diffusion models, such as Stable Diffusion (Rombach et al., 2022), FLUX.1 (Greenberg, 2025), Nano Banana (Raisinghani, 2025), and many others, have enabled the generation of high-fidelity and contextually relevant images. These models are now routinely embedded in advertising workflows, fashion and product design, e-commerce content creation, and media illustration (H. Chen et al., 2025; Leng et al., 2025; Shen et al., 2025), thereby reshaping content production paradigms. They can respond to fine-grained linguistic prompts and modifiers, including sensory adjectives that describe gustatory, tactile, auditory, and olfactory. Because these systems are trained on large-scale web-based corpora, they likely absorb and amplify existing biases in how sensory language is visually depicted (Afreen et al., 2025). However, most current research focuses on technical aspects such as text-image alignment and image quality while paying relatively little attention to the perceptual and cognitive profiles of these models. Consequently, the cross-sensory dimension of these systems, specifically how non-visual sensory language maps onto visual attributes like color and composition, remains largely unexplored.

Figure 1: Example of cross-sensory and demographic bias in T2I models. The top row shows samples from several mainstream T2I models, which consistently generate young female rhythmic-gymnastics outfits with very similar colors and silhouettes. The bottom row illustrates more varied and less stereotyped interpretations of "rhythmic wear", including different body types, genders, and clothing styles.

Such cross-sensory biases may carry over into downstream applications and have potentially harmful societal consequences (Girrbach et al., 2025; Zhang et al., 2024). Culturally, as T2I models are increasingly adopted in creative and marketing pipelines, they are beginning to function as an emerging aesthetic infrastructure. They can consolidate particular links between sensory language and visual form, so that contingent associations become taken-for-granted defaults of visual common sense (Smith & Southerton, 2025). Over time, this process risks compressing local aesthetic practices and contributing to a broader homogenisation of how sensory experiences are visually represented across cultures (H. Chen, 2024; Varghese, 2025). Socially, if adjectives such as sweet, soft, or fragrant are, to a large extent, rendered as young, feminine figures, whereas terms such as bitter, freezing, or musty are more often realised as adult, masculine figures, the

[‡]Corresponding author.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

model would be tying particular sensory qualities to gendered appearances. In addition, some sensory adjectives may become racialised in practice; for instance, prompts involving loud might disproportionately produce Black or dark-skinned subjects, whereas prompts with melodic may predominantly yield white, Western-looking women. Such patterns risk reinforcing existing stereotypes about which social groups are expected to embody which kinds of sensory experience (Naik & Nushi, 2023). Investigating cross-sensory biases in T2I models is therefore not only a matter of model assessment, but also a necessary step toward understanding and, ultimately, mitigating their broader cultural and social impact.

To address this gap, we extend crossmodal correspondence research to contemporary T2I models by introducing a systematic framework to analyze five recurring forms of bias: color, demographic, entity, situational, and layout bias. Our study employs a controlled modality × carrier design, pairing adjectives from four sensory modalities (gustatory, tactile, auditory, and olfactory) with five visual contexts such as 3D spheres, rooms, and clothing. This structure allows us to systematically probe how the same sensory term is visualized across varied carriers while also revealing how different terms manifest within a consistent visual frame. To operationalize this analysis, we design a set of indicators to characterize how these multi-dimensional biases are instantiated in the outputs. Finally, we develop an automated evaluation procedure using a vision–language model (VLM) to annotate images along these indicators, enabling a large-scale, quantitative, and consistent analysis of cross-sensory biases in T2I systems.

Contributions. The contributions can be summarized as:

- To our knowledge, we are the first to systematically study crossmodal correspondences in T2I generative models.
- We develop a carefully designed modality × carrier prompting scheme and apply it to several mainstream T2I models to generate over 6,000 images for analysing similarities and differences in their cross-sensory bias patterns.
- We define five types of cross-sensory bias in T2I models and develop a VLM-based metric that enables automated quantitative assessment of these biases.

Related Work

Crossmodal Correspondences in Human Perception. Crossmodal correspondences refer to systematic associations between features in different sensory modalities—for example, linking high-pitched sounds with bright colors (Knöferle & Spence, 2012), or rounded shapes with sweet tastes. (Spence, 2011) identifies several possible origins, including environmental regularities, linguistic mediation, shared affect, and neural structure. These associations have been demonstrated in both controlled experiments using geometric stimuli (Wan et al., 2014) and applied contexts such as food design and packaging (Cardello et al., 2019; Velasco et al., 2016). Studies report consistent mappings: sweet with pink and round shapes, bitter with black and angular forms, sour with green, and salty with white (Spence & Levitan, 2021; Velasco et al., 2015). Recent work further distinguishes between language-mediated and perceptually grounded correspondences (Guedes et al., 2023; Ohtake et al., 2023), offering a foundation for comparing human and model behavior.

Biases in Text-to-Image Models. Prior work reveals systematic patterns of gender, racial, and occupational bias (Naik & Nushi, 2023): prompts involving caregiving or aesthetic terms often yield young, slim, white women, while technical or labor-related prompts tend to produce male or darker-skinned figures (Bianchi et al., 2022; Girrbach et al., 2025; Schneider & Hagendorff, 2024). Cultural defaults are also common (Nayak et al., 2025), with Western aesthetics dominating outputs even for culturally neutral prompts (Elsharif et al., 2025). While some studies explore mitigation strategies, most focus on explicit demographic attributes. Few studies have explored how sensory language such as sweet or spicy may give rise to stereotyped visual representations. We address this gap by analysing how such terms shape bias in T2I systems.

Crossmodal Mapping in Generative Models. Recent work has begun probing crossmodal associations in generative systems. (Alper & Averbuch-Elor, 2023) examined the bouba–kiki effect in Stable Diffusion and found sound–shape pairings consistent with human data. (**motoki2024color**) showed that ChatGPT could associate taste adjectives with expected shapes and colors. Other studies find that CLIP exhibits weaker or less consistent crossmodal mappings (Kouwenhoven et al., 2025; Verhoef et al., 2024). While these efforts offer early insights, they typically focus on isolated modality pairs. To our knowledge, no study has examined how non-visual sensory adjectives such as soft, bitter, or fragrant are rendered by T2I models across visual contexts.

Define Cross-Sensory Bias in Models

This section defines cross-sensory bias in T2I models. Since Spence's work on crossmodal correspondences (Spence, 2011), cross-sensory mappings have become a framework in cognitive science and experimental psychology, and they increasingly inform research in design, marketing, and sensory studies. However, only a few studies apply this lens to AI, and almost none focus on multimodal generative models.

Work on crossmodal perception and so-called "weak synaesthesia" shows that associations across sensory modalities follow stable patterns in human cognition (Lee & Spence, 2022; Spence et al., 2015). Research on social stereotypes and gender schemas documents how abstract traits are unevenly mapped onto particular social bodies (Eagly et al., 2020; Sczesny et al., 2018), such as specific genders, races, and body types. Prototype theory in categorisation argues that concepts are organised around a small set of salient exemplars (Rosch et al., 1976) rather than treating all instances as equally representative. Script and schema theories, together with frame semantics, describe how linguistic cues activate default situations and event structures (Fillmore et al., 2006; Schank & Abelson, 2013). Visual semiotics and visual grammar further show that camera angle, distance, and composition

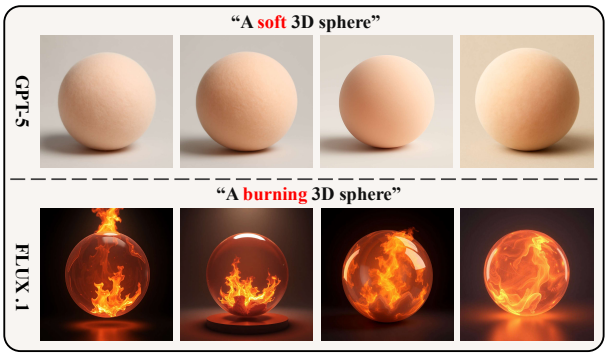

Figure 2: Example of Color Bias.

Figure 3: Example of Demographic Bias.

function as systematic resources for encoding power, emotion, and interpersonal stance in images (Kress & Van Leeuwen, 2020). Building on these strands of work, we use them as a conceptual basis to define and analyse cross-sensory bias in T2I models in the remainder of this section.

Color Bias: Models shrink cross-sensory language into a small set of color choices. Adjectives such as sweet, spicy, or salty are repeatedly mapped to pink pastels, saturated reds, and stark whites. Over time, these pairings turn historical and cultural links between taste and color into new visual defaults and make other readings less visible. For example, burning is almost always rendered as red flames, even though flames can also appear purple, blue, or other colors.

Demographic Bias: Models often align sensory adjectives with particular gendered and racialised bodies. As shown in Figure 3*, prompts such as "sweet wear" predominantly yield young, light-skinned, Western-looking women, whereas "loud wear" more often produces Black or dark-skinned subjects in highly saturated, patterned outfits. "Bitter wear" tends to return older-looking, masculine figures with worn, rugged styling. Repeated across prompts, these patterns turn sensory qualities into demographic cues and risk reinforcing stereotypes about who is expected to embody particular sensations.

*Ethics Statement: All visuals are model outputs from neutral prompts, included to reveal—NOT endorse—algorithmic bias.

Figure 4: Example of Entity Bias.

Entity Bias: Open-ended sensory prompts are often narrowed to a small set of prototypical entities in generative outputs. Figure 4 shows that "a fresh dish" repeatedly converges on a salad-bowl template, typically featuring sliced eggs, tomatoes and cucumbers. "A fragrant shop window display" defaults to a perfume-style window with glass bottles and floral arrangements, and "a prickly room" is rendered as a room filled with indoor cacti or succulents. These recurrent defaults turn sensory adjectives into triggers for a few familiar icons and reduce the range of objects and interpretations the model produces.

Situational Bias: Sensory language is repeatedly realised through a narrow set of scene schemas. As shown in Figure 5, "a loud shop window display" is consistently staged as a high-salience retail spectacle, with oversized SALE signage, bright lighting, and attention-grabbing displays, whereas "a freezing room" is rendered as an industrial cold-storage space, with frosted surfaces, metal shelving, and a walk-in freezer atmosphere. These scripts crowd out less stereotypical contexts and make a small set of "expected" situations feel like the default settings in which particular sensations should appear.

Layout Bias: Certain sensory adjectives are consistently linked to specific viewpoints, establishing a default "way of seeing". In Figure 6, "an echoing shop window display" typically adopts a frontal, highly symmetrical composition with strong one-point perspective and repeated frames that pull the eye into depth, while "a full body shot of silky wear" converges on a centred full-body fashion pose with similar camera distance and staging across samples. These regularities constrain the model's compositional range, presenting a single visual syntax as the natural representation of the sensation.

These cross-sensory defaults amount to a new mode of

Table 1: Brief introduction of Modality × Carrier.

Carriers	Gustatory	Tactile	Auditory	Olfactory
a [Adjective] 3D sphere	Sweet	Burning	High-pitched	Fragrant
a [Adjective] room	Sour	Freezing	Loud	Musty
a full body shot of [Adjective] wear	Bitter	Silky	Melodic	Pungent
a [Adjective] dish	Spicy	Prickly	Echoing	Fresh
a [Adjective] shop window display	Salty	Soft	Rhythmic	Synthetic

Figure 5: Example of Situational Bias.

Figure 6: Example of Layout Bias.

aesthetic knowledge production. As a result, T2I pipelines can scale culturally learned sensory associations into visual conventions that come to function as default representations.

Experimental Setup

Model Choices. We considered eight T2I models: SDXL (Podell et al., 2023), FLUX.1 (Greenberg, 2025), Nano Banana (Raisinghani, 2025), GPT-5 (OpenAI, 2025), Seedream 4.5 (Seedream et al., 2025), Qwen3-VL-235B-A22B (Wu et al., 2025), ERNIE 4.5 (Baidu-ERNIE-Team, 2025) and Kling 2.1. For the main evaluation, we focused on six models that are widely adopted and representative of current mainstream use. ERNIE 4.5 and Kling 2.1 were included in preliminary trials but were not retained for the full-scale analysis.

Modality × Carrier. Our experiments are organised around a controlled **modality** × **carrier** design. We focus on four non-visual sensory modalities, namely *Gustatory*, *Tactile*, *Auditory*, and *Olfactory*. For each, we selected five common adjectives that are unambiguously sensory, elicit cross-sensory imagery, and possess culturally stable meanings. We also prioritise words with relatively stable meanings across common cultural contexts. This selection is intentionally not exhaustive. For example, we include *soft* and *silky* as canonical positive tactile descriptors, but do not aim to cover all polarity pairs such as *hard* or *rough*. Our goal is to probe robust and widely shared sensory cues, rather than to build a complete lexicon. To examine how the same adjective is visualised across contexts, we pair each adjective with five everyday visual carriers that cover a broad range of common generation settings. These carriers are drawn from routine life domains, including food, clothing, housing, and retail. Table 1 gives a brief introduction of modality × carrier. Each prompt follows a fixed template, with only the adjective and carrier varied. The full design therefore consists of 4 modalities × 5 adjectives per modality × 5 carriers, resulting in 100 sensory–carrier prompt types. This structure supports two complementary comparisons: within-adjective analyses across carriers, and within-carrier analyses across adjectives.

Configurations. All experiments were conducted with standardised configurations to ensure consistency and reproducibility. For each sensory–carrier prompt type, we generated $N = 10$ images. This repeated sampling reduces sensitivity to stochastic variation in the generation process and allows us to estimate the stability of each model's cross-sensory tendencies beyond single-run idiosyncrasies. Overall, this procedure yields over 6,000 generated images across models.

Evaluation. We quantify cross-sensory bias as a lack of diversity within each sensory–carrier category. For each category, we measure how strongly a set of generated images converges on a recurring visual template. Using a fixed rubric, a subfactor is treated as locked if it remains effectively unchanged in more than 80% of images; otherwise, it is treated as varied. Each bias dimension is rated on a 0–10 scale, with higher values indicating stronger within-category uniformity. We implement this rubric in a VLM-based evaluation procedure (Gemini 3.0 Pro) that produces structured annotations and per-dimension bias scores. Each bias dimension is decomposed into multiple sub-factors; for instance, demographic bias considers repeated patterns in gender presentation, age group, and skin-tone cues when people are depicted. The overall cross-sensory bias score for a category is computed as the mean of

Table 2: Lower scores indicate more diverse images within each sensory–carrier category, whereas higher scores indicate a stronger tendency to repeat the same visual template and produce more stereotyped depictions.

Model	Overall		Color		Demographic		Entity		Situational		Layout	
	Score	σ	Score	σ	Score	σ	Score	σ	Score	σ	Score	σ
Nano Banana	7.68	1.39	6.82	1.18	**8.46**	**1.11**	7.77	1.45	8.19	1.51	7.68	1.44
Seedream 4.5	6.60	1.59	5.38	1.45	7.70	1.26	6.46	1.03	7.25	1.45	7.76	1.59
Qwen3-VL-235B	7.40	1.50	6.83	1.39	8.15	1.01	7.30	1.29	7.85	1.43	8.08	1.05
Flux.1-schnell	7.36	1.08	6.50	1.64	7.40	1.30	7.26	1.49	8.43	1.33	8.17	1.40
SDXL	7.02	1.46	6.47	1.32	7.40	1.10	6.80	1.61	8.26	1.65	7.55	1.53
ChatGPT-5	**8.38**	**0.67**	**7.69**	**1.39**	8.12	1.28	**8.00**	**1.12**	**9.26**	**0.66**	**8.73**	**1.05**

all applicable bias scores, excluding non-applicable cases.

Experimental Results

Main Results

In this section, we report the main quantitative results of our cross-sensory bias evaluation across six mainstream T2I models. Across models, Table 2 shows clear separation in overall cross-sensory bias. Seedream 4.5 achieves the lowest overall score (6.60), indicating the weakest bias on average, whereas ChatGPT-5 exhibits the highest overall score (8.38) together with the smallest overall variance ($\sigma = 0.67$). This suggests that its cross-sensory defaults are not only strong but also relatively stable across categories. In contrast, models with lower means but larger σ (e.g., Seedream 4.5, $\sigma = 1.59$) appear more heterogeneous, with some sensory–carrier categories remaining diverse while others show pronounced bias.

Breaking down by dimension, situational and layout biases are consistently high across models and often dominate the overall signal. Notably, ChatGPT-5 reaches 9.26 in Situational and 8.73 in Layout, indicating strong convergence on stereotyped scene scripts and camera grammars. By comparison, color bias is generally weaker (e.g., Seedream 4.5 at 5.38), suggesting that palette variation is easier to preserve than diversity in depicted entities, scenes, or composition. Entity bias also shows a clear spread, with Seedream 4.5 lowest (6.46) and ChatGPT-5 highest (8.00), implying that some models more readily reduce sensory adjectives to a small set of prototypical objects or character types. Demographic bias is substantial for several models, with Nano Banana peaking at 8.46, reflecting strong regularities in how sensory adjectives are embodied in human depictions; even the lowest-scoring models on this dimension remain relatively elevated (e.g., FLUX.1-schnell and SDXL at 7.40).

Figure 7 further decomposes the overall score by sensory modality and carrier. Averaged across models, olfactory prompts show the strongest overall bias, followed by tactile, whereas auditory is lowest. This pattern also holds within individual systems: Nano Banana and Qwen3-VL peak on olfaction (8.03 and 7.81), while ChatGPT-5 peaks on touch (8.62). Across carriers, Sphere and Shop tend to elicit stronger bias than Dish/Room/Wear. In particular, Sphere is the most

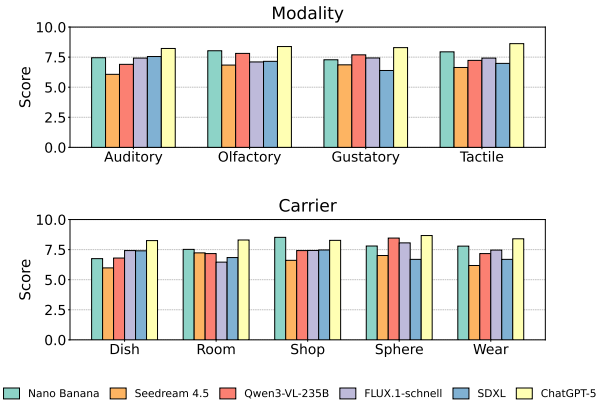

Figure 7: Mean overall cross-sensory scores for each model, broken down by sensory modality and by carrier category.

bias-prone carrier for several models (e.g., Qwen3-VL at 8.46; ChatGPT-5 at 8.67), consistent with the idea that abstract or display-like frames encourage reliance on learned visual defaults. By contrast, Dish is often among the lowest-scoring carriers for most models, suggesting comparatively more room for variation in food depiction, although ChatGPT-5 remains high even in this setting.

Human-LLM Alignment

We quantify model alignment by measuring the distance between model centroids (μ_{model}) and human-preference proxies (v_{human}). For each category, μ_{model} is the mean CLIP (Radford et al., 2021) embedding of the generated images. To establish a ground truth for human intuition, we conducted a behavioral study where participants selected the most representative images from a curated pool for each sensory prompt. These selections are then averaged in CLIP space to construct v_{human}, allowing us to define the alignment gap as:

$$D_{\text{gap}} = 1 - \cos\left(\mu_{\text{model}}, v_{\text{human}}\right) \quad (1)$$

where lower values indicate stronger alignment between the model's typical output and human selections.

Figure 8 illustrates these gaps in a shared 2D projection of the CLIP space, revealing two primary systematic regularities.

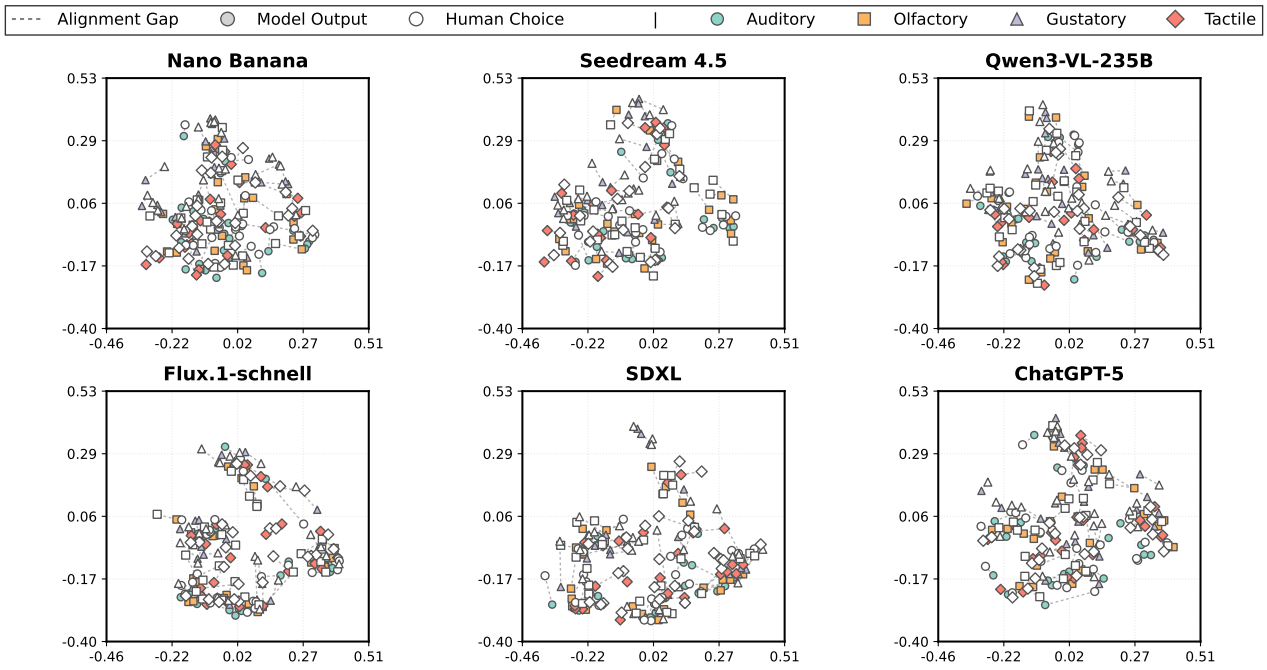

Figure 8: **Visualization of Cross-Sensory Semantic Alignment.** For each model (one panel), solid markers denote the model centroid (μ_{model}) and hollow markers denote the human-choice proxy (ν_{human}) for each sensory subclass; dashed connectors indicate the alignment gap between them. Marker shape and color encode the sensory modality.

First, many categories form local clusters in the embedding space, implying that models and humans share a coarse semantic organization of sensory concepts. Second, the lengths of the dashed connectors vary systematically across categories and models, reflecting that misalignment is not random noise but concentrates in particular modality–carrier combinations.

Quantitatively, models differ reliably in their alignment with human intent (see Table 3). ChatGPT-5 exhibits the smallest mean gap ($\bar{D} = 0.066$), suggesting its typical outputs consistently fall close to what human raters select as prompt-consistent. In contrast, Seedream 4.5 exhibits the weakest alignment overall ($\bar{D} = 0.115$). The modality breakdown reveals that Auditory subclasses generally show the largest gaps—peaking at $D_{\text{gap}} = 0.290$ for Nano Banana in the *dish* carrier, whereas Tactile subclasses show the smallest (e.g., $D_{\text{gap}} = 0.017$ for *soft wear* in Seedream 4.5). This disparity suggests that mapping sound-related adjectives into visual forms is more weakly constrained by shared human preferences than touch-related descriptors, which possess stronger physical and visual grounding.

The most striking finding is the paradoxical performance of ChatGPT-5. Although it achieves the highest alignment scores, it also exhibits the strongest cross-sensory bias. This suggests that RLHF prioritizes "safe" visual defaults that track entrenched social schemas. These defaults mirror average human expectations but are often rooted in cultural stereotypes. As a result, ChatGPT-5 functions less as a nuanced interpreter of sensory concepts and more as a mirror of central human preferences. By sacrificing long-tail diversity, the model effectively absorbs and stabilizes the biases and fixed associations already present in human society. This underscores the need to evaluate not only whether T2I systems agree with human preferences, but also what kinds of diversity and social neutrality are lost when they do.

Table 3: Quantitative Comparison of Semantic Alignment Gap (D_{gap}). All values are in percentage (%). Lower scores indicate better alignment with human intent.

Model	Mean Gap	Std.	Best Sense	Worst Sense
ChatGPT-5	**6.66**	4.29	Tactile (5.26)	Auditory (8.28)
Qwen3-VL-235B	7.80	4.76	Olfactory (6.14)	Gustatory (9.40)
Flux.1-schnell	8.13	3.55	Auditory (7.35)	Gustatory (9.28)
Nano Banana	8.65	4.64	Tactile (7.66)	Auditory (10.22)
SDXL	10.75	4.85	Tactile (9.63)	Auditory (11.22)
Seedream 4.5	11.52	5.00	Tactile (9.91)	Auditory (14.08)

Conclusion

In this study, we introduce a novel application of crossmodal correspondence theory to T2I models. Through this approach, we reveal the widespread perceptual templates of mainstream models across multiple dimensions, and quantify their systematic biases toward stereotypical sensory depictions. By proposing the modality × carrier design and integrating VLM-based evaluation, we provide effective tools for identifying AI biases. This work deepens our understanding of cross-sensory behaviour in T2I models and supports efforts to promote sensory diversity and fairness in generated content.

References

Afreen, J., Mohaghegh, M., & Doborjeh, M. (2025). Systematic literature review on bias mitigation in generative ai. *AI and Ethics*, 1–53.

Alper, M., & Averbuch-Elor, H. (2023). Kiki or bouba? sound symbolism in vision-and-language models. *Advances in Neural Information Processing Systems*, *36*, 78347–78359.

Baidu-ERNIE-Team. (2025). Ernie 4.5 technical report.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, *59*(1), 617–645.

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J. Y., & Caliskan, A. (2022). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*.

Cardello, A. V., Chheang, S. L., Hedderley, D. I., Guo, L. F., Hunter, D. C., & Jaeger, S. R. (2019). Toward a new scale to measure consumers'"need for uniqueness" in foods and beverages: The 31-item fbnfu scale. *Food Quality and Preference*, *72*, 159–171.

Chen, H., Xiang, Q., Hu, J., Ye, M., Yu, C., Cheng, H., & Zhang, L. (2025). Comprehensive exploration of diffusion models in image generation: A survey. *Artificial Intelligence Review*, *58*(4), 99.

Chen, H. (2024). A study of artificial intelligence's impact on aesthetic standards and its potential social dilemmas. *J Sociol Ethnol*, *6*(5), 35–42.

Di Stefano, N., & Spence, C. (2024). Perceptual similarity: Insights from crossmodal correspondences. *Review of Philosophy and Psychology*, *15*(3), 997–1026.

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of us public opinion polls from 1946 to 2018. *American psychologist*, *75*(3), 301.

Elsharif, W., Alzubaidi, M., & Agus, M. (2025). Cultural bias in text-to-image models: A systematic review of bias identification, evaluation and mitigation strategies. *IEEE Access*.

Fillmore, C. J., et al. (2006). Frame semantics. *Cognitive linguistics: Basic readings*, *34*, 373–400.

Girrbach, L., Alaniz, S., Smith, G., & Akata, Z. (2025). A large scale analysis of gender biases in text-to-image generative models. *arXiv preprint arXiv:2503.23398*.

Greenberg, O. (2025). Demystifying flux architecture. *arXiv preprint arXiv:2507.09595*.

Guedes, D., Garrido, M. V., Lamy, E., Cavalheiro, B. P., & Prada, M. (2023). Crossmodal interactions between audition and taste: A systematic review and narrative synthesis. *Food Quality and Preference*, *107*, 104856.

Knöferle, K., & Spence, C. (2012). Crossmodal correspondences between sounds and tastes. *Psychonomic bulletin & review*, 1–15.

Kouwenhoven, T., Shahrasbi, K., & Verhoef, T. (2025). Crossmodal associations in vision and language models: Revisiting the bouba-kiki effect. *arXiv preprint arXiv:2507.10013*.

Kress, G., & Van Leeuwen, T. (2020). *Reading images: The grammar of visual design*. Routledge.

Lee, B. P., & Spence, C. (2022). Crossmodal correspondences between basic tastes and visual design features: A narrative historical review. *i-Perception*, *13*(5), 20416695221127325.

Leng, J., Su, X., Liu, Z., Zhou, L., Chen, C., Guo, X., Wang, Y., Wang, R., Zhang, C., Liu, Q., et al. (2025). Diffusion model-driven smart design and manufacturing: Prospects and challenges. *Journal of manufacturing systems*, *82*, 561–577.

Naik, R., & Nushi, B. (2023). Social biases through the text-to-image generation lens. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 786–808.

Nayak, S., Bhatia, M., Zhang, X., Rieser, V., Hendricks, L. A., Van Steenkiste, S., Goyal, Y., Stańczak, K., & Agrawal, A. (2025). Culturalframes: Assessing cultural expectation alignment in text-to-image models and evaluation metrics. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 20918–20953.

Ohtake, Y., Tanaka, K., & Yamamoto, K. (2023). How many categories are there in crossmodal correspondences? a study based on exploratory factor analysis. *Plos one*, *18*(11), e0294141.

OpenAI. (2025, August). GPT-5 System Card [Accessed: 2025-12-12].

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., & Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, 8748–8763.

Raisinghani, N. (2025, November). Introducing nano banana pro [Google Blog (The Keyword). Accessed: 2025-12-12].

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382–439.

Schank, R. C., & Abelson, R. P. (2013). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology press.

Schneider, M., & Hagendorff, T. (2024). Investigating toxicity and bias in stable diffusion text-to-image models. *Scientific Reports*, *15*.

Sczesny, S., Nater, C., & Eagly, A. H. (2018). Agency and communion: Their implications for gender stereotypes and

gender identities. In *Agency and communion in social psychology* (pp. 103–116). Routledge.

Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al. (2025). Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*.

Shen, F., Liang, Z., Wang, C., & Wen, W. (2025). Cider: A causal cure for brand-obsessed text-to-image models. *arXiv preprint arXiv:2509.15803*.

Smith, N., & Southerton, C. (2025). Ai and aesthetic alienation: The image and creativity in contemporary culture. *Social Science Computer Review*, 08944393251361449.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*(4), 971–995.

Spence, C., & Levitan, C. A. (2021). Explaining crossmodal correspondences between colours and tastes. *i-Perception*, *12*(3), 20416695211018223.

Spence, C., Wan, X., Woods, A., Velasco, C., Deng, J., Youssef, J., & Deroy, O. (2015). On tasty colours and colourful tastes? assessing, explaining, and utilizing crossmodal correspondences between colours and basic tastes. *Flavour*, *4*(1), 23.

Varghese, J. K. (2025). Platform vernaculars: How ai image generators create new forms of visual bias. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *8*(3), 2936–2938.

Velasco, C., Woods, A. T., Deroy, O., & Spence, C. (2015). Hedonic mediation of the crossmodal correspondence between taste and shape. *Food Quality and Preference*, *41*, 151–158.

Velasco, C., Woods, A. T., Petit, O., Cheok, A. D., & Spence, C. (2016). Crossmodal correspondences between taste and shape, and their implications for product packaging: A review. *Food Quality and Preference*, *52*, 17–26.

Verhoef, T., Shahrasbi, K., & Kouwenhoven, T. (2024). What does kiki look like? cross-modal associations between speech sounds and visual shapes in vision-and-language models. *arXiv preprint arXiv:2407.17974*.

Wan, X., Woods, A. T., van den Bosch, J. J., McKenzie, K. J., Velasco, C., & Spence, C. (2014). Cross-cultural differences in crossmodal correspondences between basic tastes and visual features. *Frontiers in psychology*, *5*, 1365.

Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.-m., Bai, S., Xu, X., Chen, Y., et al. (2025). Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Zhang, L., Liao, X., Yang, Z., Gao, B., Wang, C., Yang, Q., & Li, D. (2024). Partiality and misconception: Investigating cultural representativeness in text-to-image models. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–25.