SENSE IT WITH YOUR EYES: SENSATION GENERATION AND UNDERSTANDING FOR ADVERTISEMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sensory advertising evokes human senses through visual cues, enabling audiences to mentally simulate experiences and increasing persuasive impact. Despite the recent increase in using AI in generating and understanding creative and persuasive content, how advertisements visually evoke sensations remains largely unexplored. In this work, we introduce the first study of evaluating and generating sensory ads. We introduce the Sensory Ad dataset, and propose SenseScore, an automated evaluation metric for sensation evocation achieving strong agreement with human judgments. Further, we introduce the Sensory Ad Generation (SenseGen) task and propose SAGA, a multi-agent framework that improves message–image alignment, sensory evocation, and persuasion. Our work establishes a foundation for sensory-aware visual persuasion.

1 INTRODUCTION

> "I have left behind illusion, I said to myself. Henceforth I live in a world of three dimensions–with the **aid of my five senses**. I have since learned that there is no such world, but then, as the car turned out of sight of the house, I thought it took no finding, but lay all about me at the end of the avenue." (Evelyn Waugh, "Brideshead Revisited")

The full spectrum of senses is important for humans to navigate and experience their environments. However, humans sometimes hallucinate sensations, with very real effects: people experiencing lexical-gustatory synesthesia experience taste triggered by words (Ward & Simner, 2003), visually impaired people can "see" with their tongue through electrical signals (Nau et al., 2015), phantom limb pain can be treated with augmented reality (Prahm et al., 2025).

In marketing, sensory advertising enhances persuasion and brand effectiveness (Krishna, 2012; Lindstrom, 2006; Krishna et al., 2016; Elder & Krishna, 2022) by allowing the audience to imagine the benefit of a product or the consequence of an action in a visceral way. Since stimulating the senses in the exact sense modality is infeasible, ads resort to visual content *associated* with the target sensation. For example in Fig. 1, on a hot summer day, image (b) is more likely to convince a thirsty audience to buy the drink by evoking the cooling and refreshing sensation (through the inclusion of the ice cubes), compared to (a).

In this work, we conduct the first investigation of *how ads evoke the senses through visual means*. We focus on three facets: recognizing, scoring (evaluating), and generating sensory evocative ads. First, we develop a taxonomy of senses at different levels of granularity in which the first layer corresponds to the five fundamental sensory modalities (information perceived through the eyes, ears, nose, skin, and tongue). These senses are then further refined into more specific subcategories (e.g., "temperature" is a type of "touch"). We construct a dataset (primarily intended for evaluation) by collecting annotations on 1100 images sourced from an existing dataset of advertisements (PittAds Hussain et al. (2017)): whether the image evokes a sensation and if so, the category of sensation, the visual elements evoking it, and score of *how well* the image evokes the sensation.

Second, we propose an evaluation method, **SenseScore**, that measures how effectively an image evokes a target sensation. SenseScore first utilizes an MLLM to generate the description of the image, then fine-tunes an LLM using CPO Xu et al. (2024) and hierarchy loss. Experimental results

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

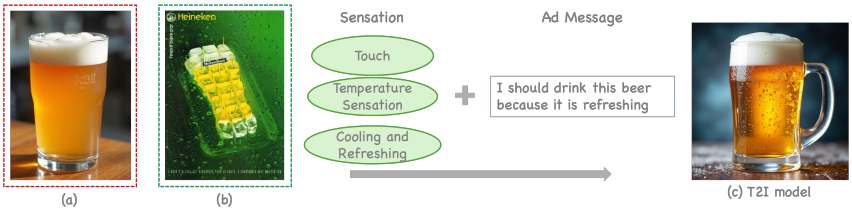


Figure 1: **Sensory Ad Generation:** The message is shown in a box ("I should... because..."). Image (a) is generated by a T2I model using the message as prompt. Image (b) is a real human-designed ad. Image (c) is an ad designed by a T2I model which also takes an intended sensation as input.

show that our evaluation metric achieves a Kappa (Cohen, 1960) agreement score of 0.69 with human annotators, representing an improvement of 30% over existing baseline metrics.

Third, we introduce the **Sensory Ad Generation (SenseGen)** task, where the goal is to generate advertisement images that both convey a given message and evoke a specified sensation. The messages are in the form *"I should* {*action*} *because* {*reason*}*"* (Hussain et al., 2017). Our results show that existing T2I models fail in generating advertisement images that evoke specific sensations while accurately conveying the intended message.

Finally, we propose a multi-agent image generation framework, **Sensory Ad Generation with Agent (SAGA)**, to generate advertisement images that both convey the message and evoke the sensation. Our framework improves sensation evocation, as well as the effectiveness of the generated ads, measured via alignment of image and message, and persuasion scores from prior work.

We observe that sensory evocation and messaging/persuasion are not necessarily correlated: sensory evocation can boost alignment with the intended message and persuasiveness, but only through careful design. The complication is that sensory evocation can be exaggerated, to the point of distracting from the image's intended meaning and message. Our SAGA system cleverly avoids exaggeration through self-refinement. It optimizes the potential of sensory evocation, by effectively utilizing this evocation to message alignment and impact on the audience.

To summarize, we introduce: (1) the **Sensory Ad dataset**, (2) the novel task of Sensory Ad Generation. We propose (3) **SenseScore**, an automated evaluation method for sensation evocation, and (4) **SAGA**, a multi-agent framework for generating effective sensory ads.

2 Related Works

Text-to-Image Generation. Text-to-Image (T2I) models such as Flux Black Forest Labs (2024), Stable Diffusion Esser et al. (2024), Qwen-Image Wu et al. (2025), PixArt Chen et al. (2024a), DALLE3 Betker et al. (2023), etc. have advanced in generating high quality and realistic images given the explicit description of the prompt. Some existing works (Aghazadeh & Kovashka, 2025; Liao et al., 2024; Menon et al., 2024) assess the capability of models in generating images from abstract concepts and messages (like advertisement design tasks). Others (Yang et al., 2024b; Dang et al., 2025; Park & Lee, 2020) tackle emotion transfer through images. Sensation and emotion differ; the former is more visceral/physical while the latter is more interpretative/psychological.

Text to Image Evaluation. Existing metrics (Lin et al., 2024; Xu et al., 2023) assess how well an image corresponds to an explicit prompt specifying concrete objects, attributes, or visual relations. Evaluating **sensation evocation** poses a unique challenge: the sensation is not only an implicit concept but the same sensation can be represented through entirely different visual designs.

Understanding Modalities beyond Sight. Our work is part of a bigger trend, including examples such as understanding audio and touch data (Ghosh et al., 2024; Yang et al., 2024a) or semantic-taste and visual-smell mappings (Bender et al., 2023; Ozguroglu et al., 2025). Other work predicts physical properties such as density and hardness from images and descriptions (Zhai et al., 2024). However, no prior work studies how images are created to evoke specific sensations, nor predicts computationally the impact of sensations on an audience.

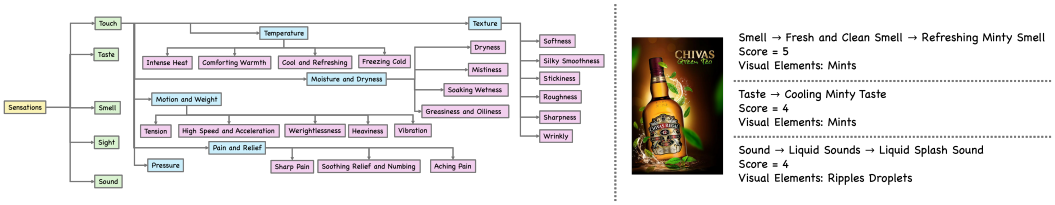


Figure 2: Left: **Sensation hierarchy** (only categorizing the Touch sensation): Green box represents first level sensation, blue represents second, and purple represents third. Right: **Annotation example** from our Sensory Ad dataset.

Understanding and Generating Advertisements. Hussain et al. (2017) pioneer computational visual ad understanding, but do not capture sensory information. Kumar et al. (2023); Singh et al. (2025); Qiu et al. (2025) study persuasion strategies. Prior work has investigated T2I models for generating advertisements, focusing on criteria such as creativity and persuasion (Aghazadeh & Kovashka, 2025), depicting specific metaphorical relationships (Akula et al., 2023), or personalized effects (Kim et al., 2025). However, these generation studies do not examine the models' ability to implement specific persuasion strategies, such as the evocation of specific sensations, which play a crucial role in making ads influential and memorable.

Sensory Advertising is studied in Krishna & Schwarz (2014); Petit et al. (2019); Hultén (2015); Krishna et al. (2016). Krishna (2012) define sensory marketing as "marketing that engages the consumers' senses and affects their perception, judgment and behavior." Subconscious sensory triggers may make the ad's message more compelling than explicit messaging, causing viewers to perceive specific properties of the product. Examples include product packaging (e.g., Hershey's chocolate kisses creating the sensation of a drop melting), sound symbolism (e.g., the word "frosh" evoking the sensation of creaminess more than "frish"), and the memories scents create and evoke. Cian et al. (2014) describe the dynamics encoded in similar but slightly varied imagery (e.g., a horizontal vs tilted seesaw). None of this work computationally models visually evoked sensations. Singh et al. (2024) focus on observable behavioral reactions such as likes, upvotes, or memorability, which are external message recipient (viewer) outcomes and do not indicate what sensory experiences an image evokes. We instead focus on the sender of the signals (the images that evoke the sensations).

Agentic Image Generation. A growing line of work casts (multimodal) LLMs as *agents* that plan, act with tools, and refine their outputs through self-generated feedback (Yao et al., 2023; Madaan et al., 2023; Shinn et al., 2023; Wu et al., 2024a; Hong et al., 2024). Related frameworks were proposed for text-to-image (T2I) generation via iterative self-refinement (Yang et al., 2024d), MLLM planning and recaptioning (Yang et al., 2024c; Feng et al., 2023), closed-loop self-correction (Wu et al., 2024b; Wang et al., 2024; Li et al., 2025; Wan et al., 2025), and backbone selection (Qin et al., 2024). However, prior agentic T2I systems optimize *literal* objectives (prompt fidelity, object counts, spatial relations), whereas SAGA targets the *implicit, rhetorical* goals of advertising: jointly optimizing message alignment, persuasion, and evocation of a target sensation. Wang et al. (2025); Chen et al. (2025) automate poster *design*, optimizing objectives like layout, typography, and branding, whereas SAGA pursues evoking a target sensation while conveying a persuasive message. To our knowledge, SAGA is the first agentic framework for sensory-aware, persuasive ad generation.

3 Dataset, Tasks and Methods

3.1 Sensory Ad Dataset

Sensation Hierarchy (Taxonomy). Some advertisements are designed to evoke sensations that help audiences imagine a specific situation and the need for a product more vividly, an important factor in ad effectiveness (Krishna et al., 2016). In this work, we formalize the notion of sensation using a hierarchical taxonomy supported by neuroscience research (Ackerman et al., 2010; Gravina et al., 2013) (partly shown in Fig. 2; also see appendix). At the top level (left in the figure) are the five primary senses. Each is further subdivided into fine-grained categories, e.g., "Touch" is refined into "Texture", "Temperature," "Moisture and Dryness," "Pain and Relief," and "Pressure", resulting in 96 total sensation labels. By definition, if an image evokes a child sensation (e.g., "Temperature"),

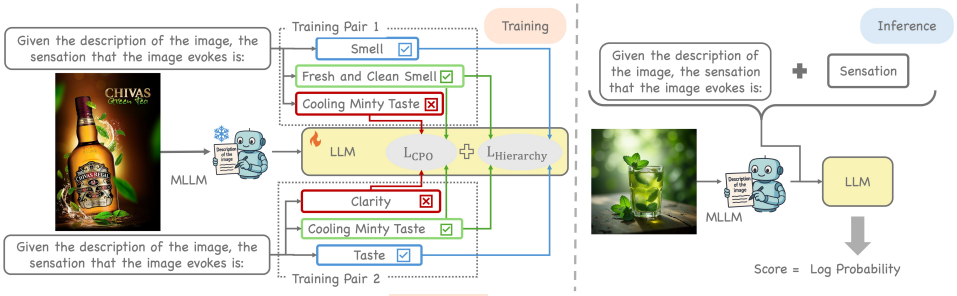


Figure 3: **SenseScore evaluation.** Left: training of LLM with two different sets of sensations for one image. Green border shows the winner sensation, blue border represents the winner parent (used in hierarchy loss), and red border denotes the loser. Each pair is derived from a triplet of annotations, A preferred over B, and B over C. Right: score computation in inference with the fine-tuned LLM.

it also evokes its parent (e.g., "Touch"). Organizing sensations in this hierarchical manner allows us to capture these dependencies and provides a structured representation for modeling and evaluating sensory understanding; we show in D.1 - Table 4 the impact of the hierarchy quantitatively, using an auxiliary classification task. We introduce a dataset of both real and generated ads annotated with (i) up to three groups (leaves and ancestors) of sensations evoked by each image, (ii) a score reflecting the strength of evocation, and (iii) visual elements that contribute to the sensation.

Data Collection. We first selected 1100 images from the PittAd dataset Hussain et al. (2017). The total number of sensations (at any hierarchy level) is 96, so we obtain 96 labels per image (most of which equal to 0 as the sensation is not chosen). We have included the data statistics including the topics diversity, sensations diversity, and human-human agreement in appendix.

Annotation Process and Quality Control. Annotation was carried out by trained crowd workers on Prolific and using forms created on Qualtrics. Before contributing, each annotator was approved/filtered based on completing a practice form after reading detailed instructions, definitions of sensations, and illustrative examples. The annotation followed a structured protocol: annotators first chose the most prominent sensation among the five top-level categories (with the option of selecting "None" if no sensation was evoked). Based on their choice, they were presented with progressively narrower subcategories until reaching a leaf-level sensation. For each selected sensation, annotators provided a strength score and listed the visual elements (e.g., colors, objects, textures) that contributed to it, using free-form text (which can be used in future work). The process was repeated up to three times per image until the annotator chose "None".

Agreement. For about 10% of samples, we collected annotations from multiple workers, to compute inter-annotator (human-human) agreement. Kappa (Cohen, 1960) was high at 0.8. The full annotation and testing forms are in the supplement, and the dataset will be released upon acceptance.

Annotations on Generated Ads. We primarily evaluate the quality of our SenseScore metric on real ad images. To test its generalizability to AI-generated advertisements, we also annotated a subset of generated ads. We used the action-reason statements (from Hussain et al. (2017)) and a sensation as inputs to five T2I models: Flux Black Forest Labs (2024), AuraFlow Fal (2024), PixArt Chen et al. (2024a), Stable Diffusion 3 Esser et al. (2024), and Qwen-Image Wu et al. (2025). From 750 images generated by each model, we randomly selected 15 and annotated them using the same procedure as for real ads. Table 1 shows consistent trends, confirming the broad effectiveness of SenseScore.

3.2 Sensation Evocation Scoring (SenseScore)

Sensory evocation can make advertisements more persuasive by helping viewers vividly imagine the intended context (Elder & Krishna, 2022). However, an ad may evoke multiple sensations to different degrees, requiring evaluation of both the sensations evoked and their relative strengths. We introduce **SenseScore**, a hierarchy-aware evaluation method that measures how strongly an image evokes a given sensation. As illustrated in Fig. 3, SenseScore consists of two stages: (1) **Description Generation**, where an MLLM (e.g, InternVL) generates a textual description of the image, and (2) **Sensation Intensity Scoring**, where an LLM uses this description to score a target sensation.

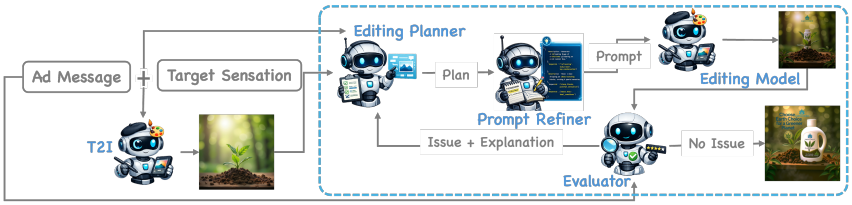


Figure 4: **SAGA framework.** Generation and editing process in multi-agent framework. Blue box represents the conversation loop among the agents, which repeats until reaching a maximum number of messages or 'No Issue' from the Evaluator Agent.

Hierarchy-Aware Preference Learning. Initial experiments show that zero-shot LLMs and MLLMs perform poorly in both retrieving correct sensations (Table 4) and estimating their intensity Table 1. We therefore fine-tune the scoring model using a subset of our annotated dataset. SenseScore represents sensory evocation through the likelihood of a sensation label conditioned on an image description, rather than by generating a numerical intensity rating. Our SFT baseline maximizes the likelihood of human-selected sensation labels, but does not explicitly constrain their relative likelihoods according to perceived intensity. For example, in Fig. 2, the image evokes both *Smell* and *Taste*, but *Smell* is evoked more strongly. To capture such relative preferences, we construct preference pairs from sensations with different human-annotated intensity scores for the same image. Each pair contains a preferred sensation y^+ with higher intensity and a rejected sensation y^- with lower intensity. Following Contrastive Preference Optimization (CPO) (Xu et al., 2024), we train the LLM to assign a higher probability to the preferred sensation. Preference learning alone does not enforce the hierarchical structure of sensations, where a fine-grained sensation implies its ancestors. We therefore introduce a *hierarchy loss* that penalizes assigning a preferred child sensation a higher probability than its parent:

$$\begin{aligned} L_{\text{CPO+Hierarchy}} = &-\log \sigma\big(\beta\big[\log \pi_\theta(y^+ \mid x) - \log \pi_\theta(y^- \mid x)\big]\big) \\ &+ \text{ReLU}\big(\log \pi_\theta(y^+ \mid x) - \log \pi_\theta(y^{\text{parent}} \mid x)\big)\,. \end{aligned} \tag{1}$$

where x is input prompt, and y^{parent} is parent of y^+, $\pi_\theta(y \mid x)$ is the model's conditional probability of y given x, β is temperature scaling, and $\sigma(\cdot)$ is the sigmoid function. The first term learns relative sensation preferences; the second enforces hierarchical consistency of parent and child sensations.

Inference. Given an image and a target sensation, we prompt the fine-tuned LLM with the MLLM-generated description followed by "Given the description of the image, the sensation that the image evokes is:" SenseScore is the average token log-probability of the target sensation, with higher scores indicating stronger predicted evocation.

3.3 Sensory Ad Generation (SenseGen)

We introduce the **Sensory Ad Generation**, a task where the input consists of an advertisement message (action-reason statement Hussain et al. (2017)) and a target sensation. The objective is to generate an image that effectively conveys the message while also evoking the specified sensation. Examples outputs from prior models are shown in Fig. 1.

Sensory Ad Generation with Agent (SAGA). Generating sensory advertisements requires balancing ad's message communication with the evocation of a target sensation through appropriate visual cues. Achieving this balance is challenging, as emphasizing sensory elements may produce exaggerated imagery that obscures the intended message. To address this challenge, we develop SAGA, an iterative multi-agent framework that refines generated advertisements via targeted feedback. Rather than relying on a single prompt-to-image step, SAGA automates the iterative refinement typically performed by human designers. As shown in Fig. 4, SAGA consists of three agents: an Editing Planner, a Prompt Refiner, and an Evaluator, together with a text-to-image generator and an image-editing model. The process begins with an initial image generated from the advertisement message and target sensation. The three agents then iteratively refine the image as follows. (Prompts and loop examples are included in Sec. F and Fig. 12).

Editing Planner Agent. Given the current image, target sensation, and ads message, the Planner proposes a structured list of editing actions (e.g., adding/removing visual elements, modifying at-

Table 1: Kappa agreement between human annotators and evaluation metrics real images and generated ads. Best results are shown in bold.

Category	Metric	Real Ads						Gen. Ads
		Touch	Smell	Sound	Taste	Sight	All	
Baselines	VQA-score	0.57	0.55	0.41	**0.70**	0.61	0.53	0.52
	Image-Reward	0.41	0.39	0.51	0.37	0.35	0.40	0.40
	CLIP-score	0.35	0.33	0.43	0.39	0.28	0.31	0.45
	Pick-score	0.28	0.48	0.15	0.35	0.31	0.33	0.41
LLM/MLLM-as-a-Judge	InternVL	0.51	0.34	0.41	0.47	0.51	0.46	0.48
	QwenVL	0.44	0.28	0.30	0.48	0.46	0.41	0.43
	LLAMA3 + D_{InternVL}	0.42	0.34	0.41	0.51	0.49	0.43	0.47
	QwenLM + D_{InternVL}	0.40	0.24	0.25	0.45	0.48	0.36	0.48
SenseScore	LLAMA3 + D_{InternVL}	**0.69**	0.70	**0.57**	0.67	**0.74**	**0.69**	**0.65**
	LLAMA3 + D_{QwenVL}	**0.69**	**0.76**	0.53	0.63	0.72	**0.69**	0.64
	QwenLM + D_{InternVL}	0.55	0.57	0.41	0.51	0.66	0.55	0.53
	QwenLM + D_{QwenVL}	0.53	0.61	0.40	0.45	0.68	0.55	0.53

tributes like color, texture, or style) aimed at improving both conveying message and evoking the target sensation. In subsequent iterations, the Planner also incorporates feedback from the Evaluator to address the identified shortcomings.

Prompt Refiner Agent. The Prompt Refiner converts the Planner's structured action (JSON format) into a coherent prompt for the image editing model. It merges redundant actions and resolves inconsistencies to ensure that the proposed modifications are clearly communicated. The image-editing model then generates the revised image accordingly.

Evaluator Agent. The Evaluator assesses each image according to 3 criteria in priority order. First, it checks visual element consistency by identifying artifacts, glitches, or contradictory elements. If any are detected, it flags visual consistency as the primary issue and explains the inconsistency. Second, if the image is visually consistent, it assesses image–message alignment, prioritizing the intended message to avoid exaggerated sensory cues that obscure its meaning. If the product is not prominent or the visual elements fail to reinforce the message, it identifies what is missing or unclear. Third, if the previous criteria are satisfied, it evaluates sensation evocation through visual cues such as objects, colors, lighting, and atmosphere. If the target sensation is deemed weak or not effectively evoked, it explains the deficiency. If all three criteria are satisfied, the Evaluator returns "No Issue", terminating the refinement process.

4 Results

4.1 Sensation Evocation Scoring

We train SenseScore on 700 human-annotated real ad images. For each image, we pair every two of the 96 sensations that received different human scores, treating the higher-scoring sensation as preferred. Summed over more than 700 images, this gives 297,708 data points. To measure SenseScore's effectiveness, we measure its agreement with human judgments on more than 300 held-out human-annotated real images. We further evaluate SenseScore generalization to synthetic data by evaluating on 50 generated ads. Specifically, for each image, annotators select up to nine evoked sensations and assign each a score from 1 to 5. Unselected sensations receive a score of 0. We compare all pairs of sensations with distinct human scores, treating the higher-scoring sensation as the "winner". We then compare these preferences with those predicted by each metric (and humans) and compute Cohen's κ (Cohen, 1960) across all pairs, yielding approximately 30,000 and 5,000 pairs from real and generated ads, respectively.

SenseScore compared to baselines. We compare SenseScore against baseline metrics, including VQA-score (Lin et al., 2024), ImageReward (Xu et al., 2023), CLIP-score (Hessel et al., 2021), and Pick-score (Kirstain et al., 2023). We further compare SenseScore with LLM/MLLM-as-a-judge metrics. MLLM judges (InternVL (Chen et al., 2024b)/QwenVL (Bai et al., 2025)) score how strongly an image evokes a given sensation on a scale of 0-5, while LLM judges (LLaMA3 (AI@Meta, 2024)/QwenLM (Hui et al., 2024)) perform the same task using an InternVL-generated image description ($D_{InternVL}$) rather than direct visual input (see the prompts in sec. F).



Figure 5: **Examples on human agreement with SenseScore and VQA-score** on intensity of sensations. The Human row shows the chosen (✓) image(s) (including ties) and rejected (×) image. Red background indicates the model-chosen (higher-scoring) option is misaligned with human choice, and green background shows it is aligned.

Table 2: **SenseScore Ablation**. Kappa agreement with human annotators for LLAMA3 + D_{InternVL}. Best results are bolded.

Method	Real Ads					
	Touch	Smell	Sound	Taste	Sight	All
0-shot	-0.20	0.13	-0.34	-0.09	0.00	-0.14
SFT	0.36	0.52	0.60	0.38	0.51	0.33
CPO	0.58	0.65	0.43	**0.69**	0.70	0.57
Hierarchy + CPO	**0.69**	**0.70**	**0.57**	0.67	**0.74**	**0.69**

Table 3: **Sensory Ad Evaluation**. Gen. ads using AR + Sensation as input. AIM and P_{comp} measure alignment and persuasion.

T2I Model	SenseScore	AIM	P_{comp}
Flux	0.68	0.60	0.56
SD3	0.65	0.62	0.59
AuraFlow	0.68	0.60	0.57
PixArt	0.59	0.60	0.61
Qwen-Image	0.74	0.61	0.57
DALLE-3	0.79	0.63	0.61
SAGA (FLUX)	**0.83**	**0.66**	**0.62**

As shown in Tab. 1, among baselines, VQA-score achieves the highest overall agreement with moderate performance (κ = 0.53/0.52 on real/gen ads). SenseScore achieves κ = 0.69/κ = 0.65 on real/gen. ads, representing relative improvements of 30% and 25% (vs. strongest baseline) overall and substantially outperforms VQA-score on 4 (out of 5) sensations. This confirms **effectiveness of SenseScore on real images as well as its generalization on generated ads**. Additionally, SenseScore outperforms LLM/MLLM judges similar backbones significantly, indicating the effectiveness of our method. For a qualitative result, see the examples in Fig. 5.

Effectiveness of Preference Learning and Hierarchical Supervision. As shown in Tab. 2, zero-shot inference using the same scoring procedure as SenseScore (Inference in Sec. 3.2), but without fine-tuning, exhibits poor agreement with human judgments, **confirming that zero-shot is ineffective to capture sensation and its intensity**. While SFT increases agreement to $\kappa = 0.33$, CPO exhibits substantially higher agreement ($\kappa = 0.57$) by explicitly learning the relative strengths of evoked sensations rather than treating selected sensations as equally correct. Adding the hierarchy loss further boosts agreement to $\kappa = 0.69$, a relative increase of 21% over CPO alone. These results show the **complementary** contributions of preference learning in capturing differences in sensation intensity, and hierarchical supervision in encouraging consistency of parent/child sensations.

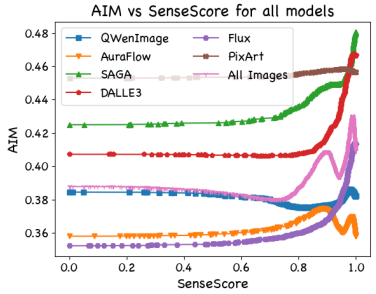
4.2 Sensory Ad Generation

We benchmark existing T2I models: Flux (Black Forest Labs, 2024), Stable Diffusion 3(SD3) (Esser et al., 2024), AuraFlow Fal (2024), PixArt (Chen et al., 2024a), and Qwen-Image (Wu et al., 2025) on SenseGen, which requires generating advertisements that convey an intended message while evoking a target sensation to make the images more persuasive. We further compare T2I models against SAGA on SenseGen (all receive an action-reason advertisement message and a target sensation). Appendix includes results on message-only (AR), message and sensation (AR+Sensation), and results on sensory images beyond ads.

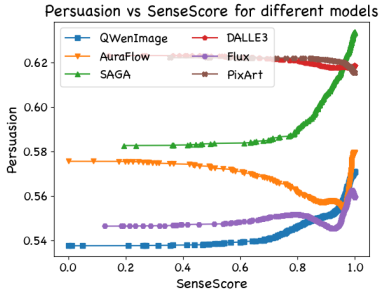
SAGA Improves Sensory Evocation and Message Alignment. We evaluate generated advertisements using SenseScore for target sensation evocation and two metrics from Aghazadeh & Kovashka (2025): AIM for image–message alignment and P_{comp} for estimated persuasiveness. Table 3 shows that SAGA achives the highest score across all metrics. Specifically, SAGA achieves the highest



Figure 6: **Sensory Ad examples:** real ads, and ads generated by Flux, SD3, AuraFlow, PixArt, DALLE3 and SAGA (ours) given the action-reason message and sensation annotation for the real advertisement. Green border represents the sensation used in the prompt of T2I models.



(a) Relation between SenseScore and AIM.



(b) Relation between SenseScore and P_{comp}.

Figure 7: Analysis of the relation between sensation intensity (SenseScore), text-image alignment (AIM), and persuasion (P_{comp}) in generated and real advertisements.

SenseScore (0.83) and message-alignment score (AIM = 0.66), compared with SenseScore = 0.79 and AIM = 0.63 for DALLE 3. It also obtains the highest P_{comp}. These results show that SAGA improves target sensation evocation while also achieving stronger alignment with the intended advertisement message.

Fig. 6 shows qualitative comparisons between SAGA and baselines. **SAGA effectively conveys both the implicit ad message and the intended sensory experience by integrating appropriate visual cues**. In contrast, baselines often produce overly literal interpretations of sensations (e.g., DALLE3 depicts literal "touch" in 1st row while our method connects touch to freshness), over-rely on text rather than rhetoric visual cues (e.g., DALLE3 in 2nd row), or miss the core ad message entirely (e.g., AuraFlow generating a generic splash scene in 3rd row).

We note that while the goal is to evoke specific sensations, sometimes models exaggerate in evoking the sensation, overlook the advertisement message, and only show sensation-associated objects; we discuss this shortly. This explains why SenseScore and AIM/P_{comp} are not always correlated.

Human Evaluation of SAGA. We conduct human evaluation using 100 images from SAGA and DALLE3 and four annotators (non-authors). For conveying the message, on 76% of sets annotators preferred SAGA, on 10% preferred DALLE3, and on the rest chose equal. For evoking the sensation, on 48% of sets they preferred SAGA, on 24% DALLE3, and on 28% chose equal.

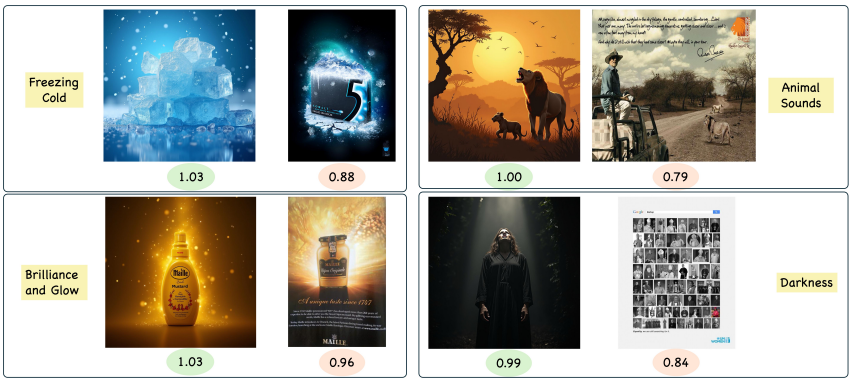


Figure 8: **Examples of exaggeration in sensation evocation.** Per group, left image is generated, right is real. Green SenseScores are higher but not necessarily better.

4.3 Connection of alignment, persuasion and sensation evocation

We investigate the relationship between sensation intensity (SenseScore), image–message alignment (AIM), and persuasion (P_{comp}), from (Aghazadeh & Kovashka, 2025) in generated advertisements. We plot the alignment over sensation intensity in Fig. 7a (Gaussian filter is applied on AIM to smooth the visualization), and persuasion score over intensity in Fig. 7b. Fig. 7a reveals that for several methods, alignment initially increases with sensation evocation, reaches a peak, and subsequently decreases as sensation evocation continues to increase. This indicates that sensation evocation does not always improve message alignment (i.e., AIM). Additionally, there is an optimal level of sensation evocation, which our method successfully finds (as shown by the superior performance on all three metrics in Tab. 3). In Fig. 7b, we observe that the persuasion score for the images generally increases with the increase in the sensation intensity. **This suggests that sensory evocation generally boosts persuasiveness.**

In Fig. 8, generated images evoke the input sensation with higher intensity than real ads; however, this exaggeration results in overlooking the advertisement message and failing in conveying it. For example, in Fig. 8, the top left image is supposed to convince the audience to buy the Five gum by showing the cooling and refreshing feature of the product; however, while the image intensely evokes the sensation it fails to convey the message. Our proposed SAGA copes with over-exaggeration by receiving specific feedback and correcting its outputs. SAGA addresses this failure mode: its Evaluator checks message alignment before sensation, so an image that evokes strongly but loses the message is flagged and revised.

5 Conclusion

We addressed the challenging, previously unexplored task of generating and understanding visual content that evokes specific human sensations, a crucial element of persuasive advertising. To facilitate research in this area, we introduced the *Sensory Ad benchmark* including the Sensory Ad dataset with a detailed hierarchical taxonomy for sensations, and the new Sensory Ad Generation task. We propose *SenseScore*, an evaluation metric that accurately measures the intensity of evoked sensations. We also propose SAGA, an agentic framework that proposes and evaluates edits to an image to improve sensory evocation, message alignment, and persuasiveness. This work lays the foundation for developing a new generation of sensation-aware models and expanding the scope of understanding sensory content beyond advertising.

AI use statement

We used AI-based tools to polish writing, for partial literature review (verified by the authors), and code debugging.

REFERENCES

Joshua M Ackerman, Christopher C Nocera, and John A Bargh. Incidental haptic sensations influence social judgments and decisions. *Science*, 328(5986):1712–1715, 2010.

Aysan Aghazadeh and Adriana Kovashka. Cap: Evaluation of persuasive and creative image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16970–16980, 2025.

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Arjun R Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T Freeman, et al. Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23201–23211, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Thoranna Bender, Simon Sørensen, Alireza Kashani, Kristjan Eldjarn Hjorleifsson, Grethe Hyldig, Søren Hauberg, Serge Belongie, and Frederik Warburg. Learning to taste: A multimodal wine dataset. *Advances in Neural Information Processing Systems*, 36:7351–7360, 2023.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Haoyu Chen, Xiaojie Xu, Wenbo Li, Jingjing Ren, Tian Ye, Songhua Liu, Ying-Cong Chen, Lei Zhu, and Xinchao Wang. POSTA: A go-to framework for customized artistic poster generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28694–28704, 2025. arXiv:2503.14908.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Luca Cian, Aradhna Krishna, and Ryan S Elder. This logo moves me: Dynamic imagery from static images. *Journal of marketing research*, 51(2), 2014.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Shengqi Dang, Yi He, Long Ling, Ziqing Qian, Nanxuan Zhao, and Nan Cao. Emoticrafter: Text-to-emotional-image generation based on valence-arousal model. *arXiv preprint arXiv:2501.05710*, 2025.

Ryan S Elder and Aradhna Krishna. A review of sensory imagery for consumer psychology. *Journal of Consumer Psychology*, 32(2):293–315, 2022.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Fal. Introducing auraflow v0.1, an open exploration of large rectified flow models, 2024. Available at: https://blog.fal.ai/auraflow/ [Accessed: 2024-11-07].

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in neural information processing systems*, 36: 18225–18250, 2023.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6288–6313, 2024.

Stephen A Gravina, Gregory L Yep, and Mehmood Khan. Human biology of taste. *Annals of Saudi medicine*, 33(3):217–222, 2013.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2308.00352.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Bertil Hultén. *Sensory marketing: Theoretical and empirical grounds*. Routledge, 2015.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1705–1715, 2017.

Junseo Kim, Jongwook Han, Dongmin Choi, Jongwook Yoon, Eun-Ju Lee, and Yohan Jo. Pvp: An image dataset for personalized visual persuasion with persuasion strategies, viewer characteristics, and persuasiveness ratings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19209–19237, 2025.

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

Aradhna Krishna. An integrative review of sensory marketing: Engaging the senses to affect perception, judgment and behavior. *Journal of consumer psychology*, 22(3):332–351, 2012.

Aradhna Krishna and Norbert Schwarz. Sensory marketing, embodiment, and grounded cognition: A review and introduction. *Journal of consumer psychology*, 24(2):159–168, 2014.

Aradhna Krishna, Luca Cian, and Tatiana Sokolova. The power of sensory marketing in advertising. *Current Opinion in Psychology*, 10:142–147, 2016.

Yaman Kumar, Rajat Jha, Arunim Gupta, Milan Aggarwal, Aditya Garg, Tushar Malyan, Ayush Bhardwaj, Rajiv Ratn Shah, Balaji Krishnamurthy, and Changyou Chen. Persuasion strategies in advertisements. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 57–66, 2023.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Arsh Koneru, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Reflect-DiT: Inference-time scaling for text-to-image diffusion transformers via in-context reflection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15657–15668, 2025. arXiv:2503.12271.

Jiayi Liao, Xu Chen, Qiang Fu, Lun Du, Xiangnan He, Xiang Wang, Shi Han, and Dongmei Zhang. Text-to-image generation for abstract concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3360–3368, 2024.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *ECCV*, 2024.

Martin Lindstrom. Brand sense: How to build powerful brands through touch, taste, smell, sight and sound. *Strategic Direction*, 22(2), 2006.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2303.17651.

Samia Menon, Sitong Wang, and Lydia Chilton. Moodsmith: Enabling mood-consistent multimedia for ai-generated advocacy campaigns. *arXiv preprint arXiv:2403.12356*, 2024.

Amy C Nau, Christine Pintar, Aimee Arnoldussen, and Christopher Fisher. Acquisition of visual perception in blind adults using the brainport artificial vision device. *The American Journal of Occupational Therapy*, 69(1):6901290010p1–6901290010p8, 2015.

Ege Ozguroglu, Junbang Liang, Ruoshi Liu, Mia Chiquier, Michael DeTienne, Wesley Wei Qian, Alexandra Horowitz, Andrew Owens, and Carl Vondrick. New york smells: A large multimodal dataset for olfaction. *arXiv preprint arXiv:2511.20544*, 2025.

Chanjong Park and In-Kwon Lee. Emotional landscape image generation using generative adversarial networks. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Olivia Petit, Carlos Velasco, and Charles Spence. Digital sensory marketing: Integrating new technologies into multisensory online experience. *Journal of Interactive Marketing*, 45(1):42–61, 2019.

Cosima Prahm, Korbinian Eckstein, Michael Bressler, Zhixing Wang, Xiaotong Li, Takashige Suzuki, Adrien Daigeler, Jonas Kolbenschlag, and Hideaki Kuzuoka. Phantomar: gamified mixed reality system for alleviating phantom limb pain in upper limb amputees—design, implementation, and clinical usability evaluation. *Journal of NeuroEngineering and Rehabilitation*, 22(1):21, 2025.

Jie Qin, Jie Wu, Weifeng Chen, Yuxi Ren, Huixia Li, Hefeng Wu, Xuefeng Xiao, Rui Wang, and Shilei Wen. DiffusionGPT: LLM-driven text-to-image generation system. *arXiv preprint arXiv:2401.10061*, 2024.

Haoyi Qiu, Yilun Zhou, Pranav Narayanan Venkit, Kung-Hsiang Huang, Jiaxin Zhang, Nanyun Peng, and Chien-Sheng Wu. Mmpersuade: A dataset and evaluation framework for multimodal persuasion. *arXiv preprint arXiv:2510.22768*, 2025.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2303.11366.

Somesh Singh, Harini SI, Yaman K Singla, Veeky Baths, Rajiv Ratn Shah, Changyou Chen, and Balaji Krishnamurthy. Teaching human behavior improves content understanding abilities of llms. *arXiv preprint arXiv:2405.00942*, 2024.

Somesh Kumar Singh, Yaman Kumar Singla, Balaji Krishnamurthy, et al. Measuring and improving persuasiveness of large language models. In *ICLR*, 2025.

Xingchen Wan, Han Zhou, Ruoxi Sun, Hootan Nakhost, Ke Jiang, Rajarishi Sinha, and Sercan Ö. Arık. Maestro: Self-improving text-to-image generation via agent orchestration. *arXiv preprint arXiv:2509.10704*, 2025.

Heng Wang, Yotaro Shimose, and Shingo Takamatsu. BannerAgency: Advertising banner design with multimodal LLM agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. arXiv:2503.11060.

Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. GenArtist: Multimodal LLM as an agent for unified image generation and editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. Spotlight. arXiv:2407.05600.

Jamie Ward and Julia Simner. Lexical-gustatory synaesthesia: linguistic and conceptual factors. *Cognition*, 89(3):237–261, 2003.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation framework. In *Conference on Language Modeling (COLM)*, 2024a. arXiv:2308.08155.

Tsung-Han Wu, Long Lian, Joseph E. Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting LLM-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6327–6336, 2024b. arXiv:2311.16090.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. In *International Conference on Machine Learning*, pp. 55204–55224. PMLR, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, et al. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26340–26353, 2024a.

Jingyuan Yang, Jiawei Feng, and Hui Huang. Emogen: Emotional image content generation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6358–6368, 2024b.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal LLMs. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56704–56721. PMLR, 2024c. arXiv:2401.11708.

Zhengyuan Yang, Jianfeng Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. Idea2img: Iterative self-refinement with GPT-4V(ision) for automatic image design and generation. In *European Conference on Computer Vision (ECCV)*, 2024d. arXiv:2310.08541.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.03629.

Albert J Zhai, Yuan Shen, Emily Y Chen, Gloria X Wang, Xinlei Wang, Sheng Wang, Kaiyu Guan, and Shenlong Wang. Physical property understanding from language-embedded feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28296–28305, 2024.

A Appendix Outline

This appendix provides additional details on the Sensory Ad dataset, experimental setup, and implementation, followed by extended results and analyses. We also discuss ethical considerations and provide the prompts used in our experiments to support reproducibility. The annotation survey is provided as a separate supplementary file.

The appendix is organized as follows:

- **Dataset** (Section B)
  - Sensation taxonomy and data collection
  - Sensation and advertisement-topic diversity
  - Annotation protocol and quality control
  - Human–human agreement
- **Experimental Setup and Implementation Details** (Section C)
  - SenseScore training and preference-pair construction
  - SenseScore inference and score normalization
  - Image-description generation and examples
  - SenseGen generation settings and SAGA refinement
- **Additional Results and Analysis** (Section D)
  - **Sensation Classification (SenseClass)**
    - Hierarchical and single-level classification
    - Evaluation protocol and results
  - **Sensory Ad Generation (SenseGen)** (Section D.2)
    - Impact of sensation conditioning on image–message alignment and persuasion
    - Fine-tuning SD3 for sensory ad generation
    - Additional qualitative comparisons between real and generated advertisements
  - **Generation Performance Across Sensations** (Section D.3)
    - Sensory image generation across the taxonomy
    - Variation in evocation strength across sensation categories
- **Ethical Considerations** (Section E)
  - Potential risks of sensory advertising
  - Sensations that may be inappropriate for particular audiences
- **Prompts** (Section F)
  - Image-description and sensation-classification prompts
  - Sensory image and advertisement-generation prompts
  - SAGA agent prompts
- **Annotation Survey** (provided as a separate supplementary file)

B Dataset

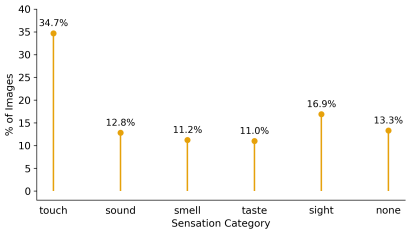
Taxonomy and Data Collection. To collect the dataset, we first defined the taxonomy shown in Fig. 10. We then randomly sampled 1100 images from the PittAd dataset Hussain et al. (2017), covering 96 sensations and more than 40 topics.

Sensation and Topic Diversity. Fig. 9a and Fig. 9b show the distribution over the 5 main sensations and the 10 most frequent topics, exhibiting diversity across both sensory modalities and topics.

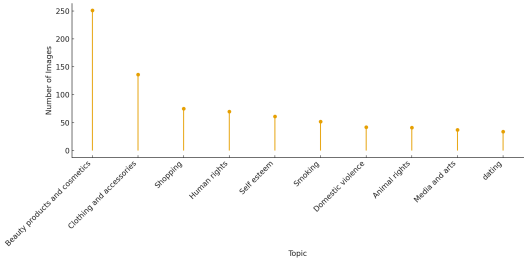
Annotations. For data annotation, we first had a test phase study on Prolific, gave the annotators detailed instruction with examples of images evoking each sensation, and selected a group of annotators based on the quality of their responses to do the main study. We used Qualtrics to create dynamic forms showing different options based on annotators choice in each step. The form is uploaded as the supplementary file.

The final annotations were collected from 200 annotators of different genders, aged 25–60, all with at least a high school diploma, an approval rate above 90% on more than 1000 prior annotations, and located in the United States. Each image was annotated by one annotator and subsequently reviewed by a skilled evaluator for quality assurance. In cases of disagreement (which were rare), the annotator was asked to justify their choice. If the justification was insufficient, the annotation was discarded and the image was returned to the annotation pool.

Human–Human Agreement. To assess reliability, we collected two independent annotations for approximately 10% of the images and computed κ. The human–human agreement is 0.8, indicating strong inter-annotator consistency.



(a) Diversity of images over 5 main sensations.



(b) Diversity of images over 10 most frequent topics in the SensoryAd dataset.

Figure 9: Image Distribution. Left: percentage of images per sensation category (sensation diversity). Right: distribution over the 10 most frequent topics (topic diversity).

C Experimental Setup & Implementation Details

In this section we explain the experimental setup. Hugging Face implementation of models are utilized, and code will be released upon the acceptance.

C.1 SenseScore Training

We fine-tune LLMs using LoRA Hu et al. (2022) on 297708 image-sensation pairs. To train our proposed evaluation metric, we randomly selected above 700 images from annotations to create our training data. In our proposed training, we pair each two sensations with different intensity (scores chosen by human annotators) as chosen and rejected. Each data point in our training, included description of the image, chosen sensation, rejected sensation, and parent of chosen sensation. This training data setting resulted in 297708 data point. We fine-tuned the LLMs - LLAMA3 (Meta-Llama-3-8B-Instruct), and QwenLM (Qwen2.5-7B-Instruct) - using LoRA Hu et al. (2022) with batch-size of 64, and learning rate 5e-5. Our evaluation of SenseScore performance was on a subset of the above 300 images not selected for training.

C.2 SenseScore Inference

To facilitate score interpretation, we first apply min-max normalization using the score range observed across all 96 sensations and all images in our initial experiments. We then standardize the resulting scores using the mean and standard deviation of target-sensation scores in generated advertisements.

C.3 Description Generation

We generate descriptions of images with 0-shot InternVL, Gemma, and QwenVL and utilize the same descriptions in assessing LLMs' capabilities on sensation classification tasks, and SenseScore evaluation. Fig. 11, represents two examples of descriptions generated by each of the MLLMs. As shown in the examples, given the prompt in Table 10 the models generate accurate descriptions of the image without interpreting the image. This prevents the information leakage in SenseScore

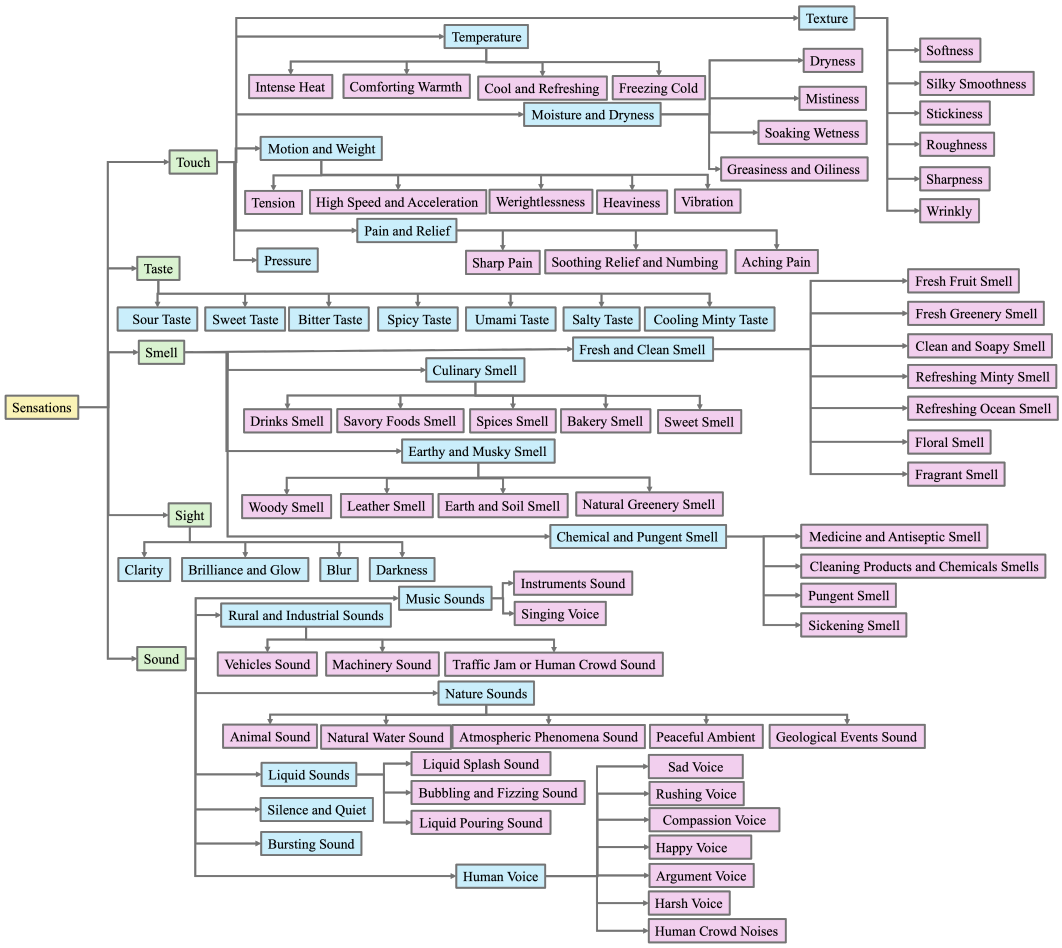


Figure 10: **Sensation Hierarchy.** First level, represents the main five human sensations, and each sensation is categorized into different set of sensations.



InternVL

Q1: Yes, there are objects in the image. The objects include: 1. A waterfall 2. Two wooden posts 3. A bird on the left wooden post 4. A bird in flight above the waterfall 5. Rocks at the base of the waterfall
Q2: The image features a large waterfall cascading down a cliff into a body of water below. Two wooden posts stand on either side of the waterfall, with a bird perched on the left post. Another bird is flying in the sky above the waterfall. The water at the base of the waterfall is turbulent, with visible rocks protruding from the water. The sky is partly cloudy, and the overall scene has a serene and natural appearance. In the bottom right corner, there is text that reads "Great wash! Water purification systems by SIEMENS."

Gemma

Q1: Waterfall, rocks, sky, clouds, clothespins, bird.
Q2: The image depicts a large waterfall cascading down a rocky cliff face into a body of water. The water is white and turbulent, creating a misty spray. Darker rocks are visible at the base of the waterfall. Above the waterfall, a blue sky with white, fluffy clouds dominates the upper portion of the image. Two wooden clothespins are attached to the waterfall at approximately the same height. A bird is flying across the sky. The bottom of the image shows the water flowing into a wider, calmer section of the body of water. The text "Great wash! Water purification systems by SIEMENS" is positioned at the bottom left corner of the image.

QwenVL

Q1: Yes, there are objects in the image, excluding text-only logos, and text. The objects are: - A bird flying in the sky - Two wooden clothespins
Q2: The image depicts a dramatic waterfall cascading over a rocky cliff into a body of water below. The water appears turbulent, creating mist as it falls. Two wooden clothespins are humorously placed at the top of the waterfall, one on each side, as if they are holding back the water. The sky above is partly cloudy with a soft blue hue. In the bottom right corner, there is text that reads "Great wash! Water purification systems by SIEMENS."



InternVL

Q1: Bottle, leaves, liquid splash, cap, label.
Q2: The image features a bottle of Chivas Regal 12 Years Old Blended Scotch Whisky. The bottle is prominently displayed against a warm, golden background. Green leaves are scattered around the bottle, with some leaves and a splash of liquid appearing to emerge from the bottle. The cap is purple, and the label on the bottle includes text and decorative elements. The text "CHIVAS Green Tea" is displayed at the top right of the image.

Gemma

"Q1: Bottle, green leaves, water splashes, dark background, and light reflections.
Q2: The image shows a brown bottle of Chivas Regal 12-year-old blended Scotch whisky. The bottle is positioned on a dark, slightly reflective surface. Several green leaves are scattered around the bottle, some with water splashes emanating from them. The background is dark, with hints of green and light reflections. The words "CHIVAS Green Tea" are displayed in a green, stylized font above the bottle. The bottle has a red-colored cap. There are also several more green leaves near the bottom of the image."

QwenVL

Q1: Yes, there are objects in the image, excluding text-only logos, and text. The objects are: - A bottle of Chivas Regal whisky - Green leaves - Water splashes
Q2: The image features a bottle of Chivas Regal 12-year-old Blended Scotch Whisky with a label prominently displayed. The bottle is set against a dark background with a warm, glowing light effect. Surrounding the bottle are green leaves and water splashes, creating a dynamic and refreshing visual. The text "CHIVAS Green Tea" is written in a stylized font, suggesting a fusion of the brand's signature whisky with green tea. The overall composition emphasizes the product's premium quality and the addition of green tea as a unique element.

Figure 11: **Image Description Examples.** Two examples of descriptions generated by InternVL, Gemma, and QwenVL. Both images are real advertisements from PittAdHussain et al. (2017) dataset.

while providing the accurate description of the image for LLMs in both classification and evaluation tasks. Negative agreements of zero-shot LLMs (LLAMA3-instruct) in the main text further reject the hypothesis of information leakage from MLLM description generation.

C.4 SensoryAd Generation

We benchmark different T2I models including Stable Diffusion 3 (Stable-diffusion-3-medium-diffusers), PixArt (PixArt-alpha/PixArt-XL-2-1024-MS), AuraFlow (AuraFlow-v0.3), Flux (FLUX.1-dev), and QwenImage (Qwen-Image) with 4-bit quantization on QwenImage and 8-bit quantization on rest of the models. We set the seed to 0 and number of time-steps as 28. For the rest of the model setting we use the default values. To generate the Sensory Ads, we utilized the sensation group (different level in hierarchy) evoked by the image with highest intensity generating an image for each sensation.

We show an example of conversation loop in Fig. 12.

D Additional Results & Analysis

D.1 Sensation Classification Tasks (SenseClass)

Interpreting sensory ads and evaluating their effectiveness hinges on understanding which sensations an image evokes and with what intensity. Moreover, certain sensations (e.g., pain) can be inappropriate for some audiences (e.g., children), making it important for automated systems to recognize the sensations conveyed by visual content. To formalize this, we introduce the Sensation Classification (SenseClass) task. We consider two formulations, hierarchical and single-level.

Hierarchical Classification. In this setting, we capture the hierarchical dependencies between sensations. Data annotation proceeds level by level: starting from the top-level categories, annotators choose the sensation best evoked by the image, then move to its subcategories, and so on until reaching a fine-grained leaf. The hierarchical classification task mirrors this process. Given an image, the goal is to predict the complete sensation path(s) from the root to the leaf node (e.g., Touch $\rightarrow$ Temperature $\rightarrow$ Freezing Cold). A model is recursively prompted to predict up to three sensations, advancing down the hierarchy by selecting among the children of each previously chosen node. To provide sufficient context, the definition of each potential sensation was included in the prompt.

Single-Level Classification. This task flattens the taxonomy and treats every sensation, regardless of its level, as a potential label. The goal is to predict the complete set of sensations that an image evokes. A critical constraint in this task is maintaining hierarchical consistency. By definition, if an image evokes a specific sensation, it must also evoke its parent sensation (e.g., if "Temperature" is evoked, "Touch" is evoked as well). To evaluate a models' understanding of these relationships, we define an additional metric: Parent Recall (R_{parent}), which measures the fraction of predicted non-root sensations for which the direct parent sensation was also predicted. It is formally defined as:

$$R_{parent} = \frac{|\{s \in S_{pred} \mid s \text{ is not a root node and } parent(s) \in S_{pred}\}|}{|\{s \in S_{pred} \mid s \text{ is not a root node}\}|} \tag{2}$$

where S_{pred} is the set of sensations predicted by the model. A high R_{parent} score indicates that the model understands the hierarchical dependencies of sensations.

D.1.1 Sensation Classification Evaluation

Our evaluation follows distinct protocols based on the model's input modality. For MLLMs, the ad image was provided as direct visual input. The model was then tasked with classifying the corresponding sensations based on a task-specific prompt. To assess the performance of text-only LLMs, we employed a two-stage pipeline. First, we utilized different MLLM (InternVL, QwenVL, and Gemma) to generate a description for the image (D_{MLLM}). These generated descriptions were utilized as input context for the LLMs to perform sensation classification. This approach allows us to isolate and evaluate the language-based reasoning capabilities of LLMs for this specific task. We report Recall (R), Precision (P), and F1-score (F1). For Single-Level Classification, we also report the Parent Recall (R_{parent}) to assess understanding the hierarchical relations.

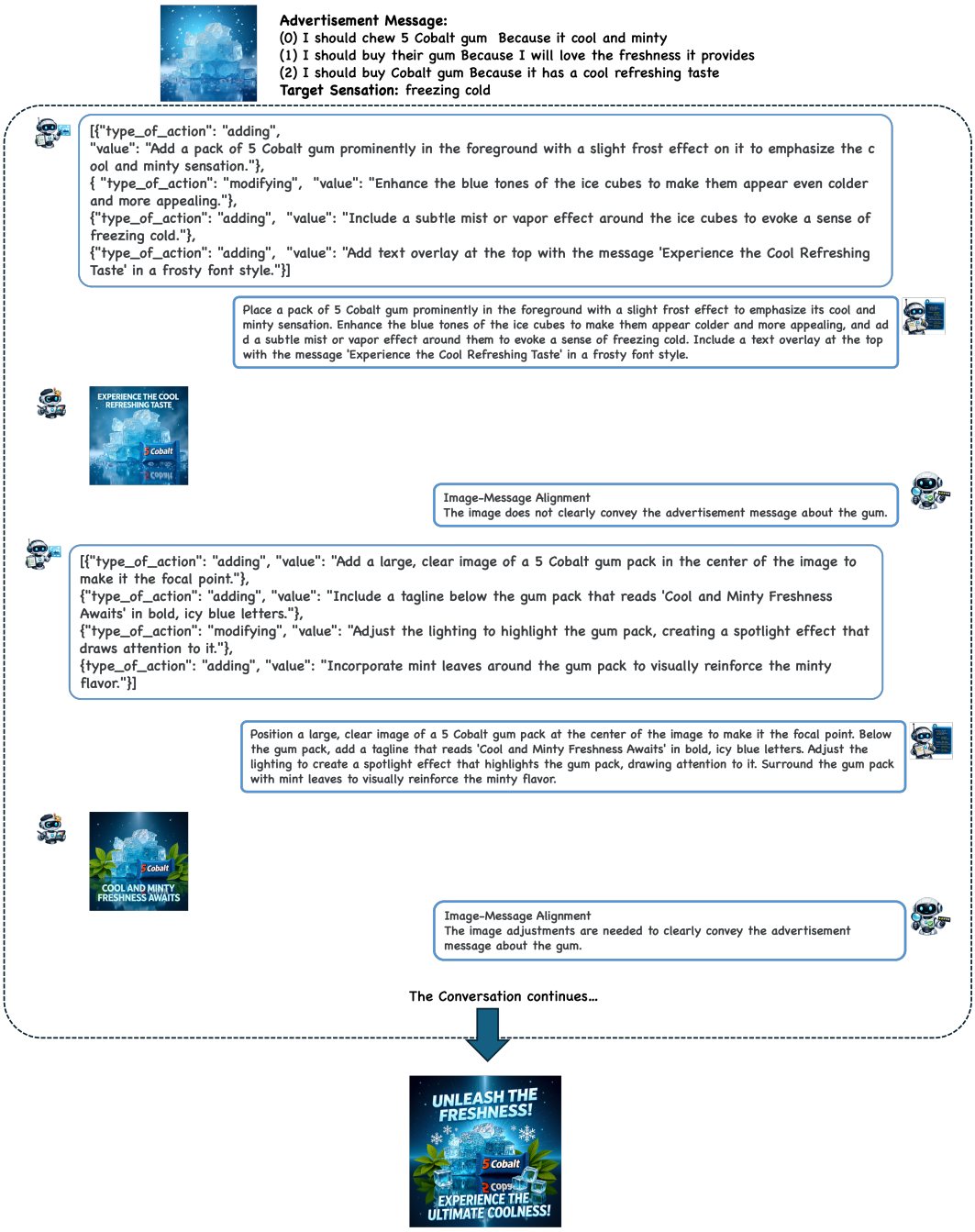


Figure 12: **Conversation loop in SAGA.** An example of two steps of conversation loop among agents with FluxKontext image editing. The input of the conversation loop is the advertisement message, target sensation, and the image generated by Flux for conveying the advertisement messages and targeting the sensation.

Table 4: Results on classification. "-" denotes model did not follow instructions.

Model	Hierarchical Classification			Single-Level Classification			
	P	R	$F1$	P	R	$F1$	R_{parent}
MLLMs							
QwenVL	0.17	0.62	0.27	**0.33**	0.18	0.23	0.45
InternVL	0.13	0.60	0.21	0.18	0.44	0.26	0.41
LLAVA-Next	0.10	0.60	0.17	-	-	-	-
GEMMA	0.17	**0.66**	0.27	0.11	0.39	0.17	0.49
LLMs							
QwenLM + D_{QwenVL}	0.18	0.45	0.26	0.18	0.42	0.25	0.24
QwenLM + $D_{InternVL}$	0.18	0.44	0.26	0.18	0.42	0.25	0.22
QwenLM + D_{GEMMA}	0.18	0.45	0.26	0.19	0.44	0.27	0.24
GEMMA + D_{QwenVL}	0.16	0.54	0.25	0.13	**0.54**	0.21	0.65
GEMMA + $D_{InternVL}$	0.15	0.54	0.23	0.13	**0.54**	0.21	0.64
GEMMA + D_{GEMMA}	0.15	0.55	0.24	0.14	**0.54**	0.22	**0.68**
LLAMA3 + D_{QwenVL}	0.19	0.43	0.26	0.15	0.47	0.23	0.48
LLAMA3 + $D_{InternVL}$	**0.21**	0.43	**0.28**	0.13	0.47	0.20	0.45
LLAMA3 + D_{GEMMA}	0.20	0.43	0.27	0.13	0.48	0.20	0.46

Hierarchical Classification. Table 4 reveals a consistent trend across all models: significantly higher recall than precision. This imbalance indicates that while models are proficient at identifying potentially relevant sensations, they struggle to reject incorrect ones. The best F1 numbers are found among MLLMs, and MLLMs often outperform their LLM counterparts on this task (e.g., QwenVL / QwenLM, GEMMA / GEMMA). This suggests that direct visual input provides crucial cues that may be lost or distorted in text-only descriptions.

Single-Level Classification. The results in Table 4 show that while MLLMs achieve higher precision and F1-scores, some LLMs (GEMMA) have a stronger performance on Parent Recall (R_{parent}). This suggests that LLMs, operating on textual descriptions and definitions, develop a better understanding of the abstract, semantic relationships between sensations in the hierarchy.

D.2 Sensory Ad Generation

This section presents additional experiments on sensory advertisement generation, including an analysis of sensory exaggeration, fine-tuning SD3 for SenseGen, and qualitative comparisons between real and generated advertisements.

D.2.1 Exaggeration

To further evaluate the impact of exaggeration of sensation by T2I models, we generate images both with sensation and without sensation and compare the alignment of generated images with advertisement messages in each case. As observed in Table 5, when adding sensation evocation AIM score drops in Flux, AuraFlow, PixArt and Qwen-Image, and stays unchanged in SD3. The drop in AIM score suggest that adding the sensation model ignores the advertisement message. However, adding sensation generally positively impacts persuasion (last column).

D.2.2 Fine-tuning SD3 on SensoryAd Generation

To further analyze the capability of T2I models, we fine-tuned the SD3 model on SensoryAd data. After fine-tuning text-image alignment of images increases by 0.01 compared to 0-shot SD3 and sensation intensity stays unchanged. We hypothesize, this is the result of implicitness of the text input as previously suggested by Aghazadeh & Kovashka (2025) making the generation task more challenging.

Table 5: Evaluating generated sensory ads.

T2I model	Sensory Ad Input	AIM	P_{comp}
Flux	AR	0.66	0.54
	AR + Sensation	0.60	0.56
SD3	AR	0.62	0.58
	AR + Sensation	0.62	0.59
AuraFlow	AR	0.62	0.53
	AR + Sensation	0.60	0.57
PixArt	AR	0.65	0.61
	AR + Sensation	0.60	0.61
Qwen-Image	AR	0.66	0.55
	AR + Sensation	0.61	0.57
SAGA (FLUX)	AR	0.62	0.61
	AR + Sensation	**0.66**	**0.62**

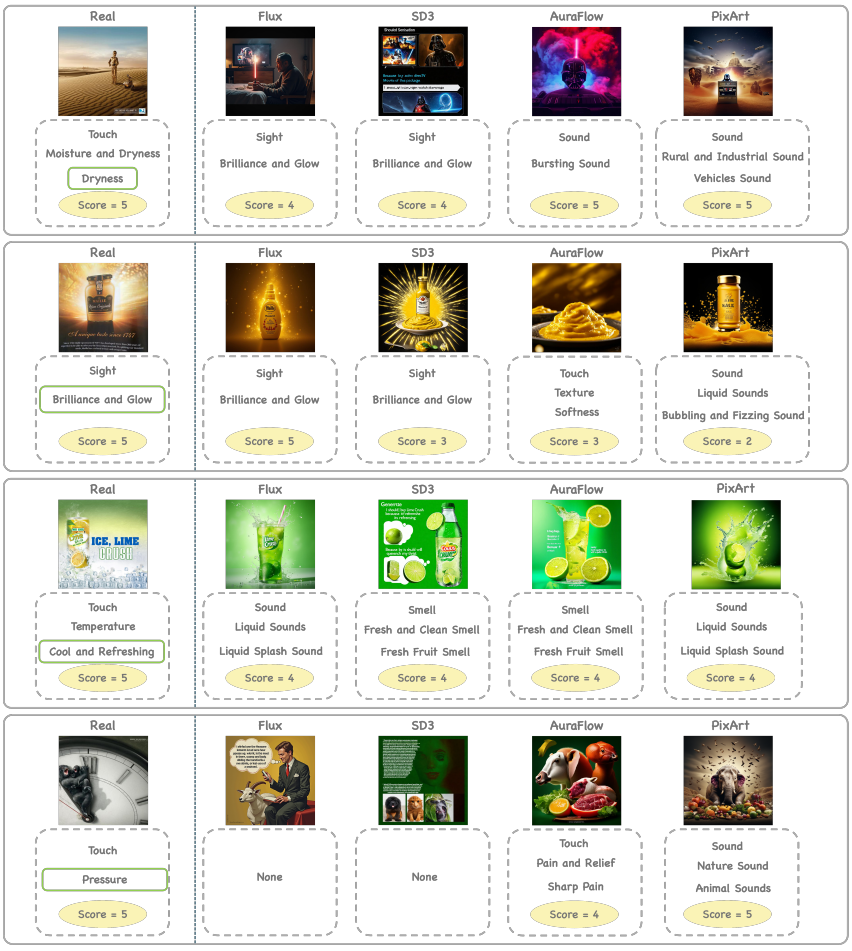


Figure 13: **Sensory Ad examples.** Four examples of real advertisement and generated advertisements by Flux Black Forest Labs (2024), SD3 Esser et al. (2024), AuraFlowFal (2024), and PixArt Chen et al. (2024a) given the action-reason interpretation and sensation annotation for the real advertisement. Green border represents the sensation used in the prompt of T2I models.

D.3 Variation in generation performance across sensations

To analyze the capability of the T2I model in generating images evoking each sensation in our taxonomy, we isolated the sensation and only prompted the model to 'Generate an image that evokes {sensation}' with seeds from 0 to 9 resulting in 960 images per model and 4800 images in total. Fig. 14 represents the intensity of different sensations evoked in Sensory Image generation task. As shown in Fig. 14, models struggle more in evoking sensations with less common visual representation such as different types of human voices, or in overall different sounds. In contrast, models can evoke visual sensations - Sight and its children - with high intensity.

Fig. 13, shows an example of advertisements generated evoking four different sensations. Fig. 13, further represents the difference between capabilities of T2I models in evoking visual sensations like "Brilliance and Glow" and more abstract sensations like "Pressure".

E Ethical concerns & AI Usage

E.1 Ethical concerns around sensory advertisements

There are two main implications: First, the generation of adversarial persuasive content such as encouraging the audience to drink alcohol more often. This concern is not unique to our approach but is inherent to any T2I systems. Second, the model might generate sensitive content for a certain group of audiences and this is one of the motivations for classification tasks. While automatically generating the Sensory Ads can be helpful, some sensitive sensations (for example pain) should be detected and prevented from being shown to a specific audience groups. This is why it is also important to be able to classify the sensations evoked by the image.

F Prompts

We have included prompts in Tables 10 (for image description generation), 9, 8, 7, 6 (for Sensation Classification tasks), and 12, 11 (for Sensory Image Generation tasks).

Table 6: Prompt for LLM Hierarchical Sensation Classification

Prompt
System: You are a helpful assistant, choosing the sensations evoked by the described image given the following definition in ordered form. You can choose up to 3 sensations evoked by the image ranked in order of how well the sensations are evoked. If the image does not evoke any sensation, you can choose None. **Context**: Sensation is the process of detecting and receiving information from the environment or the body through specialized sensory organs, which send signals to the brain for interpretation. **Definition of the sensations in the options**: {{context}} **User**: What are the sensations evoked the most by the described image? Only return the indices of maximum of 3 options in ordered form without any further explanation. **Image Description**: {{description}} **Options**: {{options}} Your answer must follow the following format: **Answer:** <indices of maximum of 3 correct options separated by comma>

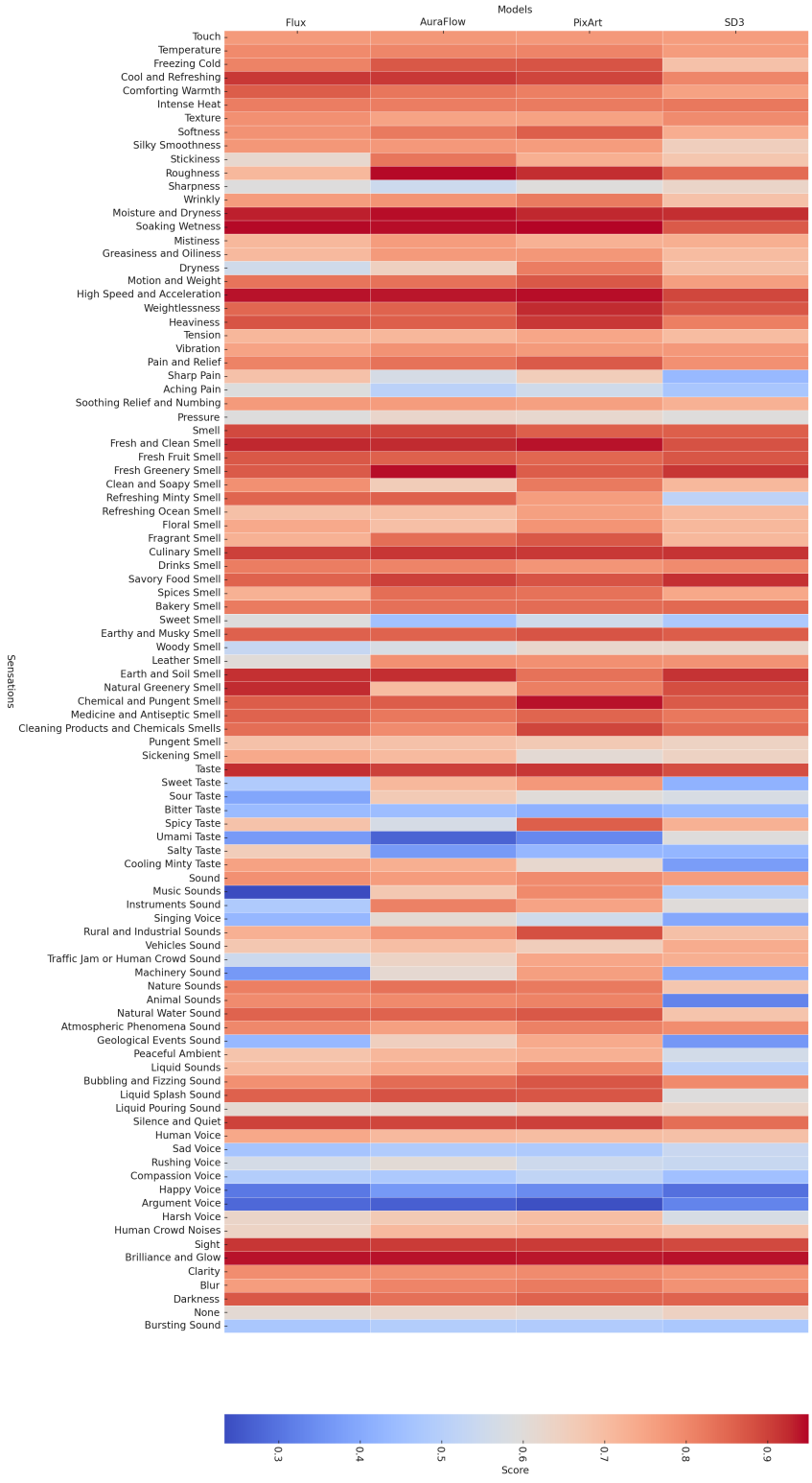


Figure 14: **Sensation Heatmap.** Average SenseScore score for images generated by each model for each sensation. Each model generates ten images evoking each sensation.

Table 7: Prompt for MLLM Hierarchical Sensation Classification

Prompt
System: You are a helpful assistant, choosing the sensations evoked by the input image given the following definitions. You can choose up to three sensations, ranked by how strongly they are evoked. If the image does not evoke any sensation, choose None. **Context**: Sensation is the process of detecting and receiving information from the environment or the body through specialized sensory organs, which send signals to the brain for interpretation. **Definition of the sensations in the options**: {{context}} **User**: Which sensations are most strongly evoked by this image? Return only the indices of up to three options, ordered by evocation strength, without any further explanation. **Options**: {{options}} Your answer must follow this format: Answer: <indices of up to three options separated by commas>

Table 8: Prompt for LLM Multi-choice Sensation Classification

Prompt
System: You are a helpful assistant tasked with identifying the sensations evoked by a described image. Select all sensations evoked by the image and rank them by how strongly they are evoked. If the image does not evoke any sensation, choose None. **Context**: Sensation is the process of detecting and receiving information from the environment or the body through specialized sensory organs, which send signals to the brain for interpretation. **Definition of the sensations in the options**: {{context}} **User**: Which sensations are evoked by the described image? Return only the indices of all applicable options, ordered by evocation strength, without any further explanation. **Image Description**: {{description}} **Options**: {{options}} Your answer must follow this format: Answer: <indices of all applicable options separated by commas>

Table 9: Prompt for MLLM Multi-choice Sensation Classification

Prompt
System: You are a helpful assistant tasked with identifying the sensations evoked by the input image. Select all sensations evoked by the image and rank them by how strongly they are evoked. If the image does not evoke any sensation, choose None. **Context**: Sensation is the process of detecting and receiving information from the environment or the body through specialized sensory organs, which send signals to the brain for interpretation. **Definition of the sensations in the options**: {{context}} **User**: Which sensations are evoked by this image? Return only the indices of all applicable options, ordered by evocation strength, without any further explanation. **Options**: {{options}} Your answer must follow this format: Answer: <indices of all applicable options separated by commas>

Table 10: Prompt for Structured Description Generation

Prompt
Carefully analyze the image and respond only in the specified format, without any interpretations or inferences. Focus on only the visible elements in the image. Ensure that any object seen in the image is included in Q1, even if it is described in more detail in Q2. **Response Format:** Q1: ${answer to Q1} Q2: ${answer to Q2} **Questions:** Q1: Are there any objects in the image, excluding text-only logos, and text? List at most 5 such objects if present. Q2: Describe the image in detail, focusing only on visible objects and elements without adding any interpretation, opinion, or analysis in a single paragraph.

Table 11: Prompt for Sensory Image Generation

Prompt
Generate an image that evokes {{sensation}} sensation.

Table 12: Prompt for Image Generation with Action-Reason and Sensation

Prompt
Generate an advertisement image that evokes {{sensation}} sensation and conveys the following messages: {% for statement in action_reason %} - {{statement}} {% endfor %}

Table 13: Prompt for Image Editing Planner Agent

Prompt

System: You are an image-editing instruction planner agent. Given an image of an image, your task is to generate a sequence of concrete visual edits that should be applied to the image in order to:

1. Convey the intended advertisement message, and

2. Evoke the specified sensation (e.g., refreshment, heat, softness, luxury).

When you receive an issue from the critic, you MUST focus your edits on addressing that SPECIFIC issue:

Image-Message Alignment: The image does not clearly convey the advertisement message.

Focus on making the product or brand more prominent, ensuring the image directly relates to the message, adding visual elements that reinforce the message, and improving composition to highlight the key message.

Sensation Evocation: The image does not effectively evoke the target sensation.

Focus on adding visual cues that directly evoke the sensation (heat, cold, softness, etc.), adjusting colors, lighting, or texture to create the sensation, and adding atmospheric elements that reinforce the sensation.

Output Format Requirement:

You MUST output ONLY a valid JSON array in the following format:

```
[ { "type_of_action":  "<adding|removing|modifying|changing_style>",
"value":  "<editing instruction>" } ]
```

Guidelines:

Actions must be image-grounded, realistic, and minimal. Describe what to change rather than how to technically implement it. Be explicit about visual attributes such as color, texture, lighting, scale, position, motion cues, and atmosphere. When an issue is identified, all actions must directly address that specific issue type. If previous attempts failed, generate completely different actions and never repeat previous approaches.

Table 14: Prompt for Sensation Finder Agent

Prompt

System: You are a sensation finder agent.

Given an advertisement message and a list of sensations, your task is to choose a creative and relevant sensation that should be evoked by the image to improve the impact of the image.

First explain which sensation should be evoked and why it improves the impact of the image. Then choose only one sensation that should be evoked to improve the image impact.

Output Format:

```
<explanation>
The best sensation to evoke is:  <Sensation>
```

Table 15: Prompt for Sensation and Advertisement Message Finder Agent

Prompt

System: You are a sensation and advertisement message finder agent.
 	
Given an advertisement message and a list of sensations, your task is to choose a creative and relevant sensation that should be evoked by the image to improve the impact of the image. Next, choose the single best advertisement message that is most descriptive of the possible image and has the strongest message.
 	
First explain which sensation should be evoked and why it improves the impact of the image. Then choose only one sensation that should be evoked. Next, explain the advertisement message that should be used and why it is the best message to improve the impact of the image. Finally choose only one advertisement message.
 	
Output Format:
`<explanation>`
`Advertisement Message and Sensation: <Advertisement Message>, <Sensation>`

Table 16: Prompt for Text Refiner Agent

Prompt
System: You are a text refiner agent. Your task is to convert structured image-editing instructions in JSON format into a single, clear, concise, and visually grounded natural language prompt suitable for guiding an image editing model. **Requirements**: You will receive JSON instructions containing actions such as `adding`, `modifying`, `removing`, and `changing_style`. Convert all instructions into one cohesive natural language description of the final edited image. Do not output JSON. Output only plain text. Do not start with phrases such as "create an image" or "generate an image". Write as if describing what the edited image should look like and combine all actions into a single flowing description. **Guidelines**: Preserve factual consistency with the provided instructions, use precise visual language describing objects, attributes, colors, lighting, textures, and spatial relations, and do not invent elements not implied by the instructions. The output must only contain the refined prompt text with no explanations, JSON, commentary, or markdown. Write in present tense describing the final state of the image.

Table 17: Prompt for Image Evaluation Critic Agent

Prompt

System: You are a strict image evaluation agent working in a multi-agent environment.

Your task is to evaluate the image and output the issue of the image using the following format:

```
<Issue>
<one sentence explanation>
```

You must never copy or paraphrase previous message content, describe the image in full sentences, or add unnecessary reasoning or commentary.

Possible Issues:

- Visual Element Inconsistency
- Image-Message Alignment
- Sensation Evocation

Evaluation Criteria:

Visual Element Inconsistency refers to incoherent or conflicting visuals such as artifacts, glitches, or contradictory elements.

Image-Message Alignment refers to cases where the advertisement message is not clearly conveyed, the product is not prominent, or the image does not reinforce the message.

Sensation Evocation refers to cases where the target sensation is weak or not effectively evoked through visual cues, colors, lighting, objects, or atmosphere.

Priority Rule:

If visual inconsistencies exist, choose Visual Element Inconsistency. Otherwise if the message is unclear choose Image-Message Alignment. Otherwise if the sensation is weak choose Sensation Evocation.

Output Requirement:

Output exactly one issue label followed by a single sentence explanation.