Binding Touch to Everything: Learning Unified Multimodal Tactile Representations

Supplementary Material

A.1. Datasets and Metrics

We provide more details of datasets used in our paper, all of which are publicly available.

Touch and Go [12]. The Touch and Go dataset is a recent, real-world visuo-tactile dataset featuring human interactions with various objects in both indoor and outdoor environments using a GelSight tactile sensor. It comprises 13,900 instances of touch across approximately 4,000 distinct object instances and 20 types of materials. Since it is the only real-world in-the-wild dataset, we apply it to multiple tasks including material classification, image synthesis with touch, Touch LLM, and X-to-touch generation. We use the official train/test split of [12] where the dataset is split by touches, not by frames to avoid similar touch images between the train and test set. For Touch-LLM and X-to-touch applications, we label 400 visual images by asking turkers to provide their captioning to describe the object, touch feeling, and texture from it.

The feeling of success [2]. The Feeling of Success is a robot-collected visuo-tactile dataset of robots grasping objects on a tabletop. The tactile images are all captured by GelSight tactile sensors. It contains 9.3k paired vision and touch images. We apply this dataset to robotic grasping stability predictions. As there is no official split of train/val/test, following [6, 12], we split the dataset by objects in the ratio of 8:1:1.

YCB-Slide [10]. The YCB-Slide dataset comprises DIGIT sliding interactions on YCB objects. The dataset is in the video format where we take all 180k frames for our experiments. The dataset contains 10 YCB objects including a sugar box, a tomato soup can, a mustard bottle, a bleach cleanser, a mug, a power drill, scissors, an adjustable wrench, a hammer, and a baseball. While the tactile images are collected via sliding interaction, the visual input is generated by simulation of the YCB objects. In our experiment, we treat each of the objects as an individual material and our goal is to classify 10 classes. We apply this dataset to material classification.

ObjectFolder 1.0 [4]. The ObjectFolder 1.0 dataset is a simulation dataset containing 3D models of 100 objects from online repositories. The touch images are simulated by TACTO simulators. As the raw dataset is a 3D model with infinite points, we randomly sample 200 points for each object. We apply this dataset to material classification and grasping stability prediction experiments. It is worth noting that for grasping stability prediction experiments, we select 6 objects suitable for grasping following their setting and achieve relatively balanced successful and failure outcomes for grasping. Following [4], all materials can be categorized into 7 material categories including wood, steel, polycarbonate, plastic, iron, ceramic, and glass. These categories are also applied to ObjectFolder 2.0 and ObjectFolder Real datasets.

ObjectFolder 2.0 [5]. The ObjectFolder 2.0 dataset extends [4] to 1000 objects and improves the acoustic and tactile simulation pipelines to render more realistic multisensory data. For the tactile simulation, it utilizes the Taxim simulator instead of TACTO. Similar to the preprocessing of ObjectFolder 1.0, we sample 200 points for each object. To avoid overlapping with [4], we only take the 101-1000 objects. We apply this dataset to material classification, cross-modal retrieval, robot grasping stability prediction, and Touch-LLM. For cross-modal retrieval and Touch-LLM tasks, we annotate text descriptions that depict the contact point of the object from its visual input, *e.g.* "`The corner of a wooden table.`"

ObjectFolder Real [6]. ObjectFolder Real is an object-centric multimodal dataset containing 100 real-world household objects. The touch images are captured by the GelSlim tactile sensor. Similarly, we sample 200 points for each object thus containing in total of 20k visuo-tactile pairs. We apply this dataset to a material classification task, which is considered an out-of-domain dataset.

SSVTP [7]. SSVTP dataset is a recent human-collected visuo-tactile dataset containing 4.9k paired visuo-tactile images. The touch images are collected via the DIGIT tactile sensor. The objects in this dataset are mainly from garments but also contain materials of metal. We apply this dataset to material classification. As the dataset does not contain material labels, we annotate material labels from the visual images. In total, we classify all images into 6 material categories including cotton, metal, denim fabric, plastic, wood, and nylon.

A.2. Implementation Details

We show more implementation details in this section.

Image synthesis with touch. We used a pretrained stable diffusion-2.1 unclip [8] to perform zero-shot touch-to-image generation by replacing the text condition with our aligned UniTouch embedding. Specifically, we keep the simple text `"high quality"` as the condition while using our touch embedding as an additional condition. We use DDIM

sampler [9] with a guidance scale of 9 and denoising steps of 50. Additionally, we set an embedding strength of 0.75 for our touch embedding condition. Synthesized images are at the resolution of 768×768.

As for tactile-driven image stylization, similarly, we still keep the simple text "high quality" as the condition. However, we use both touch and image embeddings as extra conditions to conduct image stylization. We perform a linear combination of touch and image embeddings, the weights for touch and image are set to 0.3 and 0.7 respectively. We use DDIM sampler [9] with a guidance scale of 9 and denoising steps of 50. The strength for linear combination embedding is set to 1 and edited images are at the resolution of 768×768.

Touch-LLM. We adapt our model from [3, 14], which leverages an adapter to connect our touch encoder and an open-source large language model LLaMA [11]. We replace RGB image embedding with our aligned UniTouch embedding. Concretely, we denote the global touch feature encoded by our touch encoder as $F_T \in \mathbb{R}^{1\times C_T}$, where C_T is the dimension of the touch embedding. Inspired by prior work [3, 14], we use a projector f, which encodes F_T to have the same dimension as the token embedding in LLaMA [11]:

$$F'_T = f\left(F_T\right). \tag{1}$$

Then we repeat F'_T and add it to all text tokens across all layers in language model LLaMA [11] with a zero-initialized learnable gate function:

$$T_j^q = h_{\text{zero}} \cdot F'_T + T_j^q, \tag{2}$$

where j and q denotes the layer and sequence index respectively, T_j^q is the text token embedding, and h_{zero} is the zero-initialized learnable gate function. In our experiments, we use pretrained h_{zero}, and plug our UniTouch embedding in.

X-to-touch generation We conduct our X-to-touch generation model based on stable diffusion. While most existing multimodal tactile datasets only contain vision and touch, we first train an image-to-touch diffusion model and we are able to conduct text-to-touch and audio-to-touch *zero shot* by replacing the image conditioning as they are already aligned. We use the Adam optimizer with a base learning rate of 1e-6. Models are all trained with 30 iterations using the above learning rate policy. We train our model with a batch size of 48 on 4 RTX A40 GPUs. Since we want to use the aligned condition embeddings, the conditional model is frozen during training. The condition embeddings are integrated into the model using cross-attention. We use the frozen, pretrained VQGAN to obtain our latent representation, with a spatial dimension of 64×64. During the inference, we conducted the denoising process for 200 steps and set the guidance scale s = 7.5.

A.3. Evaluation Details

Touch-to-image generation Following [13], we use three evaluation metrics of Frechet Inception Distance (FID), Contrastive Visuo-Tactile Pre-Training (CVTP), and Material Classification Consistency. FID is a standard evaluation metric in image synthesis that compares the distribution of real and generated image activations using a trained network. CVTP [13] is a metric similar to CLIP but measures the cosine similarity between the visual and tactile embeddings learned for the generated images and conditioned tactile signals, which used an off-the-shelf network. Material classification consistency [13] uses a material classifier to categorize the predicted and ground truth images and measure the rate at which they agree, where we use CLIP as the zero-shot material classifier by feeding the prompt of "material of [CLS]".

Touch-LLM. We feed each vision language model (including our Touch-LLM) with a touch image and text prompt: "You will be presented with a touch image from an object/surface. Can you describe the touch feeling and the texture?". In the end, we use GPT-4 to perform the automatic evaluation for each model following prior work [1]. Specifically, we provide GPT-4 with: 1) a system prompt describing the desired evaluation behavior; 2) the question; and 3) a human-crafted reference response; 4) each model's generation result (more details see supp.). We instruct GPT-4 to rate each model's generations on a scale of 1 to 5 given the reference response. The template is shown in Fig. 2.

X-to-touch. We test the effectiveness of the x-to-touch model on the Touch and Go dataset, which is the only real-world dataset that contains objects and scenes in the wild. As the objects in this dataset are closely related to the material properties, we measure the material classification consistency between different touches generated from different modalities. We use our UniTouch embedding as the off-the-shelf zero-shot material classifier. For quantitative results for text-to-touch generation, we use the 400 human-labeled text captions as the input. For audio-to-touch generation, as there is no impact sound correlated to this dataset, we manually select audios from ObjectFolder 2.0 as the input that have the same material properties or geometry with the visual image for qualitative evaluations, as shown in Fig. 5.

A.4. Additional Experiments

In-batch sampling mix rate selection. We evaluate different choices of σ for in-batch sampling, where σ denotes the percentage of the data that comes from the same dataset while the rest from others. We set σ to $\{0, 0.5, 0.75, 1.0\}$ and evaluate their zero-shot material classification performance on all six datasets, as shown in Fig. 1. We observe

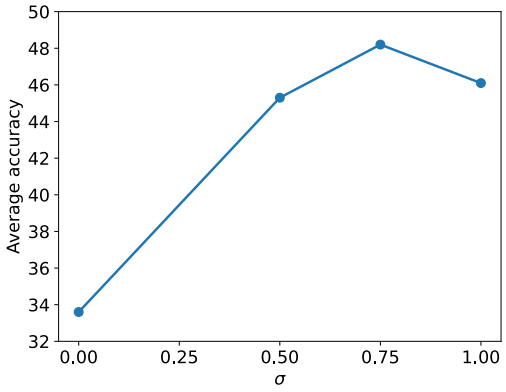

Figure 1. **Effect of σ for in-batch sampling.** We compare the average zero-shot material classification accuracy from six datasets using different σ of 0, 0.5, 0.75, 1.

that if we select $\sigma = 0$, the ability to distinguish between intra-sensor samples is significantly undermined thus leading to inferior performance. As the σ is increasing, the model is able to better distinguish between intra-sensor samples. In the extreme case when $\sigma = 1.0$ where all samples come from the same dataset, the model will have no exposure to the inter-class negatives. We observe that the performance in this case is actually decreasing. This demonstrates the effectiveness of design to balance between inter-sensor and intra-sensor negatives. We empirically found that selecting $\sigma = 0.75$ obtains a good trade-off between these factors.

Image synthesis with touch. We leverage our aligned UniTouch embedding and pretrained text-to-image stable diffusion model [8] to generate more qualitative results of touch-to-image generation and tactile-driven image stylization as presented in Fig. 3. It shows that our UniTouch embedding can guide image synthesis successfully in a zero-shot manner.

X-to-touch generation. We show more examples of X-to-touch generations on the Touch and Go [12] dataset in Fig. 5, where we generate touch images using image, text, and audio.

Touch-LLM. We show more touch image question answering examples in Fig. 4.

system prompt (human authored)

You are EvaluationGPT, an expert language model at judging whether or not a response adequately answers the question. More specifically, you will be given the following:

1. A question: This is a question which requires machine to generate a reasonable answer.
2. A high-quality reference answer: This is a human crafted answer, which should address the question.
3. A candidate response: This is a response generated by model trying to answer the question.

Your job is to judge whether the response adequately answers the question given the reference answer. Please output the number from 1 to 5.

- 1 represents the response is totally irrelevant to the question. 5 means the response fully addresses the question given reference answer.

user (human authored)

I will provide question to you. Then, I will provide a reference answer which is an example of a high quality output for that question in the context of the image (image is not provided). Then I will offer you a candidate response that tries to address the question. Your job is to rate the response on a scale of 1 to 5.

assistant (human authored)

Sure, please provide the question, reference answer, and the candidate response. Then I will rate the response on a scale of 1 to 5 to judge whether the response adequately address the question.

user (human authored)

OK. Here is the question, the high-quality reference answer, and the candidate response.

Question: You will be presented with an touch image from a object/surface. Can you describe the touch feeling and the texture?

Reference: The surface appears to be a piece of fabric. The fabric looks soft and textured with a knit or woven pattern. Touching it would likely feel warm, soft, and slightly bumpy due to the textures. The material would be flexible and malleable, conforming easily to pressure and touch, providing a comfortable and gentle tactile experience.

Response: The touch image shows a soft, fuzzy, and cozy texture, which is characteristic of the material used to make the sweater. The sweater is likely made from a warm and comfortable material, such as cotton or a blend of cotton and other fibers. The texture of the sweater is smooth and inviting, making it an ideal choice for a comfortable and relaxed outfit.

Figure 2. **GPT-4 evaluation template.** We use this template to instruct GPT-4 for automatic evaluation of our Touch-LLM and other selected open-source VLM baselines.

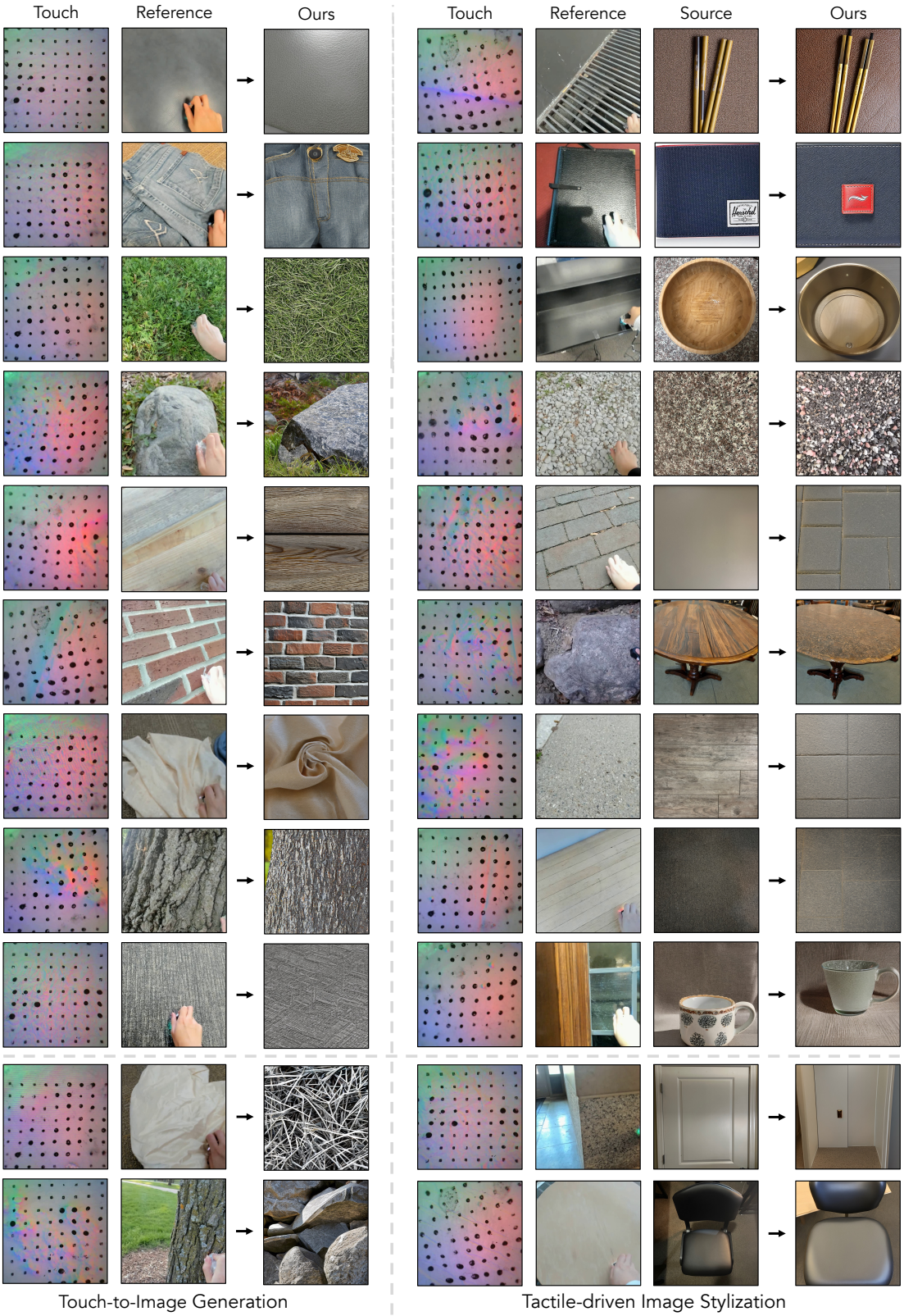

Figure 3. **More examples of zero-shot image synthesis with touch.** (Left) We generate an image of a scene given a tactile signal. (Right) We perform tactile-driven image stylization to manipulate an image to match a given touch signal. We denote "reference" as visual images paired with the input touch in the dataset, which are not seen by the model but *only* shown for demonstration purposes. The last two rows are failure cases.

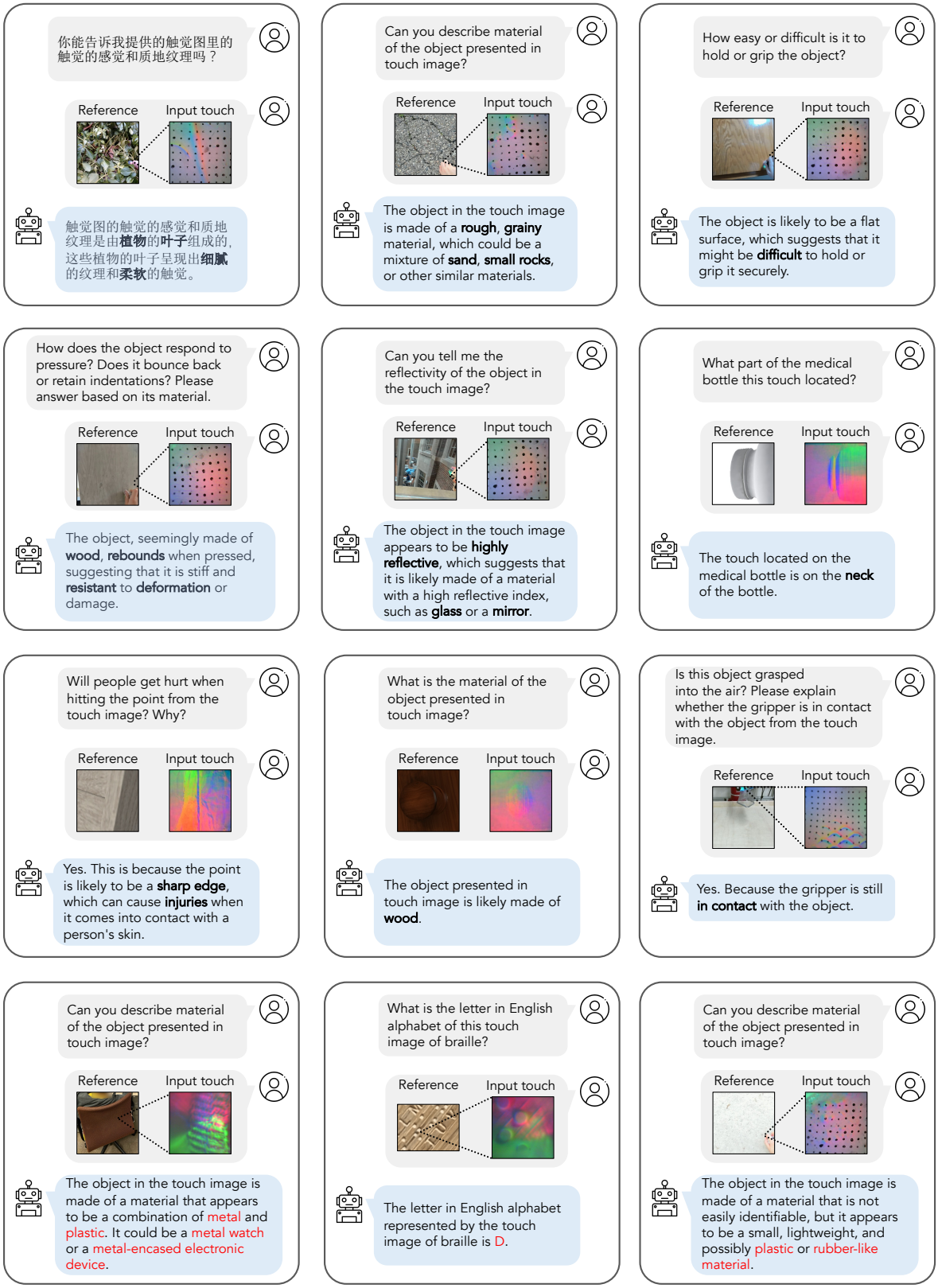

Figure 4. **More examples of Touch-LLM.** We show more question-and-answering examples for touch images using our Touch-LLM. We denote "reference" as visual images paired with the input touch in the dataset, which are not seen by the model but *only* shown for demonstration purposes. The last row is the failure case. Incorrect portion is highlighted in red.

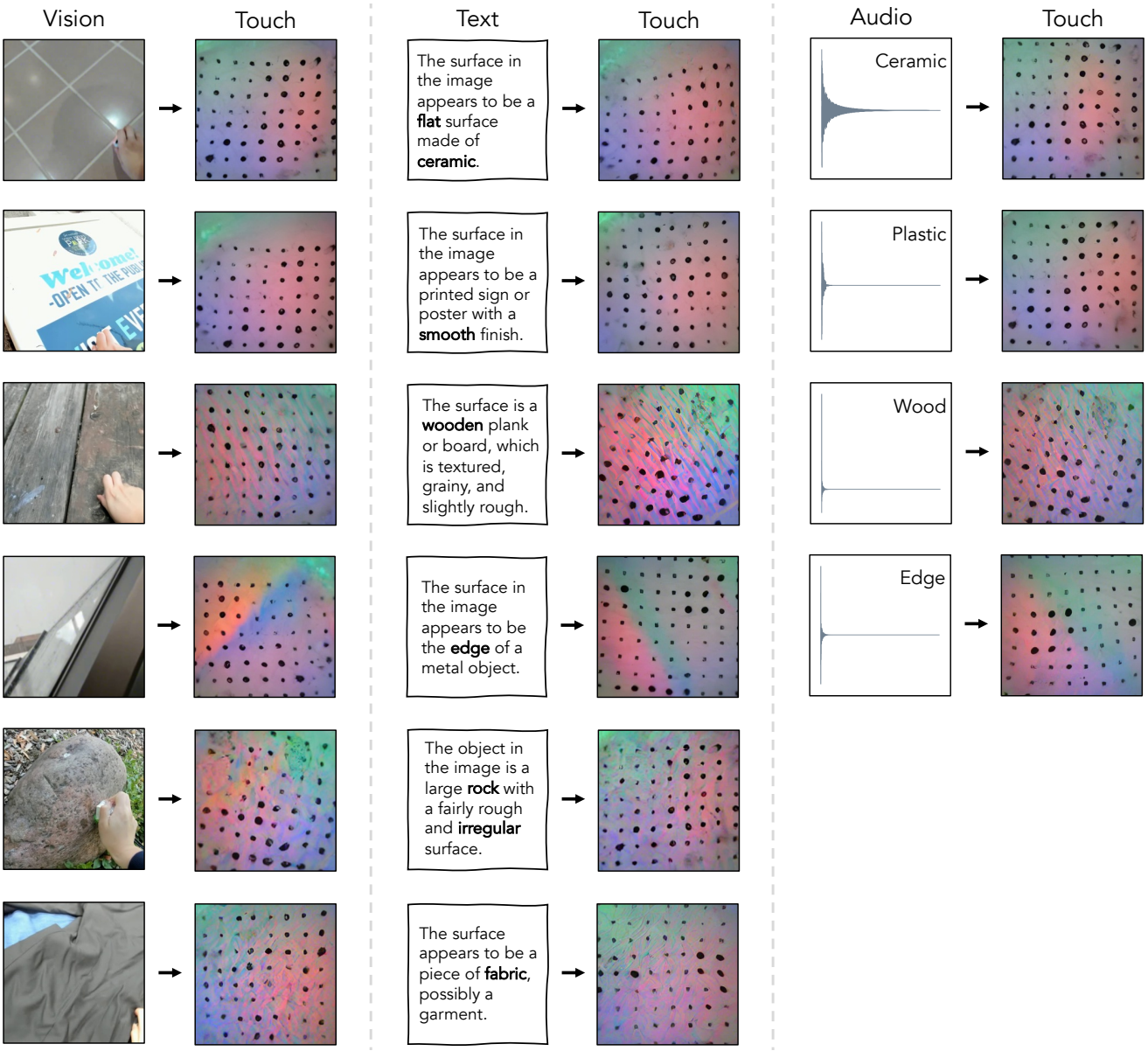

Figure 5. **More examples for X-to-touch generation.** We show more examples of x-to-touch generations on the Touch and Go [12] dataset. We manually select audios from ObjectFolder 2.0 [5] matching the vision input. Since the overlapping material categories between [5] and [12] are limited and [5] only contains rigid objects, impact sound for materials like stone and cloth can not be found.

References

[1] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023. 2

[2] Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H Adelson, and Sergey Levine. The feeling of success: Does touch sensing help predict grasp outcomes? *Conference on Robot Learning (CoRL)*, 2017. 1

[3] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 2

[4] Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. In *CoRL*, 2021. 1

[5] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *CVPR*, 2022. 1, 7

[6] Ruohan Gao, Yiming Dou, Hao Li, Tanmay Agarwal, Jeannette Bohg, Yunzhu Li, Li Fei-Fei, and Jiajun Wu. The objectfolder benchmark: Multisensory learning with neural and real objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17276–17286, 2023. 1

[7] Justin Kerr, Huang Huang, Albert Wilcox, Ryan Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. Self-supervised visuo-tactile pretraining to locate and follow garment features. In *Robotics: Science and Systems*, 2023. 1

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3

[9] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[10] Sudharshan Suresh, Zilin Si, Stuart Anderson, Michael Kaess, and Mustafa Mukadam. MidasTouch: Monte-Carlo inference over distributions across sliding touch. In *Proc. Conf. on Robot Learning, CoRL*, Auckland, NZ, 2022. 1

[11] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[12] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *Neural Information Processing Systems (NeurIPS) - Datasets and Benchmarks Track*, 2022. 1, 3, 7

[13] Fengyu Yang, Jiacheng Zhang, and Andrew Owens. Generating visual scenes from touch. *International Conference on Computer Vision (ICCV)*, 2023. 2

[14] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 2