Supplementary Material for "AID-AppEAL: Automatic Image Dataset and Algorithm for Content Appeal Enhancement and Assessment Labeling"

Sherry X. Chen[1⋆], Yaron Vaxman[2], Elad Ben Baruch[2], David Asulin[2], Aviad Moreshet[2], Misha Sra[1], and Pradeep Sen[1]

[1] University of California, Santa Barbara
[2] Cloudinary

Here we elaborate on our datasets in Appendix A, including the creation process and sample images (Appendix A.1), as well as method generalizability across image domains (Appendix A.2).

In Appendix B, we compare the content appeal labels in our dataset with aesthetic scores from IAA baselines (Appendix B.1), demonstrate the generalizability of the models on amateur-taken images (Appendix B.2), and discuss the effect of technical distortions on content appeal (Appendix B.3).

Appendix C outlines the configuration of our content appeal enhancer, followed by more enhancement results and ablation studies in Appendix C.2, while Appendix C.3 provides further setup details of enhancement baselines we compared in the paper.

Finally, Appendix D furnishes more information regarding our user study, including the questionnaire and analysis of data collected from the participants.

A Dataset Details

A.1 Dataset creation details and samples

To show that AID-AppEAL can be generalized across different image domains, we create two datasets, one with food images and the other with room interior images. Here we present them in detail.

Food: Search queries were generated from the following sets of words:

- $\mathbb{N}_F$ = {"burger," "cake," "chicken," "cookie," "food," "rice," "pizza," "pasta," "salad," "steak," "yogurt"}
- $\mathbb{A}_F^+$ = {"delicious"}
- $\mathbb{A}_F^-$ = {"burnt," "moldy," "rotten"}

We generated search queries and retrieved 189,477 image thumbnails from image hosting sites stock.adobe.com and shutterstock.com. We used our filtering method with $\gamma = 0.4$, which gave us 80,067 images, all of which were upscaled and zero-padded to 512×512 resolution.

⋆ Corresponding author email: xchen774@ucsb.edu

Fig. 11: Dataset samples. We show 4 sample images from each of the food and room interior dataset, where the label next to each row indicates the content appeal score and image aesthetic score level of images in the corresponding row. Images with scores above the 75^{th} percentile in each dataset or IAA baseline predictions are considered to have high (H) scores. Images with scores below the 25^{th} percentile in each dataset or IAA baseline predictions are considered to have low (L) scores.

We selected 50 "delicious food" images as $\mathbb{T}_F^+$, 50 "burnt food" images as $\mathbb{T}_{F_1}^-$, as well as a total of 50 "moldy food" and "rotten food" images as $\mathbb{T}_{F_2}^-$ for textual inversion. We generated two $\mathbb{T}_F^-$'s because burnt food and moldy/rotten food have distinctly different features (blackened food vs. hairy mold) that rarely appear in the same image in real life. Mixing them will generate images with both characteristics together, which is not very realistic. All selected images appear at the top of search results by search engines using the corresponding queries to ensure maximum relevance between image content and search queries. We train z_F^+, $z_{F_1}^-$, and $z_{F_2}^-$ with $\mathbb{T}_F^+$, $\mathbb{T}_{F_1}^-$, and $\mathbb{T}_{F_2}^-$ respectively using Stable Diffusion with batch size 1 and learning rate $lr = 5e^{-3}$.

Following that, we select a different set of 1,000 images with balanced content appeal levels and food types as the starting point of our synthetic dataset. Specifically, we choose 50 images retrieved from each $q = a + n$ where $+$ means appending to $a \in \mathbb{A}_F^+, n \in \mathbb{N}_F - \{\text{"food"}\}$, which gives us 500 images with appealing content and balanced food types as I_F^+. Similarly, we choose 50 images retrieved using $a \in \mathbb{A}_F^-$, which gives us a total of 500 images with unappealing content as $\mathbb{I}_F^-$. All selected images appear at the top of search results by search engines using the corresponding queries to ensure maximum relevance between image content and search queries. We use $n \in \mathbb{N}_F - \{\text{"food"}\}$ to help constrain object types and keep them balanced. For each $i \in \mathbb{I}_F^+ \cup \mathbb{I}_F^-$, we first augment it to generate three versions of $I' = SD(I, \text{" "}, 1 - M_F(I), \text{seed}())$, where $1 - M_F(I)$ is the inverse of domain-relevancy map for the food domain. For each I', we

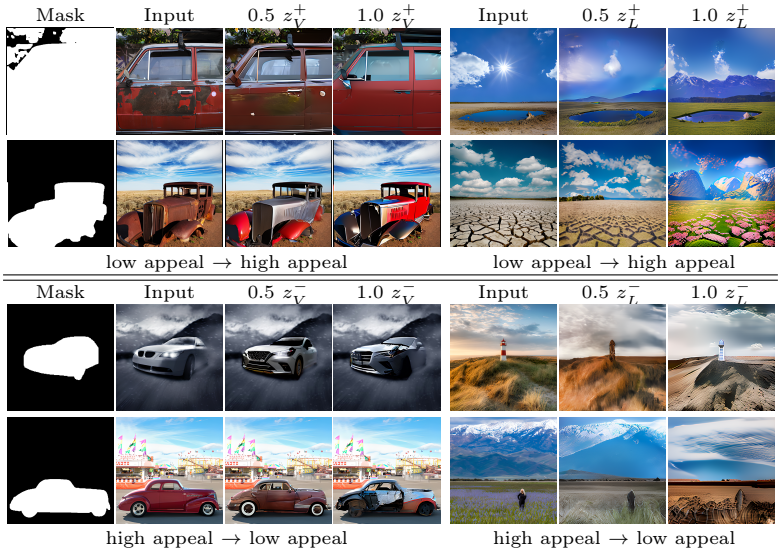


Fig. 12: We trained embeddings z_V^+/z_L^+ and z_V^-/z_L^- for vehicles (left) and landscapes (right) to adjust image appeal with different weights in the domain-relevant area ("Mask" for vehicles; for landscapes, we consider all pixels to be relevant) and created synthetic datasets samples. Although these results are not equivalent to the final output of our image enhancer (which operates with respect to the appeal heatmaps from our predictor and generates more consistent results), we can observe successful appeal changes between images necessary for training our models.

generate 6 final images $s = SD(I', BLIP(I) + f(\alpha), M_F(I), \text{seed}())$, where

$$\begin{aligned} \alpha &= max(min(k/2 + \delta, 1), 0) \\ k &\in 0, 1, 2 \\ \delta &\in uniform(-0.2, 0.2). \end{aligned} \tag{1}$$

Note that k is used to ensure that 6 images generated from each i span the entire content appeal spectrum. We use δ to add randomization and more variety in $\hat{A}(\cdot, \cdot)$ to avoid over-fitting when training our relative content appeal comparator. In the end, we generated 18,000 images as our synthetic dataset $\mathbb{S}_F$ and 78,917 remaining images for the final dataset $\mathbb{I}_F$.

Room: Search queries were generated from the following sets of words:

- $\mathbb{N}_R$ = {"bathroom," "bedroom," "kitchen," "living room," "room"}
- $\mathbb{A}_R^+$ = {"interior"}
- $\mathbb{A}_R^-$ = {"abandoned," "dirty"}

Note that we didn't include "clean" in $\mathbb{A}_R^+$ because the word can be interpreted as a verb, so images focusing on people cleaning rooms will be returned, which is outside the room interior domain. We collect 261,907 image thumbnails and

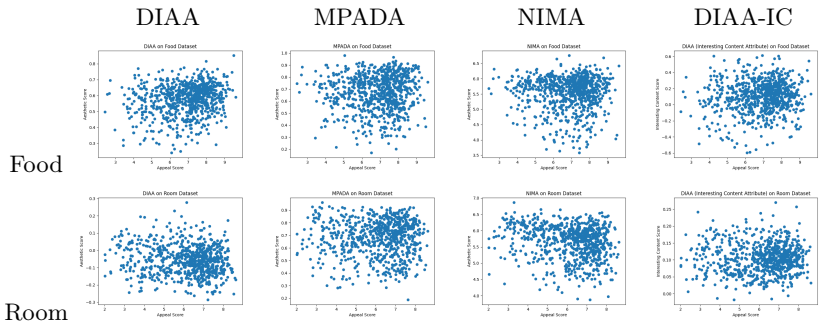


Fig. 13: Correlation between content appeal and image aesthetics. We visualize the relationship between predictions from our estimator and from three IAA models on subsets of our two datasets. We can see there is little correlation between content appeal and image aesthetics, suggesting they are indeed different image metrics. There is also little correlation between our content appeal predictions and DIAA "interesting content" (DIAA-IC) predictions, meaning that the latter cannot be readily substituted by the former.

obtain 76,387 images of size 512×512 after filtering and preprocessing. Likewise, we select 100 images to generate embeddings using textual inversion, and 1,000 images with balanced content appeal levels and room types to create the synthetic dataset. For each image, we use it to generate five different augmentations. For each augmentation, we change its content appeal level and generate three different images. In the end, we generate 15,000 images for our synthetic dataset $\mathbb{S}_R$, leaving us with 75,287 images for the final dataset $\mathbb{I}_R$.

We present image examples from each dataset with various levels of content appeal and image aesthetics in Fig. 11. Specifically, we uniformly stride one out of each 100 images in each dataset we created by image indices and estimate their image aesthetics scores using three popular open-sourced IAA baselines: DIAA, MPADA, and NIMA. We denote images with appeal scores in the 25^{th} and 75^{th} percentile in their respective datasets to have low and high content appeal respectively. Images with aesthetics scores in the 25^{th} and 75^{th} percentiles across all three IAA baselines have low and high aesthetics respectively. We can see that the content appeal and image aesthetics of an image may be very different.

A.2 Dataset creation across image domains

AID-AppEAL can be easily adapted to different domains, of which we demonstrate two new ones here: *vehicles* and *landscapes*, where we illustrate the process of creating synthetic datasets. This involves gathering 50 appealing and 50 unappealing images for each domain, which are used to train appealing and unappealing textual inversion embeddings, z_V^+/z_L^+ and z_V^-/z_L^-, following the same methodology used for food and rooms. This allowed us to manipulate the relative appeal of images to generate synthetic datasets (Figure 12). As can be seen, our

Fig. 14: Generalizability of content appeal estimator on amateur-taken images. Although being trained on professionally-taken images, the estimator can be generalized to amateur-taken images during run time and accurately distinguish appealing (predicted scores in blue and bold) and unappealing (predicted scores in red and boxes) images.

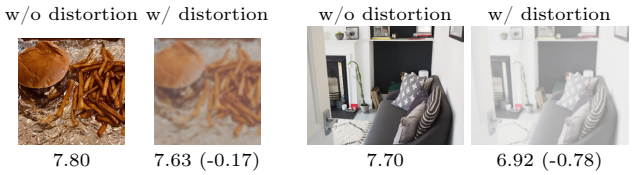


Fig. 15: When two images have the same content, technical distortions have a negative impact on content appeal scores (predicted by our model and shown below each image) as aesthetics and content appeal are not orthogonal axes.

method does a reasonable job at increasing/decreasing image appeal in these very different domains.

B Image Content Appeal Estimator Details

B.1 IAA baseline comparison

To further show the difference between content appeal and image aesthetics, we visualize the correlation between them (Fig. 13) on above strided images, where we observe little correlation between content appeal and image aesthetics (for coefficient values, please refer to the paper). Furthermore, we visualize the relationship between content appeal and DIAA "interesting content" attribute (Fig. 13 Row.4), where little correlation is presented as well. This means that DIAA 'interesting content" attribute cannot substitute ICAA either.

B.2 Performance on amateur-taken images

Although our estimator is trained on professionally-taken images, it can be generalized to amateur-taken images during inference time and accurately distinguish content appealing (predicted scores in blue and bold) and content-unappealing (predicted scores in red and boxes) images (Fig. 14).

B.3 Effect of technical distortions

When two images have the same content, their content appeal should be affected by technical distortions, which is correctly reflected in our models (Fig. 15). However, these distortions should not overshadow the inherent appeal of the image content. As illustrated in Fig. 1 and Fig. 14, images with unappealing content yet high aesthetic quality still receive low content appeal scores.

C Content Appeal Enhancer Details

C.1 Implementation details

We use Stable Diffusion v2.1 inpainting with depth-guided ControlNet for image content appeal enhancement. Specifically, here are some parameter values we use:

- prompt: "$<z_D^+>$ <object_type>"
- negative prompt: "out of frame, lowres, text, error, cropped, worst quality, low quality, jpeg artifacts, ugly, duplicate, morbid, mutilated, out of frame, extra fingers, mutated hands, poorly drawn hands, poorly drawn face, mutation, deformed, blurry, dehydrated, bad anatomy, bad proportions, extra limbs, cloned face, disfigured, gross proportions, malformed limbs, missing arms, missing legs, extra arms, extra legs, fused fingers, too many fingers, long neck, username, watermark, signature,"
- "Sampling method": "DPM++ 2M Karras"
- "CGF scale" = 7
- "denoising strength" = 0.6
- ControlNet "Preprocessor": "depth_midas",

where the prompt is constructed by concatenating the appealing embedding with the type of the object in the input image (e.g. burger, kitchen), and ControlNet preprocessor use MiDaS [Ranftl et al. 2020] to estimate a depth map from the input image.

Note that not all phrases in the negative prompt are directly related to the image domain the input image is from. Instead, we use this generic negative prompt for all image domains.

C.2 More results and ablation studies

We present a comparative display of images before and after enhancement, accompanied by their content appeal scores as determined by our absolute appeal estimator (Fig. 16). We also show the input image appeal heatmap and the estimated depth that guided the enhancement process. The visual and quantitative evidence from the increase in appeal scores clearly demonstrates that our methodology not only elevates the content appeal of images but also meticulously preserves the original color palette and structural integrity of the content.

We demonstrate the effect of different denoising strength, appeal heatmap M_D^H, and the depth map on the enhancement result in Fig. 17, where lower

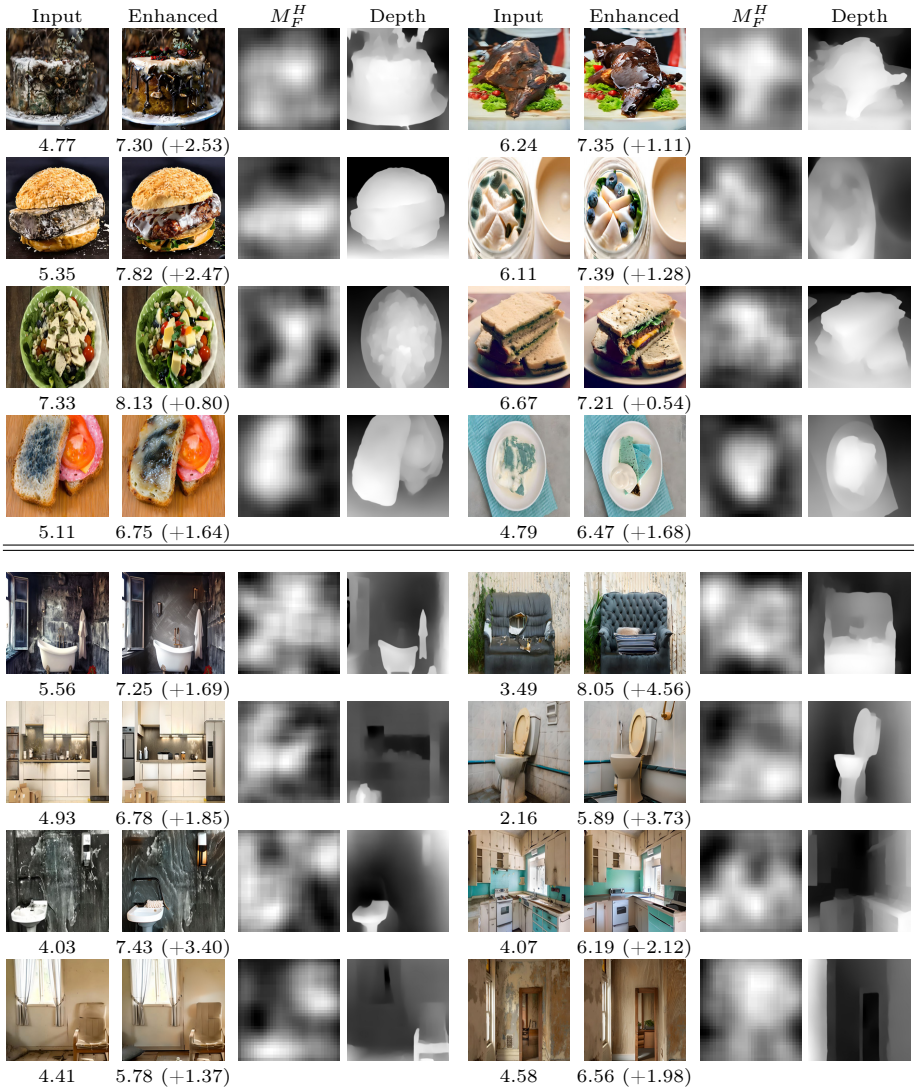


Fig. 16: Image content appeal enhancement. Corresponding to Fig. 9, we show images before/after enhancement (Col. 1/5 vs. Col. 2/6) with estimated appeal scores below each image. We use both the appeal heatmap M_F^H (Col. 3/7) and the depth map (Col. 4/8) to guide the enhancement process.

denoising strength values (e.g., 0.3, 0.45) result in marginal improvements in content appeal, indicating that such settings are insufficient for effective enhancement. Excessively high denoising strength values (e.g., 0.75, 0.9) can cause noticeable color and style discontinuities between enhanced and non-enhanced areas, as shown by the appeal heatmap M_D^H. We chose a denoising strength of 0.6 to balance enhancement impact with visual coherence. Omitting M_D^H can increase overall content appeal but may undesirably alter appealing objects. Using M_D^H helps prevent unwanted changes, and incorporating a depth map ensures the preservation of these attributes during enhancement.

C.3 Baselines details

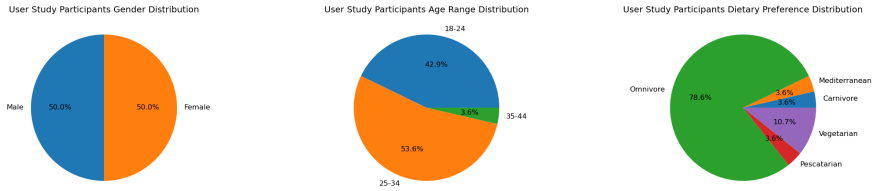
We use the following text-guided localized image editing models as baselines for image enhancement comparisons:

- InstructPix2Pix (IP2P): It takes text instructions as inputs to manipulate images. For food image, we use "turn it into a delicious [*item*]," where [*item*] is the name of the food in the image; for room images, we use "turn it into a clean [*item*]," where [*item*] is the name of the room in the image. In both cases, [*item*] is parsed from the image text description generated by BLIP.
- Null-text Inversion (N-TI): This method takes an image and its text description as inputs, inverts the image based on the description, and allows edits by inserting new words or adjusting attention weights of existing words. We use BLIP to generate text descriptions of images. For editing, we decrease the attention weight of negative adjectives to -100 and insert positive adjectives like "delicious," "tasty," "clean," or "tidy," increasing their attention weight to 100. These values were set experimentally for optimal appeal improvement with minimal artifacts.
- pix2pix-zero (P2P-0): This method enables image manipulation using a specified edit direction. We generated two sets of 1,000 captions each for unappealing (burnt, moldy, rotten food) and appealing food images. The edit direction is the mean difference between the CLIP text embeddings of these sets. Similarly, for rooms, we created two sets of 1,000 captions describing unappealing (abandoned, dirty) and appealing (clean) rooms, following the same steps as for food images to define the edit direction.
- Text2LIVE (T2L): This method takes two prompts (p_O, p_T) as inputs, where p_O describes the input image and p_T describes the target(desired) output. We take the search query that is used to retrieve the corresponding input image as p_O. For the Food dataset, we use p_T = "*delicious*[*item*]; for the Room dataset, we use p_T = "clean [item]", where [*item*] is obtained in the same manner as in IP2P.

D User Study Questionnaire and Statistics

Here is the pre-survey questionnaire we ask participants to fill out:

Fig. 17: Effect of different denoising strength (ds) values, appeal heatmap, and depth on content appeal enhancement. By enhancing the original image (the leftmost image in Cols.1 and 4 respectively) with different configurations, this analysis reveals that lower denoising strength values (e.g., 0.3, 0.45) result in marginal improvements in content appeal, indicating that such settings are insufficient for effective enhancement. Conversely, excessively high ds values (e.g., 0.75, 0.9) risk creating noticeable discontinuities in color and style between enhanced and non-enhanced areas, as delineated by the appeal heatmap M_D^H. Consequently, we opted for a denoising strength of 0.6 (highlighted in bold), balancing enhancement impact with visual coherence. Although omitting M_D^H can ostensibly further augment overall content appeal, it also introduces undesired modifications, such as altering the appearance of the burger buns or cabinet drawers next to the fridge. Employing M_D^H serves to mitigate unwarranted changes in color and structure, and the integration of a depth map further ensures the preservation of these attributes throughout the enhancement process.



(a) **(b)** **(c)**

Fig. 18: User Study Questionnaire Answers Statistics. Out of all the participants, there is an even split between males and females (Fig. 18a). The ages of most participants (27 out of 28; 96.4%) are below 35, with 12 (42.9%) of them aged between 18-24 and 15 (53.6%) between 25-34 (Fig. 18b). From Fig. 18c, we can see that the majority of participants are omnivores (22 out of 28; 78.6%); the second most common dietary preference among participants is Vegetarian (3 out of 28; 10.7%).

- Gender: M/F/Other/Prefer not to say
- Age range: 18-24, 25-34, 35-44, 45-54, 54+
- Dietary preference: Vegan, Vegetarian, Omnivore, Carnivore, Mediterranean, Keto, Paleo, Other (please specify):

Out of all 28 participants, there is an even split between males and females (Fig. 18a). The ages of most participants (27 out of 28; 96.4%) are below 35, with 12 (42.9%) of them aged between 18-24 and 15 (53.6%) between 25-34 (Fig. 18a). The majority of participants are omnivores (22 out of 28; 78.6%); the second most common dietary preference among participants is Vegetarian (3 out of 28; 10.7%).

To see how participants' personal dietary preference may affect their responses, we visualize responses by dietary preference (Fig. 19), where we observe no major distribution change of user preference in terms of image appeal across participants with different dietary preference. This suggests that the question we ask in the user study, "Which item in the image do you think the majority of the people would prefer", helps leverage individual preference.

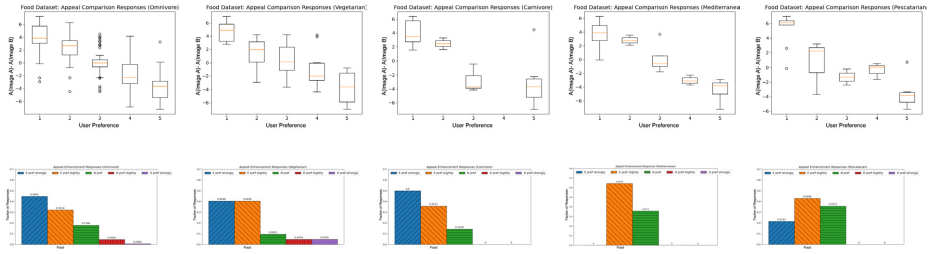


Fig. 19: Image Appeal Response Statistics by Dietary Preference. Top row is the distribution of the appeal score difference for each of the five response options in the user study. Bottom row is the percentage of image enhancement preference responses for each category, where E represents the enhanced image, O is the original image, N is neither, and "pref" stands for "is preferred." From left to right are responses from participants whose dietary preference is Omnivore, Vegetarian, Carnivore, Mediterranean, and Pescatarian. We observe no major distribution change in responses across participants with different dietary preferences.