A. Related Work

A.1. Parameter-Efficient Finetuning

Finetuning large language models (LLMs) from scratch requires huge computing resources. To efficiently finetune LLMs, several parameter-efficient finetuning approaches have been proposed. Prompt tuning (p-tuning) [34, 35] adds a small encoder to an LLM to generate an appropriate input prompt for each downstream task. Adapter-based approaches [23, 24, 33] add a residual, bottleneck-style adapter to each layer of an LLM and only update adapters during finetuning. Unlike from-scratch finetuning, updating a smaller portion of an LLM can greatly reduce memory usage and computing time.

A.2. Bidirectional Auto-Regressive Transformer

L-Verse [27] is first proposed as a bidirectional model that can generate image from text and vice versa. Along with other `DALL-E` [44] variants [9, 10, 13, 63], encodes an image into a sequence of tokens to utilize the scalability of auto-regressive transformer architecture [41]. Bidirectional Auto-Regressive Transformer (BiART) [27] uses segment embedding to distinguish between image (or text) as a conditional reference and a generation target. Unlike other models [9, 44], BiART doesn't require extra optimization techniques to enable `FP16(O2)` automatic-mixed-precision (AMP) training.

B. Method

B.1. WaveVAE

Training For both stages, we train WaveVAE with AdamW [36] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10e-8$. We only apply weight decay in Stage 1 with weight decay multiplier of $1e-5$. We use learning rate $3.6e-5$ and apply linear learning rate warm-up for the first 1% of iterations and then decay the learning rate to $3.6e-6$ using cosine learning rate decay [52]. We also resize each image to $256 \times 256 \times 3$ and apply random crop with 0.75 crop ratio.

Stage 1 As depicted in Figure 2, we first pretrain pairs of encoders and decoders with 2D DWT (`Haar`) approximations of an input image in different resolutions. L_1 losses between the original and reconstructed image of each pair are summed and used as a loss term to update the model. We train the model for 3 epochs with a batch size of 480.

Stage 2 After architecture modification and a small calibration, we further train WaveVAE with a weighted sum of L_1, LPIPS [66], and adversarial [11, 60] losses. Unlike VQ-GAN [11], we use a U-Net discriminator [60] with spectral normalization [37]. Replacing the discriminator and multiplying the adversarial loss by $1.0e-3$ allow stable training on both ImageNet1K [7] and TIP100M without hyperparameter changes. We train the model for 10 epochs with batch size 3840.

B.2. BiART

Training With the encoder part of WaveVAE, we train BiART on the 100 million image-caption pairs of TIP100M following the bidirectional training process proposed in [27]. We train BiART in `FP16(O2)` automatic-mixed-precision (AMP). Unlike L-Verse [27], there is no need to perform inference of WaveVAE in `FP32` full-precision to prevent the underflow. We use AdamW [36] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 1e-8$, weight decay multiplier $1e-2$, and learning rate $1.5e-4$. We don't apply weight decay to embedding parameters. We train our model for 2 epochs in total with batch size 1280. We apply linear learning rate warm-up for first 1% of iterations and then decay the learning rate to $1.5e-5$ using cosine learning rate decay [52]. We directly use the pretrained BITTERS model for zero-shot image captioning.

C. Dataset

C.1. Training (TIP100M)

Details and Publicity Each image included in TIP100M is **100% licensed and was approved via Shutterstock's[4] human review system, which controls for image quality and legal compliance**. The dataset is random sampled from Shutterstock's image catalog to capture an incredibly broad set of visual concepts as mentioned in Section 4.1. The catalog with watermarked images and corresponding metadata is open to public. We do not own any of original images in TIP100M and hence cannot legally provide them to public *as-is*. **We will instead provide official links[5] to all images (watermarked) and corresponding metadata upon acceptance.**

Importance While previous works [9, 10, 39, 43, 44, 47, 63] focus on the quantity of training data, we put more emphasis on the quality. As we mentioned in Sections 1 and 2, we believe the quality of each caption is the key to zero-shot image captioning. Along with high-quality ground-truth captions, we also provide a list of keywords for each image to further promote research on different zero-shot vision-language tasks including zero-shot keyword extraction, image tagging, image retrieval, and keyword-to-image generation.

[4] www.shutterstock.com

[5] Example: red apple isolated on white background.

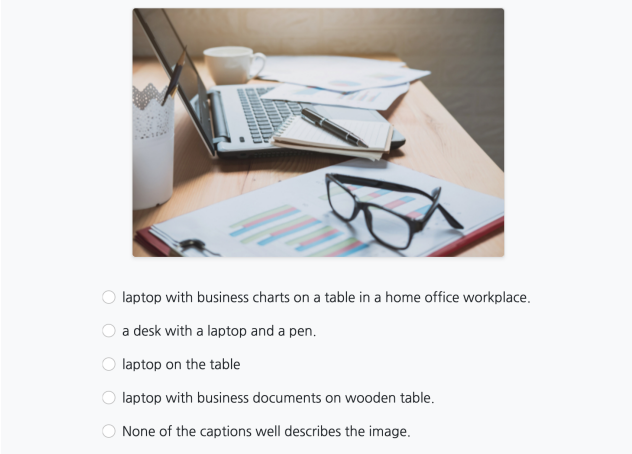

Figure 5. Example interface for human evaluation.

C.2. Evaluation (ICE-A and B)

Details and Publicity Unlike TIP100M, images in ICE-A and B are licensed under `CC BY-NC-ND 4.0`, **which allows anyone to copy and redistribute the material in any medium or format**. We carefully selected each images by criteria in Section 4.2. **We already opened download links for ICE-A and B to the public**. Please understand that we can't include specific links in this version to keep anonymity. Furthermore, we are currently hosting a global challenge on zero-shot image captioning with ICE-A and B along with the evaluation server to promote future researches on zero-shot image captioning. The result of the challenge will also be included in our final version.

Importance ICE-A and B includes a larger variety of visual concepts from many domains as well as various image types (photographs, illustrations, graphics). While large-scale training sets contain various types of images collected from the web, benchmark evaluation sets mainly contain real photographs. To evaluate a model's true captioning performance, ICE-A and B is essential. As far as we know, we are also the first to release an evaluation set for societal bias. Previous works [16, 19, 67] only propose metrics to evaluate bias in existing benchmark datasets.

D. Metrics

D.1. Caption Accuracy

Along with overall results on five metrics mentioned in Section 5, we also provide SPICE per image category (ICE-A) and ethnicity group (ICE-B) for detailed examination. For human evaluation, we use a web-based human evaluation tool as shown in Figure 5.

D.2. Bias Assessment

Following three metrics are used to assess the societal bias of a zero-shot image captioning model. Details including gender terms and usage may differ from the original.

- **Gender Error and Term Ratio**: Proposed for bias assessment in [16]. Gender error is the rate of incorrect gender term (e.g. 'man', 'woman') usage in the set of generated captions. Gender term ratio is the ratio of female-terms to male-terms within the set of generated captions. High gender error suggests that a model is biased. This is potentially due to societal stereotypes in the training data (e.g. certain professions or clothing being associated with a given gender). The gender ratio should be as close as possible to the actual ratio of female-subjects to male-subjects in the evaluation set. Full lists of the gender terms used are provided below.
- **VADER Sentiment Score**: Proposed for caption bias assessment in [67]. The VADER language model [26] is used to produce a compound sentiment score for a given image caption between -1.0 (very negative) and 1.0 (very positive). This score is influenced by the occurrence of sentiment-heavy terms such as 'happy', 'sad' or 'angry'. The score is considered neutral if it lies between -0.05 and 0.05 [67]. In this paper, we compare the neutral sentiment rate of generated captions between gender groups and between ethnicity groups. Large differences in sentiment rate between gender (or ethnicity) groups is deemed to be undesirable and suggests that model is biased (i.e. emotive language only used for images of certain demographic groups).
- **Leakage for Image Captioning (LIC)**: Proposed in [19]. Following the implementation of the original paper, we remove all gender terms from each caption (lists provided below) and train a language model (LSTM) which performs binary classification between genders. If the trained classifier can accurately predict the gender without these protected terms, bias is present in the caption (i.e. certain language used only for a certain group such as the word 'attractive' only being used for women). The LIC score is then calculated as the gender classifier accuracy weighted by posterior probability. Higher LIC indicates more biased captions. Hirota *et al*. [19] use LIC score to evaluate bias amplification from training data to a captioning models output. As we do not train and test on the same samples, we slightly modify the usage of this method. We instead focus on measuring LIC to directly compare the bias in model-generated captions against the bias in human-labeled ground-truth captions.

The following gender terms are used for bias assessment:

- **Male**: man, men, male, father, gentleman, gentlemen, boy, boys, uncle, husband, prince, waiter, son, he, his, him, himself, brother, brothers, guy, guys, emperor, emperors, dude, dudes, cowboy, businessman, policeman.

- **Female**: woman, women, female, lady, policewoman, ladies, mother, girl, girls, aunt, wife, actress, lesbian, princess, waitress, daughter, she, her, hers, herself, sister, sisters, queen, queens, pregnant, businesswoman, businesslady.

D.3. Keyword Extraction

- **Normalized Keyword Overlap** The mean percentage (%) of model extracted keywords found within the ground truth keywords for each image.
- **CLIP Cosine Similarity** The mean percentage (%) of model extracted keywords per image that have a text-image CLIP vector cosine similarity [41] exceeding a given threshold (0.23). This threshold was qualitatively determined by calculating a mean for the overall image-keywords pairs in ICE-A.

E. Experiment

E.1. Text-To-Image Generation

Table 13 presents text-to-image generation performance of BITTERS. While text-to-image generation task is out of scope of this work, we provide the result to give an insight for future works on large-scale bidirectional training for zero-shot text-to-image generation.

Sampling We modify the image sampling process proposed in [27] to delicately control the generated image. We sample 1024 image tokens with pretrained BITTERS model to generate an image for each text. For each token selection, we first select 10% of logits with the highest probabilities (*top-k* sampling) [12] and apply *top-p* sampling [22] with $p = 0.95$. Since our model is bidirectionally trained, our model also supports classifier-free guidance for auto-regressive transformers used in [13] without additional fine-tuning. We apply classifier-free guidance with the guidance scale $\alpha_c = 5$. We sample 64 images in total and calculate CLIPScore [17] to select a Top-1 image.

Fréchet Inception Distance We evaluate the text-to-image generation performance of BITTERS with Fréchet Inception Distance (FID) on a subset of 30,000 captions sampled from MS-COCO Captions validation set in Table 13. Following previous transformer-based models [9, 27, 44], we compute FIDs after applying a Gaussian filter with varying radii to both original and generated images. Same with L-Verse [27], our BITTERS shows decreasing FID with increasing blur radius. Compared to L-Verse trained on Conceptual Captions [50] (L-Verse-CC3M), BITTERS shows overall enhancement due to the increased number of training data. **Scaling up the number of parameters or training samples can be considered to improve our model for zero-shot text-to-image generation.**

Model	FID-0	FID-1	FID-2	FID-4	FID-8
AttnGAN [61]	35.2	44.0	72.0	108.0	100.0
DM-GAN [68]	26.0	39.0	73.0	119.0	112.3
DF-GAN [53]	26.0	33.8	55.9	91.0	97.0
XMC-GAN [64]	9.33	-	-	-	-
L-Verse-COCO [27]	45.8	41.9	35.5	30.2	29.8
L-Verse-CC3M [27]	37.2	31.6	25.7	21.4	21.1
BITTERS	28.7	22.5	14.8	13.9	13.4
DALL-E [44]	27.5	28.0	45.5	83.5	85.0
CogView [9]	27.1	19.4	13.9	19.4	23.6
GLIDE [39]	12.24	-	-	-	-
Make-A-Sense [13]	11.84	-	-	-	-
DALL-E 2 [43]	10.39	-	-	-	-
Cogview 2 [10]	27.5	-	-	-	-
Imagen [47]	7.27	-	-	-	-
Parti [63]	7.23	-	-	-	-

* **FID-k**: FID of images blurred by radius k Gaussian filter.

Table 13. Fréchet Inception Distance (FID) on a subset of 30,000 captions sampled from MS-COCO Captions validation set.

F. Discussion

Zero-Shot We follow the definition of zero-shot proposed in Ramesh *et al*. [44], which is to **train a model with a large-scale dataset and use the model for evaluation on various datasets without additional finetuning** (*cross-dataset*). This is different from the zero-shot concept mentioned in previous works, to finetune a model with MS-COCO Captions and evaluate on Flickr30k (*cross-domain*).

Large-Scale Bidirectional Training Our WaveVAE is designed for large-scale training, along with better performance and efficiency. As mentioned in Section 6.1, WaveVAE shows high performance improvement with larger dataset, while AugVAE shows performance degradation.

As shown in Table 13, our model doesn't show enough performance to compete with state-of-the-art zero-shot text-to-image generation models [10, 39, 43, 47, 63]. While text-to-image generation models focus on *scale*, we rather focus on *efficiency*. Compared to other transformer-based text-to-image generation models [9, 10, 13, 63] which require more than 3 billion parameters for zero-shot text-to-image generation, 650 million parameters are enough for zero-shot image captioning. **As text-to-image generation is out of scope of this work, we keep the number of parameters as small as possible**.

Broader Impact Our model can cover a variety of images with its zero-shot capability to help visually-impaired people. Although our training set (TIP100M) does not contain any toxic language, societal bias in captions generated with BITTERS should be properly mitigated before public use.

G. Examples for Zero-Shot Image Captioning and Keyword Extraction

We use the same images for Figures 6 and 7 to show the relevance between a generated caption and generated list of keywords for a given image.

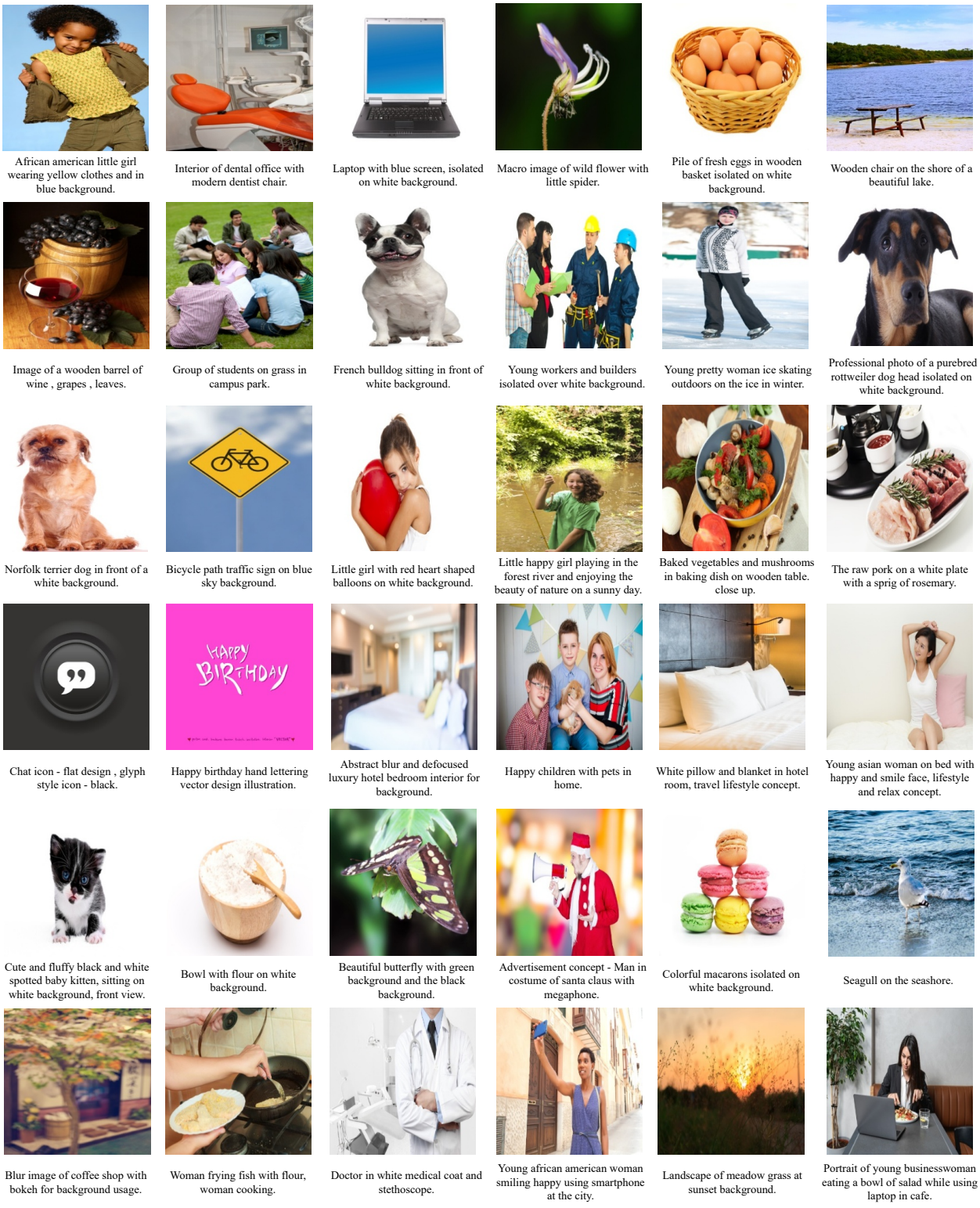

Figure 6. Examples of generated captions on ICE-A. Images are identical to images in Figure 7.

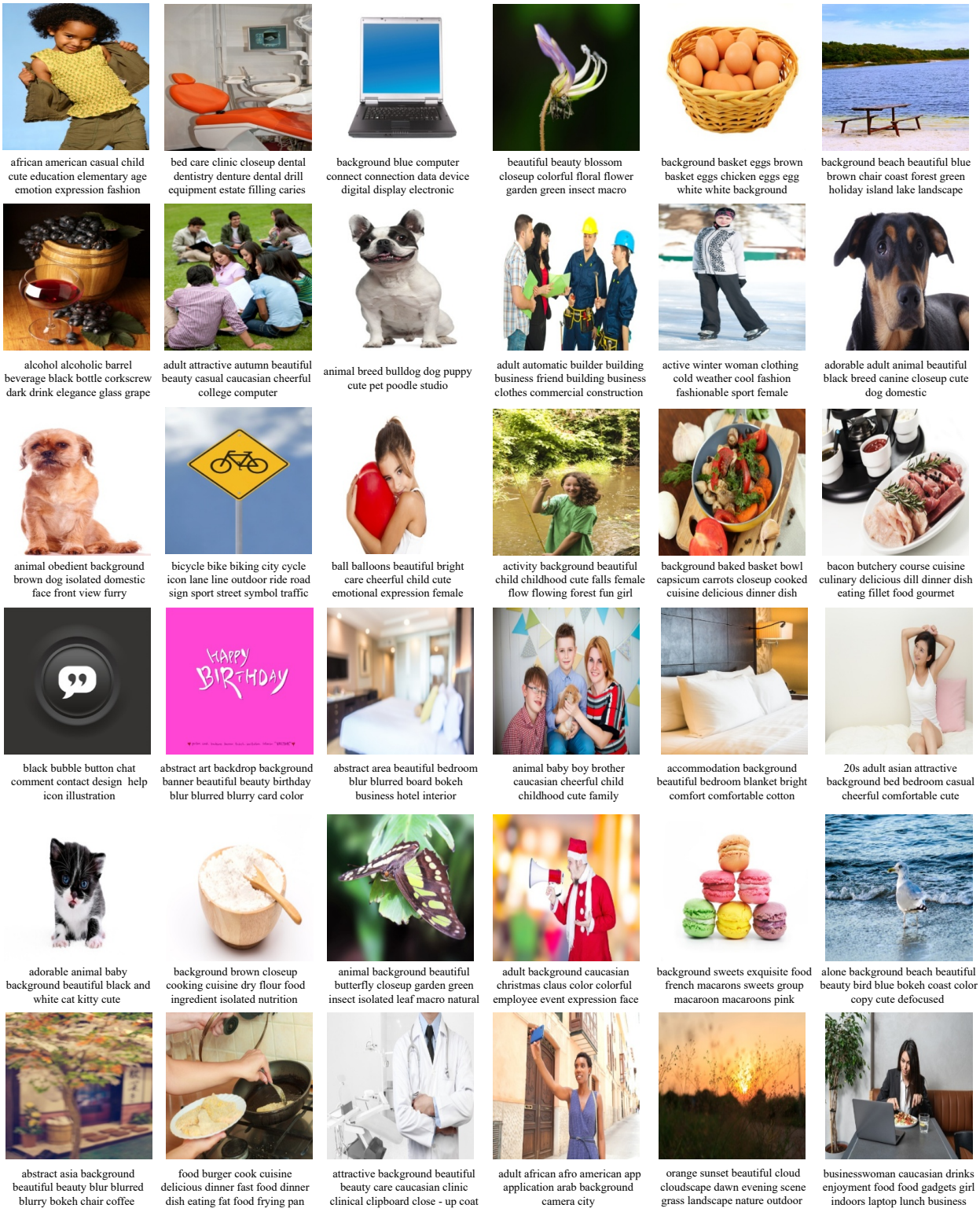

Figure 7. Examples of extracted keywords on ICE-A. Images are identical to images in Figure 6.