# Supplemental Materials for
# Toward a Visual Concept Vocabulary for GAN Latent Space

## S.1 Annotation collection and processing

**Collection** As described in Section 3.2, we collect annotations using Amazon Mechanical Turk (AMT) for layer-selective directions visualized in four classes (cottage, kitchen, lake, medina). Instructions and an example task are found in Supplementary Figure 1. We require workers to be located in the U.S., with $> 97\%$ HIT acceptance rate and $> 100$ HITs accepted. Workers were paid $\$0.06$ per annotation.

**Normalization and post-processing** We normalize direction annotations before applying the method described in Section 3.3 to decompose them into a vocabulary of primitive visual concepts. We use *pyspellchecker* to automate simple corrections, keeping the original string if there is no word with an edit distance less than 3. Lemmatizing is done with NLTK WordNetLemmatizer, discarding common terms used to describe the setting (e.g., *image*, *scene*) or image class (e.g., *house*, *lake*). We then run a basic sentiment analysis script to detect modifier words indicating whether a concept is being added (e.g., *appears, added, more*) or taken away (e.g., *disappears, removed, less*, *goes from*) in an image transformation. This simple approach worked sufficiently well to disambiguate different uses and positive vs. negative sentiment of a concept.

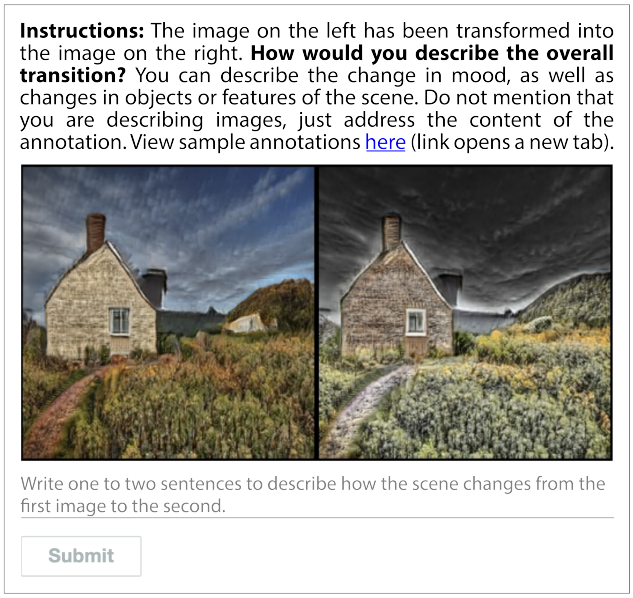


Figure S1: Example annotation HIT. Annotators are shown $G(\mathbf{z}_i; \mathbf{y})$ (left) and $G(\mathbf{z}_i + \alpha \mathbf{d}_{i,j}; \mathbf{y})$ (right) and asked to write freeform text to describe the change from L to R. We used a value of $\alpha = 6$ for all experiments.

## S.2 Concept categorization

Here we provide a breakdown of concepts into three categories reflecting their use, as described in Section 3.2 of the main paper. All concepts that appeared more than five times in each image class were categorized by the authors, as well as all concepts that appeared more than 20 times across all four classes. We sort concepts into three broad categories: *object*, including collective nouns and regions of scenes (e.g., *people, ocean, road*), *attribute* (descriptors of object and scene qualities, including color), and *geometry* (scene- and object-level geometry, including size, perspective, and position). We report results in Table S1.

Attributes are the largest category of concepts in every image class. Concepts describing color and light make up 50% of all attributes: 38% are chromatic color, 12% are related to light and dark. Attributes are also the most reliably detected concepts, both automatically and by humans (see Section S.3).

## S.3 Concept detection accuracies

Here we report human and SVM accuracies in detecting the addition of individual concepts to generated images.

| | Cottage | Kitchen | Lake | Medina | All classes |
|---|---|---|---|---|---|
| **Object** | 35% | 29% | 24% | 25% | 32% |
| **Attribute** | 48% | 50% | 54% | 54% | 48% |
| **Geometry** | 17% | 21% | 22% | 21% | 20% |
| **# of terms** | 178 | 140 | 184 | 139 | 152 |

Table S.1: Distribution of frequently used concepts across three classes: names of objects, scene- and object-level geometry, and other attributes (such as color or lighting). This shows terms that appear 5+ times within each class, and 20+ times across all classes.

| Cottage | | Kitchen | | Lake | | Medina | |
|---|---|---|---|---|---|---|---|
| tree | 0.80 | wall | 0.13 | water | 0.53 | alley | 0.60 |
| color | 0.20 | window | 0.87 | tree | 0.93 | wall | 0.60 |
| sky | 0.53 | cabinet | 0.60 | sky | 0.73 | people | 0.73 |
| building | 0.20 | color | 0.67 | cloud | 0.73 | color | 0.67 |
| grass | 0.80 | white | 0.87 | color | 0.87 | darker | 0.93 |
| green | 1.00 | lighter | 0.80 | blue | 1.00 | street | 0.20 |
| window | 0.73 | counter | 0.53 | darker | 0.93 | blue | 0.87 |
| darker | 0.73 | darker | 0.93 | green | 0.87 | sky | 0.40 |
| roof | 0.40 | brown | 1.00 | reflection | 0.80 | window | 0.67 |
| white | 0.87 | brighter | 0.73 | mountain | 1.00 | light | 0.60 |
| front | 0.67 | wood | 0.87 | land | 0.73 | brighter | 0.53 |
| red | 0.67 | floor | 0.33 | brighter | 0.47 | door | 0.33 |
| smaller | 0.53 | space | 0.27 | grass | 0.87 | white | 0.87 |
| snow | 0.93 | blue | 1.00 | background | 0.33 | red | 0.80 |
| angle | 0.53 | yellow | 0.93 | lighter | 0.40 | B&W | 0.80 |
| blue | 0.80 | smaller | 0.53 | building | 1.00 | yellow | 1.00 |
| B&W | 1.00 | angle | 0.13 | yellow | 1.00 | arch | 1.00 |
| larger | 0.47 | warmer | 0.73 | B&W | 0.93 | background | 0.47 |
| cloud | 0.53 | red | 1.00 | day | 0.13 | road | 0.60 |
| brown | 0.67 | table | 0.13 | sunset | 0.93 | wider | 0.73 |
| **Average** | 0.65 | | 0.65 | | 0.76 | | 0.67 |

Table S.2: Human accuracy detecting the 20 most frequent concepts by category in Experiment 1. Chance is 0.25. Black concepts are objects (including collective nouns and larger scene regions, e.g. water), blue concepts are attributes (adjectives, including colors), and green concepts describe scene- and object-level geometry.

**Human performance per concept.** Experiment 1 (described in Section 4.1 of the main paper) evaluated the generalizability of our vocabulary across $\mathcal{Z}$ by measuring human accuracy discriminating a target concept among three distractors. Table S.2 reports mean accuracy across 15 workers per concept, for the 20 most frequent concepts in each class. Mean human accuracy classifying *attributes* (0.79, $\sigma$ = .21) is higher than either objects (0.64, $\sigma$ = .25 ) or geometry (0.47, $\sigma$ = 0.17). All but one of the attributes shown in Table S.2 describe chromatic color or light, which we might expect to be more reliably discriminable across images and observers.

**SVM performance per concept.** As described in Section 4.1, we replicated Experiment 1 using a linear SVM to distinguish the addition of a particular concept to images from the addition of distractors. For the top 20 most frequent concepts in each of the four classes, 64 $\mathbf{z}$ were randomly sampled, and two classes of images were created to train the SVM: $G(\mathbf{z}+\mathbf{d}_*, \mathbf{y})$ where $\mathbf{d}_*$ is the target concept, and $G(\mathbf{z}+\mathbf{d}_j, \mathbf{y})$ where the $\mathbf{d}_j$ are randomly sampled from the other 19 concepts. 20% of images were held out for testing.

We report classification accuracy for the top 20 concepts in all four classes in Table S3. The color of each concept reflects its category (see Section S.2). Like in the human experiment, mean SVM accuracy classifying *attributes* (0.83, $\sigma$ = 0.09) is higher than either objects (0.75, $\sigma$ = 0.08) or geometry (0.72, $\sigma$ = 0.08).

| Cottage | | Kitchen | | Lake | | Medina | |
|---|---|---|---|---|---|---|---|
| tree | 0.77 | wall | 0.73 | water | 0.73 | alley | 0.73 |
| color | 0.92 | window | 0.81 | tree | 0.81 | wall | 0.73 |
| sky | 0.77 | cabinet | 0.65 | sky | 0.65 | people | 0.73 |
| building | 0.77 | color | 0.85 | cloud | 0.85 | color | 0.85 |
| grass | 0.77 | white | 0.62 | color | 0.69 | darker | 1.00 |
| green | 0.73 | lighter | 0.88 | blue | 0.96 | street | 0.77 |
| window | 0.77 | counter | 0.50 | darker | 0.85 | blue | 0.92 |
| darker | 0.85 | darker | 0.95 | green | 0.95 | sky | 0.65 |
| roof | 0.73 | brown | 0.80 | reflection | 0.85 | window | 0.77 |
| white | 0.73 | brighter | 0.85 | mountain | 0.69 | light | 0.77 |
| front | 0.81 | wood | 0.81 | land | 0.62 | brighter | 0.85 |
| red | 0.85 | floor | 0.69 | brighter | 0.85 | door | 0.77 |
| smaller | 0.73 | space | 0.58 | grass | 0.85 | white | 0.81 |
| snow | 0.88 | blue | 0.81 | background | 0.69 | red | 0.85 |
| angle | 0.77 | yellow | 0.92 | lighter | 0.81 | B&W | 0.62 |
| blue | 0.77 | smaller | 0.77 | building | 0.88 | yellow | 0.88 |
| B&W | 0.88 | angle | 0.65 | yellow | 0.65 | arch | 0.62 |
| larger | 0.85 | warmer | 1.00 | B&W | 0.73 | background | 0.62 |
| cloud | 0.73 | red | 0.65 | day | 0.88 | road | 0.77 |
| brown | 0.96 | table | 0.77 | sunset | 0.85 | wider | 0.77 |
| **Average** | 0.80 | | 0.76 | | 0.79 | | 0.77 |

Table S.3: SVM accuracy classifying 20 most frequent concepts by category. The same color scheme is used as in Table S.2. Chance is 0.50.

## S.4 Generalization to BigGAN-ImageNet

Our method generalizes to BigGAN-ImageNet, as referenced in the main paper. We include details in this section. The generalizability of our approach suggests that it could be used to characterize a given generator by the projection of concepts salient to humans into the set of concepts the model has learned.

**Distilling visual concepts.** We generate layer-selective directions using the method described in Section 3.1 for 64 randomly selected $\mathbf{z}$ in two classes of BigGAN-ImageNet that best resemble classes of BigGAN-Places: *lakes* (shared by both datasets) and *barns* (similar to cottages). As in our procedure for BigGAN-Places, 1280 layer-selective directions are found in each class. Annotations are then collected on AMT, normalized, and post-processed using the procedure described in Sections 3.2 and S.1. From the annotated layer-selective directions, we use the method described in Section 3.3 to distill visual concepts and associated directions in latent space. We find that 1198 unique terms are used to describe barns, with 555 repeated at least once, and 867 unique terms are used to describe lakes, with 390 repeated at least once. Selected directions in both classes are visualized in Supplementary Figure 2.

**Concept evaluation.** Following Section 4.1, we use a forced choice task on AMT to evaluate the salience of visual concepts in the latent space of BigGAN-ImageNet and their interpretability across different $\mathbf{z}_i$. Results across both classes are shown in Supplementary Figure 3.

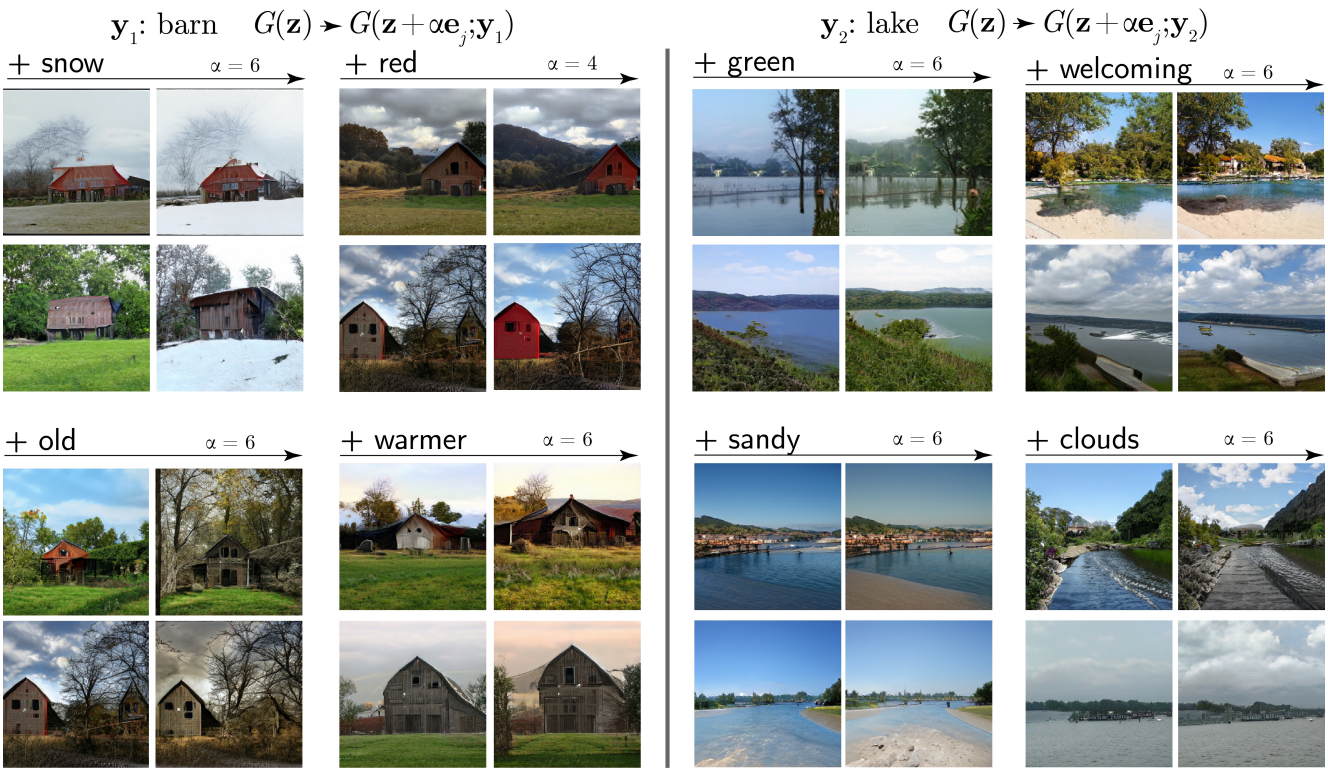


Figure S2: Example visual concepts found in the latent space of BigGAN-ImageNet using our method. The *lake* class is the only visual scene class shared by both BigGAN-ImageNet and BigGAN-Places. For the same number of annotated directions (1280), the number of distinct concepts in the *lake* class for BigGAN-ImageNet is $< 75\%$ of the number of distinct concepts in the *lake* class for BigGAN-Places (see Section 3.2, Table 1). This could reflect less scene diversity in comparable ImageNet classes due to less training data.

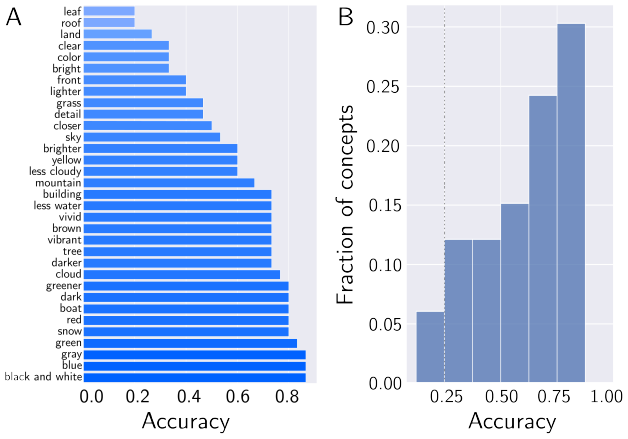


Figure S3: Task accuracies for concepts computed across $\mathbf{z}$, workers, and class. (a) Accuracies for concepts that appeared more than 20 times in the annotation dataset. Some concepts (including color changes like *green, gray, blue, black and white*) are reliably recognized across most $\mathbf{z}$, while others (such as *leaf* and *roof*) are not recognized with accuracy above chance. (b) Histogram of concept accuracies across all concepts. The dotted vertical line shows the accuracy of random guessing (0.25).

Using the procedure described in Section 4.1 and visualized in Appendix C, AMT workers are recruited to identify each concept within a set of distractors. Specifically, for each concept $\mathbf{c}_*$ and its distilled direction $\mathbf{d}_*$, we sample a novel $\mathbf{z}$ from the $\mathcal{Z}$ latent space as well as three distractor directions $\{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3\}$ sampled uniformly at random from the remaining directions. Workers are shown an initial image $G(\mathbf{z}; \mathbf{y})$ and four modified images $G(\mathbf{z} + \alpha\mathbf{d}_i; \mathbf{y})$ for $i = 1; 2; 3; *$ and are asked to discriminate which modified image corresponds to $\mathbf{c}_*$. If the direction $\mathbf{d}_*$ successfully generalized to $\mathbf{z}$, then workers should reliably choose the image change generated by that direction. We run the evaluation on 3 $\mathbf{z}$ per direction and show each ($\mathbf{z}$, $\mathbf{d}$) pair to 5 distinct workers. We use $\alpha = 6$ in our experiments. We find that workers reliably choose the correct image with $61.1\%$ overall accuracy across ($\mathbf{z}$, $\mathbf{d}$) in the *barn* class, and $61.6\%$ overall accuracy in the *lake* class. Observers only fail to discriminate about $6\%$ of concepts. For all other concepts, observers recognize the correct change more often than if they guessed randomly, demonstrating that our method is successful at discovering directions that generalize across the latent space of BigGAN-ImageNet.

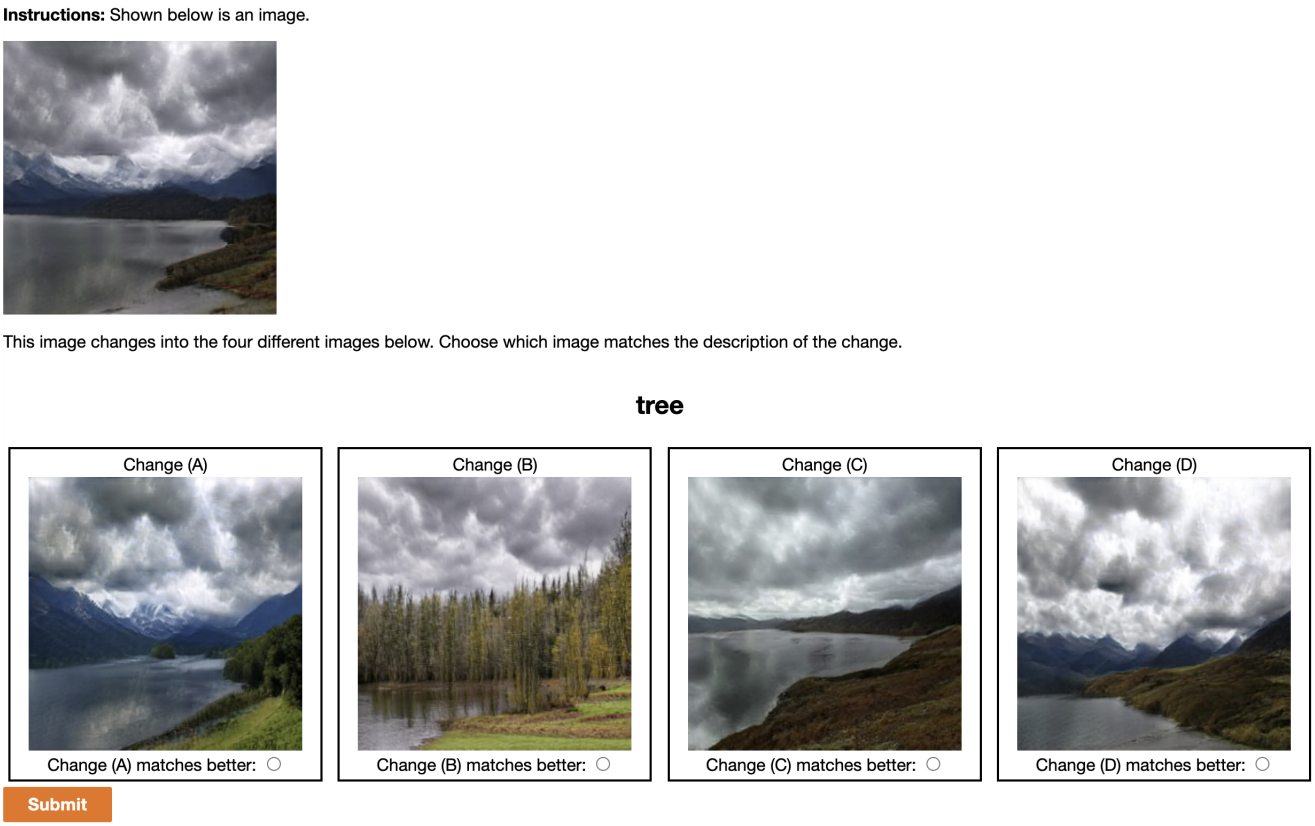


Figure S4: Example multiple choice HIT used to test concept generalization across image class. Here, the **tree** direction in $\mathcal{Z}$ latent space was learned in the *cottage* class, and is being tested in the *lake* class. The same multiple choice format is used to test concept composition, where a target composition (e.g. **tree, greener**) is described, and workers select which of four images best captures the composition.

## S.5 Generalization and composition experiments

**Experimental Paradigm.** In Figure S4 we show a screenshot of the paradigm used to collect data on AMT for analyses in Sections 4.1, 4.2, and 4.3, testing direction generalization across $\mathcal{Z}$ and image class, and composition.

Workers are shown an original image $G(\mathbf{z}; \mathbf{y})$ for randomly selected $\mathbf{z}$ and asked to identify which of four transformed images best corresponds to a named concept $\mathbf{c}_*$. $\mathbf{c}_*$ is randomly positioned among four distractors. Concepts used to create the distractor images are randomly sampled from the list of concepts that appeared more than five times in the annotation data. $\alpha$ was set to 6 for all experiments, and 5 distinct workers performed each task.

To test salience of concepts and their generalization across $\mathbf{z}$ in a given class (Section 4.1), concepts are tested for the same class $\mathbf{y}$ they were drawn from, for 3 different $\mathbf{z}$ per concept. To test generalization across class (Section 4.2), concepts are tested on a class (selected at random for each task) other than the one they were drawn from. Concept composition is tested using the method described in Section 4.3. Workers select between 4 modified images, one of which corresponds to a pair of concepts named in the task (e.g. [**tree, greener**]) and is randomly positioned among 3 other compositions (e.g. [**tree**, **brown**], [**tree**, **larger**], [**larger**, **brown**]) where **larger** and **brown** represent distractor concepts randomly selected for each task. Workers were required to be located in the U.S., with $> 95\%$ HIT acceptance rate and $> 100$ HITs accepted. Workers were paid $\$0.06$ per HIT.

## S.6 Additional qualitative results

In Figure S5 we visualize additional examples of concept subtraction as well as addition for varying $\alpha$ (compare to Figure 4 in the main paper). While most directions generalize across $\mathbf{z}$ and some across $\mathbf{y}$, others do not. Furthermore, Section 4.3 suggests we can compose concepts. This works for concepts that regularly occur in the same class, and some that do not co-occur, but some combinations do not work. In Figure S6 we show examples of concepts that fail to generalize across $\mathcal{Z}$ or class, or fail to compose with other concepts.

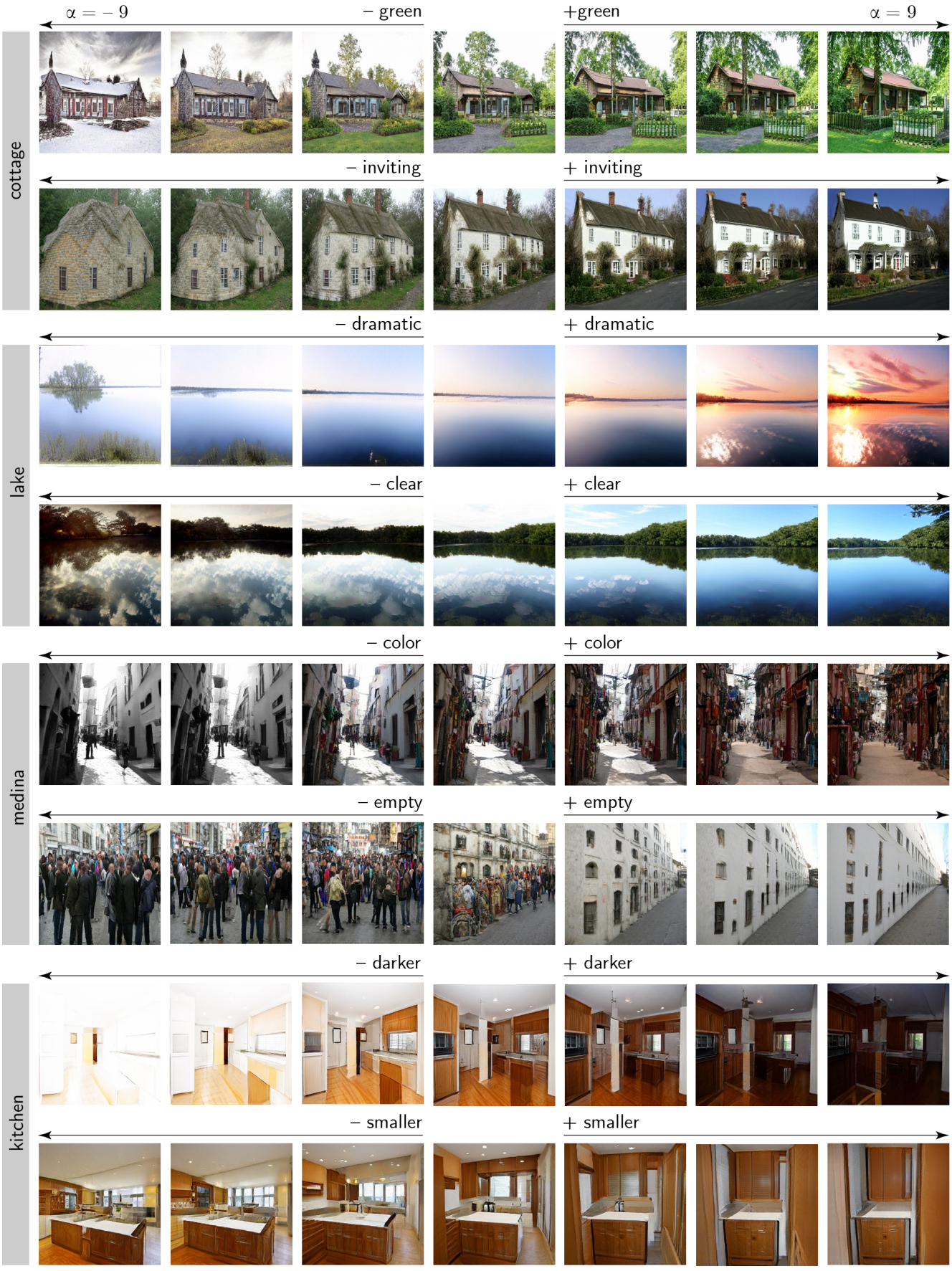


Figure S5: Additional examples of concept addition and subtraction. $\alpha$ is varied in steps of size 3.

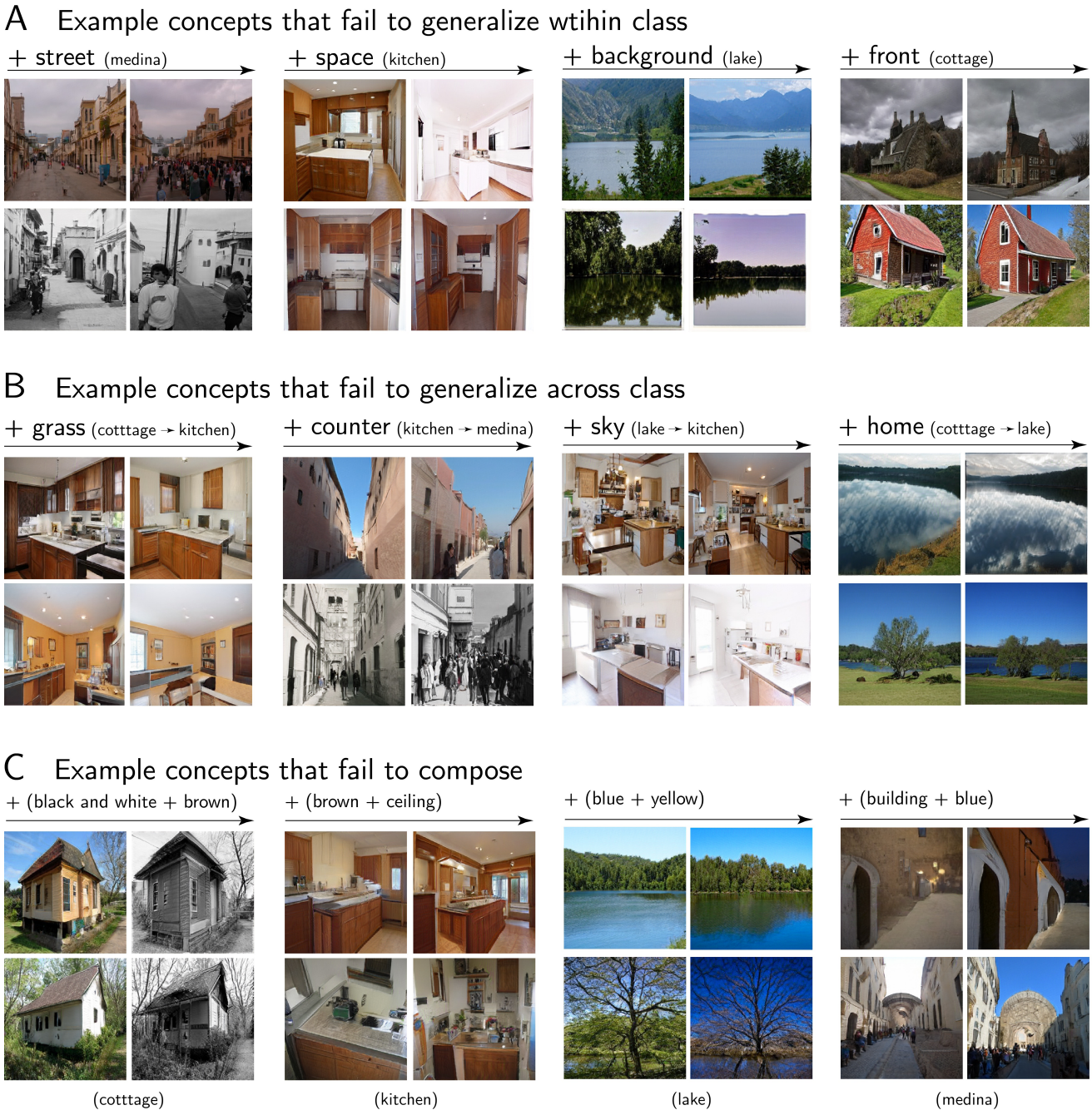


Figure S6: Example failures. (a) Shows sample concepts that did not perform at accuracy above chance in the AMT task described in Section 4.1. Many of these concepts, such as *space*, are broad in scope and could be used to describe many kinds of scene changes. (b) Shows sample concepts that each performed at accuracy above chance *within class* in the AMT task described in Section 4.2, but failed to perform above chance when tested in a *different class*. (c) Shows concepts that perform above chance individually but not when composed, in the experiment described in Section 4.3.