When Agents Steer Human Perception: How AI-Selected Images Can Convertly Alter Disagreements

Supplementary Material

1. Concept Prompt Templates

To generate interpretable visual concepts for each category pair, we used the following prompt format with GPT-4o:

> **User prompt:**
> Please list 10 distinct, visually interpretable concepts (descriptive phrases) that best distinguish images of birds and cats.
> **Guidelines:**
>
> 1. Think of observable attributes.
> 2. Do not use "bird" or "cat" themselves.
> 3. Each concept should be something CLIP could encode, and that tends to vary between birds and cats.
> 4. Give the list as bullet points, in no particular order.

This process was repeated five times per pair to generate 50 raw candidates. GPT-4o was then asked to summarize and select the 10 most representative concepts using the instruction:

> Select the 10 concepts with the most repetitions from the following 5 groups of 10 concepts. Similar concepts are also considered repetitions. If there are similar concepts, return the one you think is the best.

2. Generation method

To systematically generate a substantial dataset of images that elicit human perceptual disagreement, we adopted the following methodology:

For each pair, we used GPT-4o to produce a small preliminary set of "seed" images (approximately 20). These samples were specifically designed to blur the distinctions between Animal A and Animal B, incorporating features from both in a manner that induces perceptual uncertainty in humans.

Subsequently, these carefully selected seed images served as training data for fine-tuning a pre-trained diffusion model (e.g., Stable Diffusion) using the DreamBooth method.

Upon completion of the DreamBooth training, we then utilized the fine-tuned model to generate a larger collection of 400 images. This expanded set will constitute our official experimental image dataset, which will be used for subsequent evaluation of their efficacy in eliciting human perceptual disagreement.

For generating the initial "seed" images with GPT-4o, the following prompts were employed:

User Prompt: "Please generate an image of an animal that appears to be a blend of both Animal A and Animal B, possessing characteristics from both to an extent that makes it difficult to distinguish clearly."

Negative Prompt: "unreal, unnatural, bad"

3. Fine-tuning method

Model training was conducted on an initial experimental dataset comprising 200 image samples. We employed three distinct training methodologies for comparison:

- **Method I**: **LoRA Fine-tuning of CLIP Image Encoder:** This approach directly fine-tuned the pre-trained CLIP image encoder, incorporating Low-Rank Adaptation (LoRA) technology with Rank=8, Alpha=8.
- **Method II**: **Linear-probing + Image Loss:** A binary classification linear layer was added on top of a frozen CLIP image encoder, and training was performed solely with an image-feature-based loss function.
- **Method III**: **Linear-probing + Image and Text Loss:** Similarly, a binary classification linear layer was appended to a frozen CLIP image encoder. Distinct from Method II, this approach integrated both image and text losses.

All three methods were trained on NVIDIA GPUs using the AdamW optimizer, with a learning rate of 1×10^{-3}, for 20 epochs. Each method was independently fine-tuned on both the RN50 and ViT-L/14 versions of CLIP to ensure consistent comparison.

Loss Function Details for Method III

For Method III, the total loss function design considers both image content and text descriptions. Specifically, in addition to computing the cross-entropy loss between the predicted results of the input image (after passing through the image encoder and classifier) and the human-annotated labels, we also incorporate the text labels for each of the two animal categories as "one-hot" training samples into the loss function.

We define the following variables:

- x: Input image sample.

- y_x: Human-annotated label corresponding to image x.
- z: Text description for a certain class in the binary classification task (e.g., a photo of + animal name).
- y_z: Label corresponding to text description z.
- ϕ_i: Frozen CLIP image encoder.
- ϕ_t: Frozen CLIP text encoder.
- $\mathcal{C}$: Binary classifier (linear layer) appended after the image encoder's output.
- $\mathcal{H}$: Cross-entropy loss function.

Then, the total loss function $\mathcal{L}$ for Method III is defined as:

$$\mathcal{L} = \mathcal{H}(\mathcal{C}(\phi_i(x)), y_x) + \mathcal{H}(\mathcal{C}(\phi_t(z)), y_z) \quad (1)$$

The 36 category pairs consist of pairwise combinations among cats, dogs, fish, frogs, turtles, monkeys, insects, birds, and crabs. Below we show the cluster maps for all category pairs. Top 3 concepts for each subject are shown in *top3 concepts for all pairs* folder as csv files. We also show some images that we used in our experiment as examples.

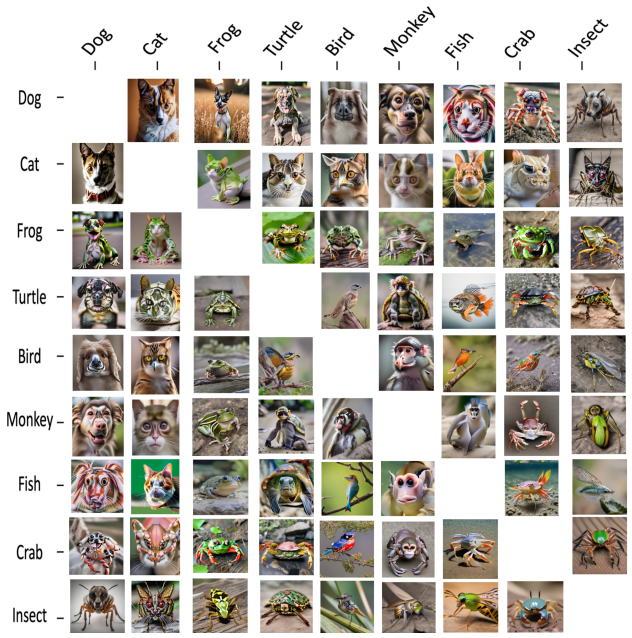

Figure 1. **Examples of stimuli in the experiment.**We only use the categories from the Restricted ImageNet

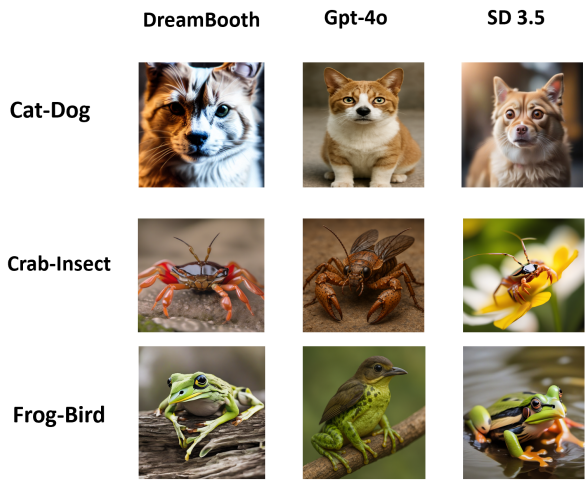

Figure 2. **Examples of images by different generating methods**

Figure 3. **Examples of images of in-class and super-class**

Concept Description
feathers with layered texture
smooth, moist skin surface
outstretched wings in motion
wide-set bulging eyes
perched on a tree branch
webbed toes gripping a surface
camouflaged among green leaves
slender beak-like mouth
long legs bent for leaping
clawed feet gripping surface

Concepts for Bird-Frog pair

Figure 4. Examples of Visually Interpretable Concepts

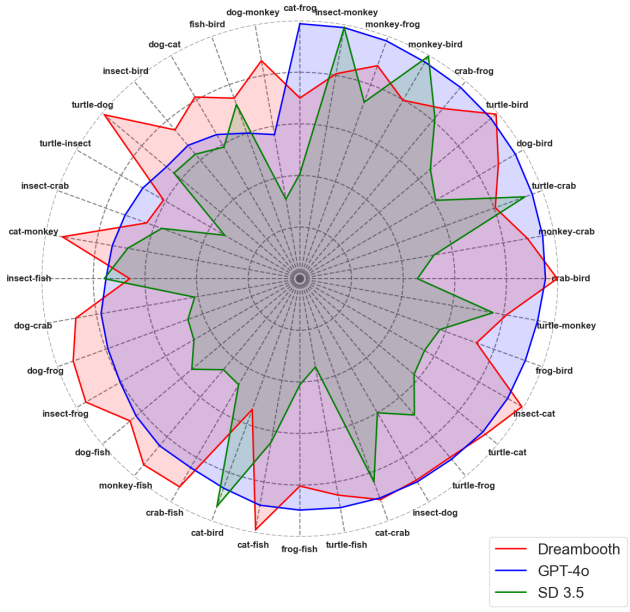

Figure 5. Comparison of three image generation methods:GPT-4o,SDV3.5,and DreamBooth.

4. Information about NaoDao

The Naodao Service Platform(www.naodao.com) is a comprehensive public platform for online psychological research, designed specifically for researchers in psychology, cognitive neuroscience, and behavioral research. It provides a one-stop solution from research design to data collection. The platform integrates core functions including online experiment programming, questionnaire design, intelligent participant recruitment, data collection and quality control, and scientific research resource sharing. It significantly lowers the technical threshold for online research while ensuring the reliability of data quality and the efficiency of research processes.

5. Participant Demographics

Here is a summary of the demographic characteristics of the participants:

- **Gender Distribution**
  The participant group was composed of 30% male and 70% female participants.
- **Age Distribution**
  Participants' ages ranged from 10 to 60 years. The mean age was 24.98, with a standard deviation of 6.94.
- **Geographic Distribution**
  The participants represented a diverse geographic spread, hailing from 120 different regions.
- **Educational Background**
  The educational attainment of the participants varied, en-

compassing the following levels:
- Vocational/Technical School
- Junior College
- Bachelor's Degree
- Master's Degree
- Doctoral Degree

Collectively, the participants majored in 21 different academic disciplines.

6. More experiment details and results for the main experiment

We conducted 36 sets of experiments, each set with data from 20 participants after filtering through sentinel trials. We also provide the results for **random filtering baseline**: the success rate is 34.60 ± 15.37%, while the targeted ratio is 49.93 ± 1.07%. For reference, our method achieves around 50% in success rate and 78% in targeted ratio, confirming the effectiveness of our method. We also provide exact correlation values for **Figure 3c**: before alignment, correlation value is -0.0917, after alignment, correlation value is 0.4522.

We also conducted a two-round causal check on participants: agents pre-selected items before any labels in that round; participants then viewed agent-selected subsets (which do not overlap with the first round) and provided their judgments. Results match our main experiment findings, with a +7% increase in success rate and a +30% increase in disagreement ratio, further confirming the effectiveness of our main setup. Stage-2 selection is performed solely by aligned agents' predictions; alignment, selection, and evaluation labels are strictly disjoint, ruling out label leakage in both the main study and this check. Regarding the **casual check process**: casual check participants were recruited offline. After completing the first round of selections, they waited for approximately 2 hours. Their individual models were then used to filter and create new image sets, which were shown to the participants. Due to the small amount of data collected and the limited category pairs tested, this may have caused differences with the main experimental results.

7. Mathematical Definition on Metrics

Disagreement Metrics: Let n be the number of agent-disagreement images, and let k be the number of **Success** images among them. We define the success rate as $\mathrm{SuccessRate} = k/n$. Moreover, let k_p and k_n be the numbers of **Positive** and **Negative** images within the successful set, respectively, with $k = k_p + k_n$. We define the targeted ratio (a.k.a. positive rate in Fig. 5) as the proportion of **Positive** outcomes within **Success**: $\mathrm{TargetedRatio} = k_p/k$.

Activation targets: For our classifier, let the two class logits be $\ell_1, \ell_2 \in \mathbb{R}$, and probabilities $p_1, p_2 \in (0, 1)$ be obtained by softmax: $p_i = \frac{\exp(\ell_i)}{\exp(\ell_1)+\exp(\ell_2)}, \quad i \in \{1, 2\}$. **(1) Single-logit target.** We use each class logit as an activation target: $T_{\text{logit-}i} = \ell_i, \quad i \in \{1, 2\}$ **(2) In-subject entropy target.** We quantify the model's within-subject uncertainty by the entropy of its predictive distribution: $T_{\text{ent}} = H(\mathbf{p}) = -\sum_{i=1}^{2} p_i \log p_i$, where $\mathbf{p} = (p_1, p_2)$. **(3) Inter-subject cross-entropy target.** Given two agents A and B with predictive distributions $\mathbf{p}^A$ and $\mathbf{p}^B$ on the same image, we define directional cross-entropy activations in both directions: $T_{\text{CE}}^{A\to B} = H(\mathbf{p}^A, \mathbf{p}^B) = -\sum_{i=1}^{2} p_i^A \log p_i^B$, $T_{\text{CE}}^{B\to A} = H(\mathbf{p}^B, \mathbf{p}^A) = -\sum_{i=1}^{2} p_i^B \log p_i^A$.

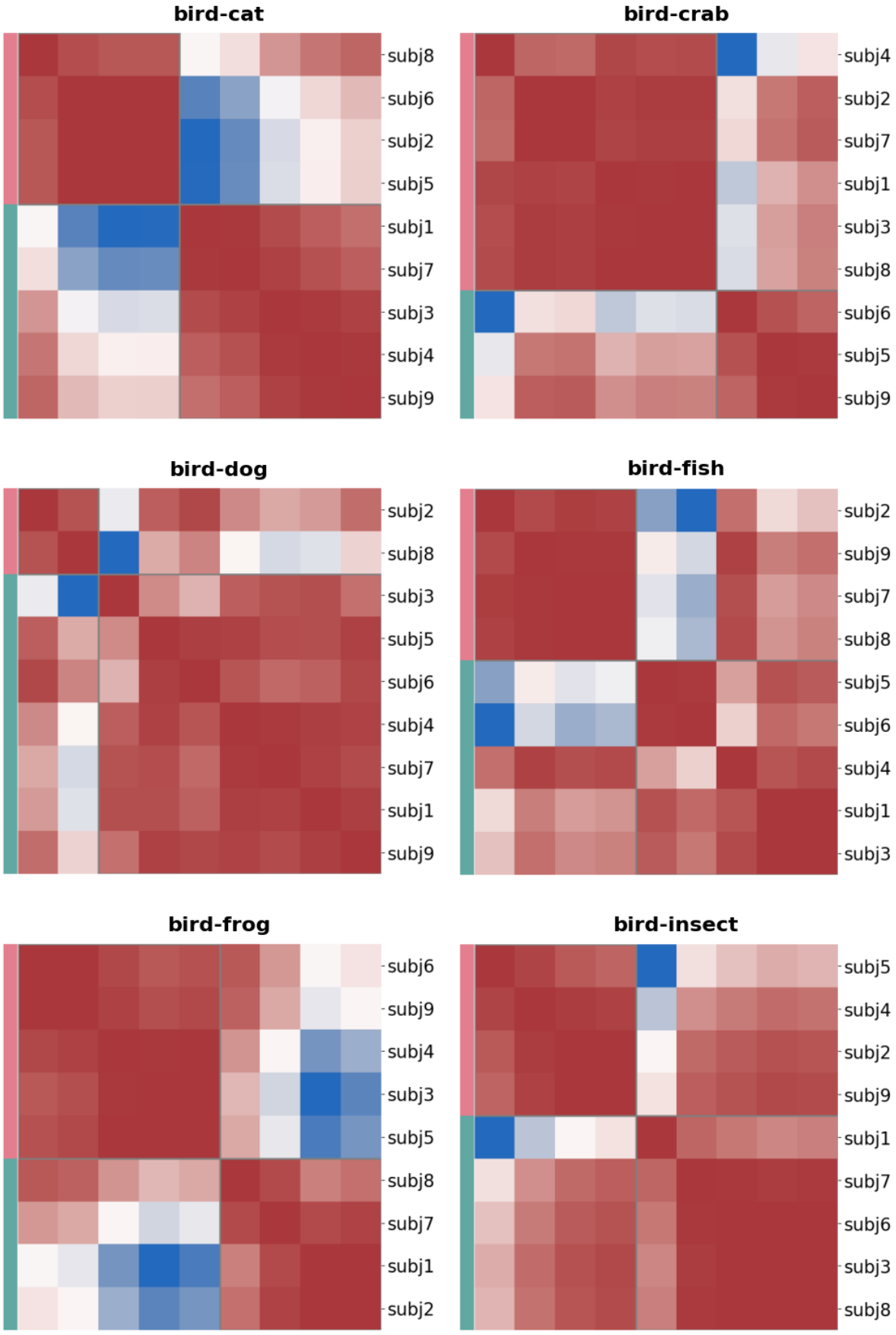

Figure 6. Clustermaps for pairs: crab-fish, crab-frog, crab-insect, crab-monkey, crab-turtle, dog-fish.

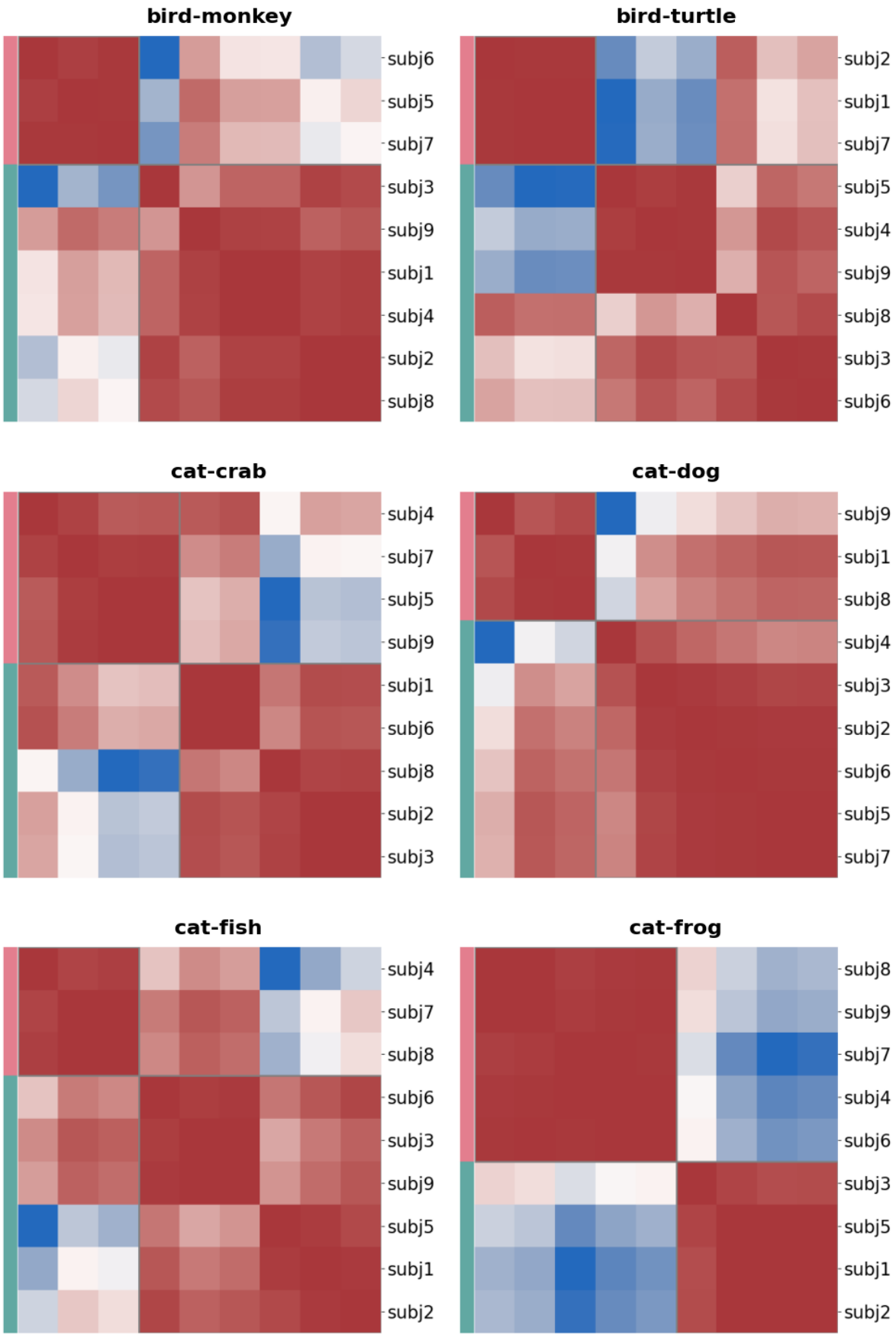

Figure 7. Clustermaps for pairs: dog-frog, dog-insect, dog-monkey, dog-turtle, fish-frog, fish-insect

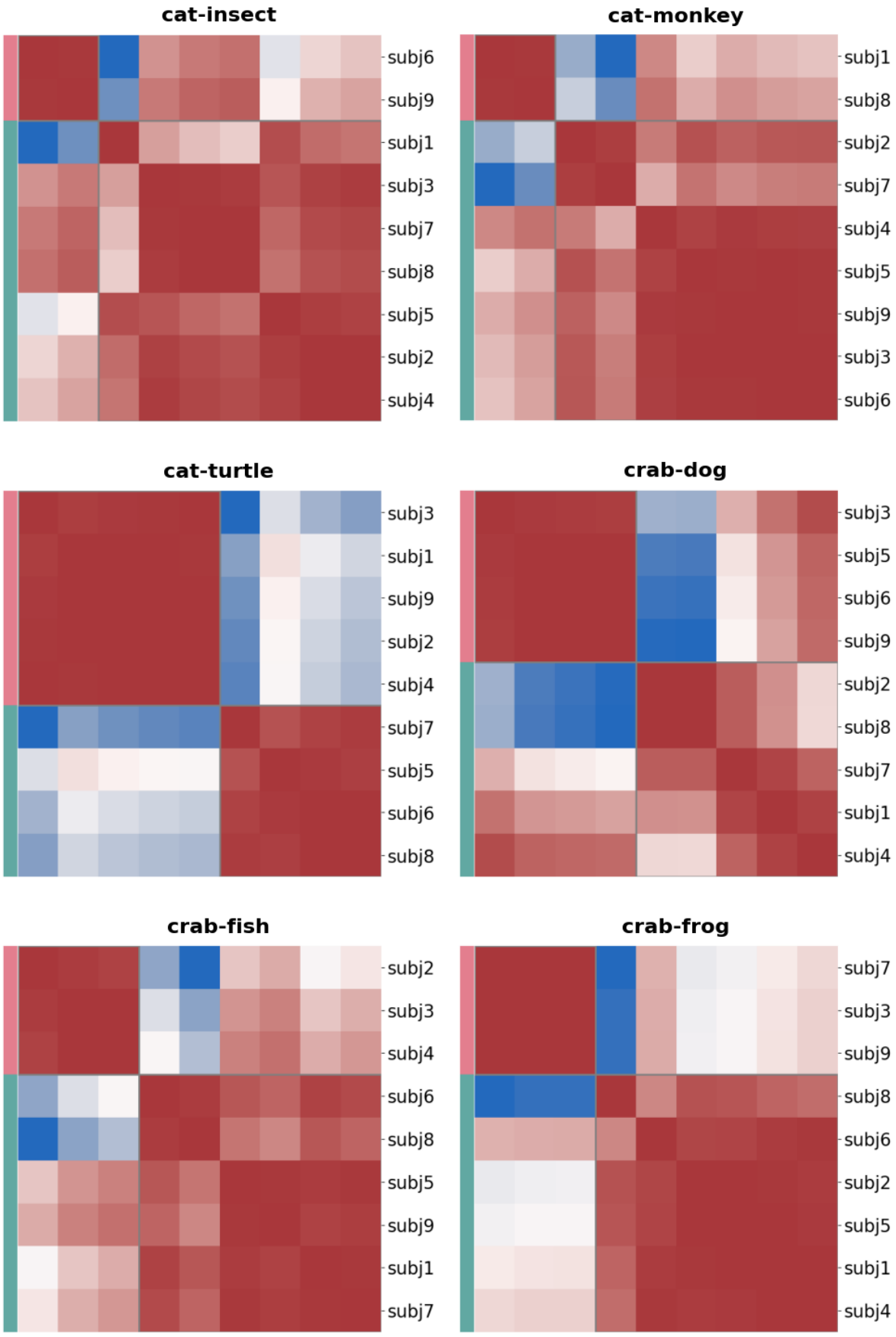

Figure 8. Clustermaps for pairs: fish-monkey, fish-turtle, frog-insect, frog-monkey, frog-turtle, insect-monkey

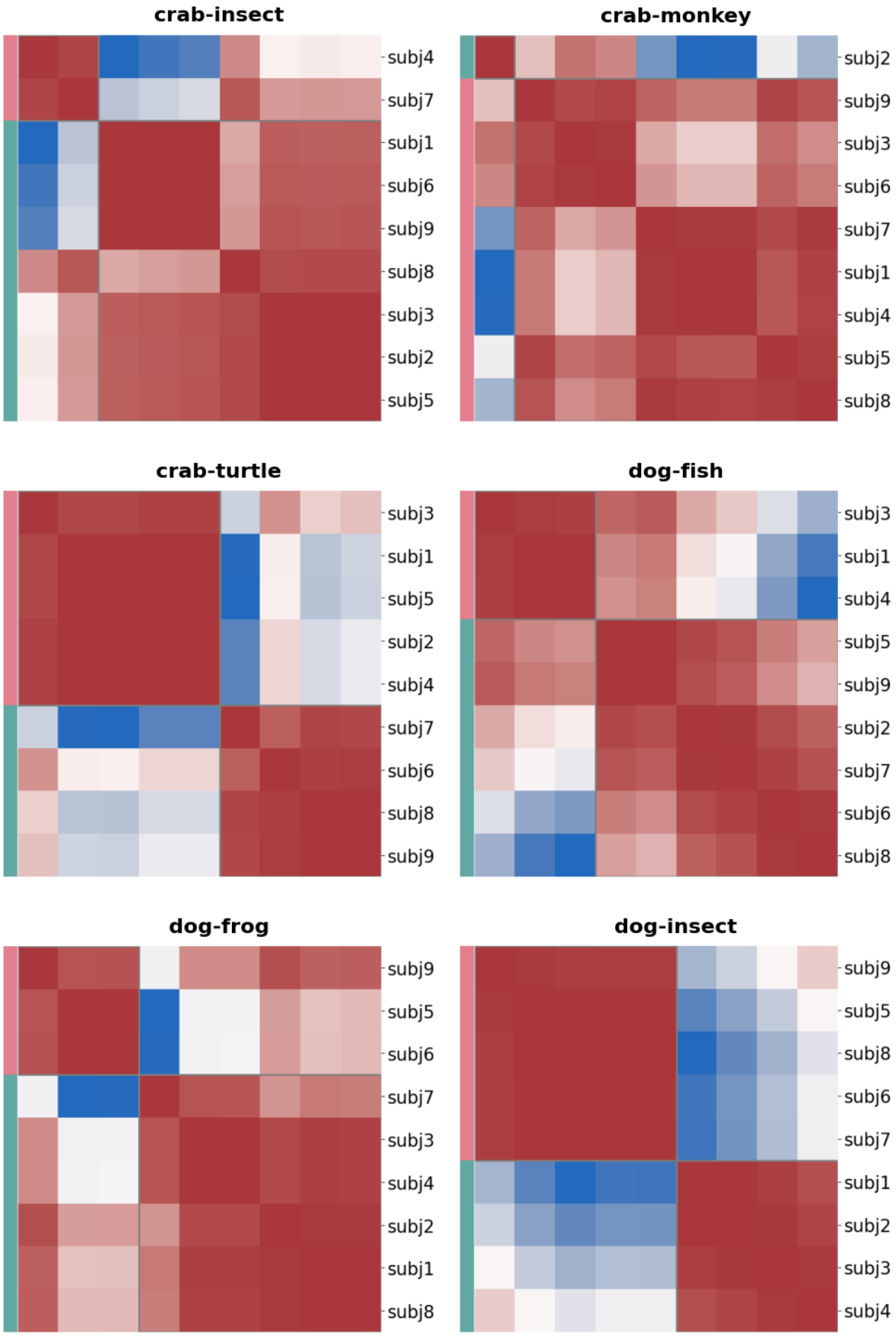

Figure 9. Clustermaps for pairs: insect-turtle, turtle-monkey, bird-cat, bird-crab, bird-dog, bird-fish

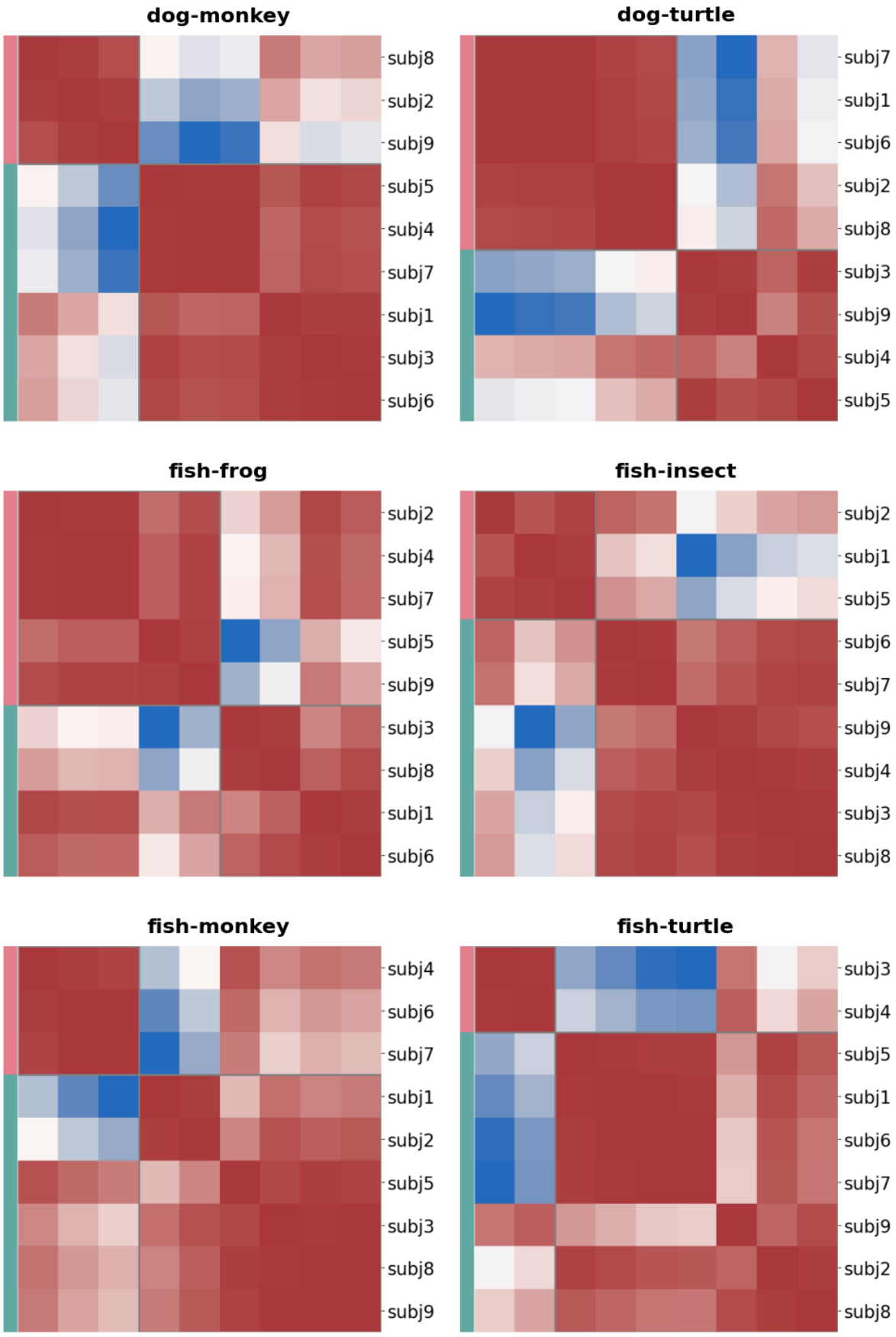

Figure 10. Clustermaps for pairs: bird-frog, bird-insect, bird-monkey, bird-turtle, cat-crab, cat-dog

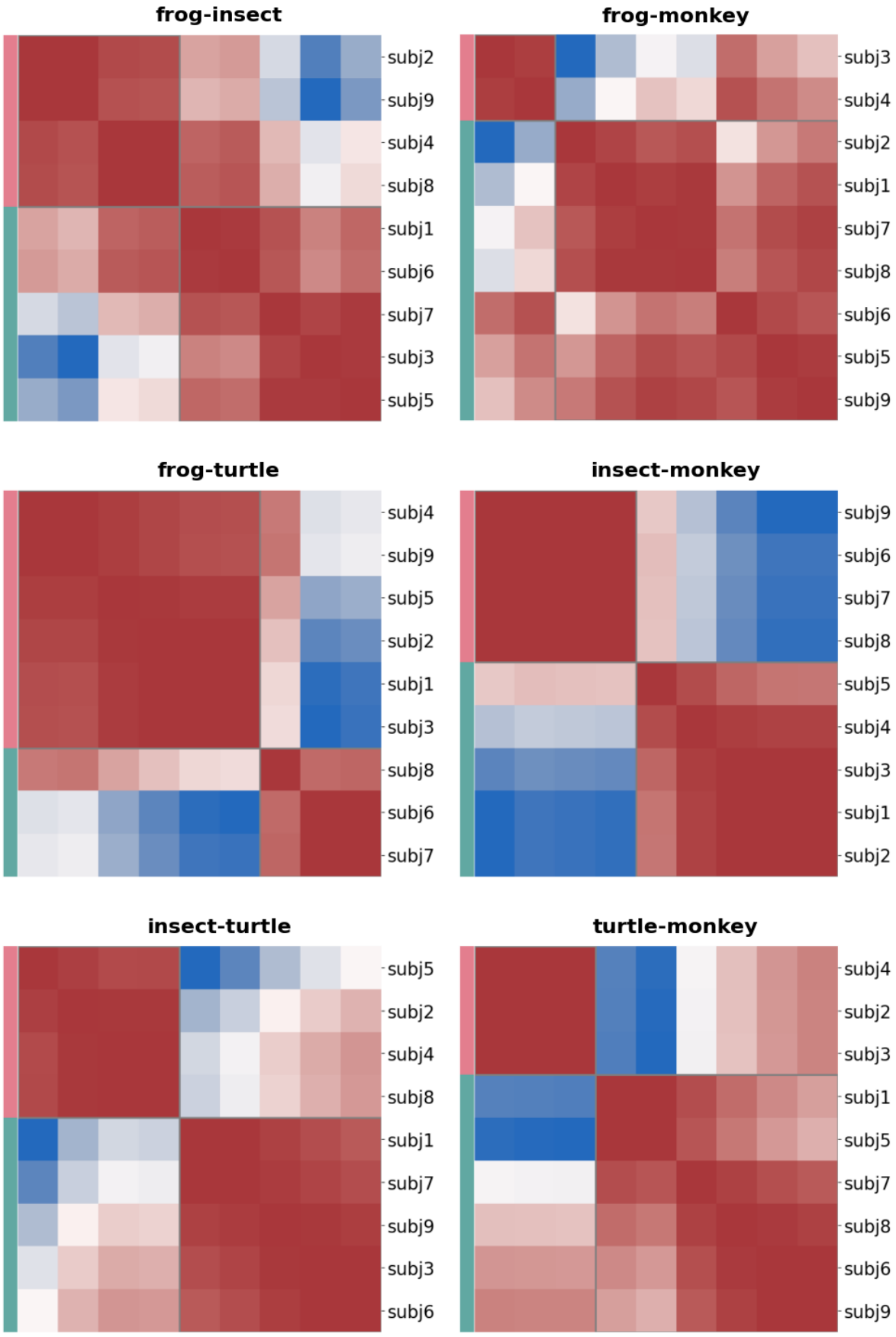

Figure 11. Clustermaps for pairs: cat-fish, cat-frog, cat-insect, cat-monkey, cat-turtle, crab-dog