Supplementary Materials for Grid Distil

Anonymous Authors

1 Primer on Submodular Optimization

Submodular functions play a central role in our grid construction framework due to their natural ability to model coverage, diversity, and informativeness under discrete selection constraints. This section provides a brief primer to contextualize our objective in main paper and to clarify why submodular maximization offers an effective and scalable mechanism for selecting representative image subsets.

1.1 Background

A set function $f : 2^{\mathcal{U}} \to \mathbb{R}$ defined over subsets of a ground set $\mathcal{U}$ is *submodular* if, for all $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{U}$ and any element $x \in \mathcal{U} \setminus \mathcal{B}$,

$$f(\mathcal{A} \cup \{x\}) - f(\mathcal{A}) \;\geq\; f(\mathcal{B} \cup \{x\}) - f(\mathcal{B}),$$

a property known as *diminishing returns*. Submodularity naturally captures scenarios in which adding new elements becomes less informative as the selected set grows—precisely the behavior desired when selecting diverse and representative exemplars.

Many classical objectives such as facility-location, DPP log-determinants, and spectral criteria exhibit submodularity or near-submodularity, enabling efficient greedy optimization with strong approximation guarantees (typically $(1 - 1/e)$ for monotone functions under cardinality constraints).

1.2 Our Submodular Objective

For grid selection we consider a candidate pool of M images with similarity kernel $\mathbf{K} \in \mathbb{R}^{M \times M}$. We seek a subset $\mathcal{S}$ of fixed cardinality $|\mathcal{S}| = L^2$, corresponding to the spatial layout of an $L \times L$ grid. Our selection is governed by the composite submodular objective:

$$\begin{aligned} \mathcal{F}(\mathcal{S}) \;&= \alpha \sum_{i \in \mathcal{U}} \max_{j \in \mathcal{S}} K_{ij} \\ &\quad + \beta \log\det(\mathbf{K}_{\mathcal{S},\mathcal{S}} + \epsilon \mathbf{I}) \\ &\quad + \gamma \sum_{i \in \mathcal{S}} s_i, \end{aligned} \tag{1}$$

where $\mathbf{K}_{\mathcal{S},\mathcal{S}}$ denotes the kernel submatrix induced by the selected samples and s_i captures the spectral energy or importance of each candidate.

Each component encodes a complementary property:

- **Coverage (Facility Location; α):** Encourages each candidate to be close to at least one selected exemplar, ensuring the selected set represents all modes of the distribution.
- **Diversity (DPP Log-Det; β):** The log-determinant term rewards sets that span large volume in embedding space, reducing redundancy and favoring orthogonal semantic modes.
- **Spectral Information (γ):** Prioritizes high-energy samples aligned with principal manifold directions, reinforcing selection of semantically informative regions.

The weighted combination of these terms produces a structured, well-behaved objective that promotes high coverage, strong semantic spread, and alignment with intrinsic dataset geometry. This combination would be difficult to achieve using purely clustering-based or purely diversity-driven approaches.

1.3 Greedy Optimization Procedure

Because $\mathcal{F}$ is monotone and submodular, the standard greedy algorithm provides a $(1 - 1/e)$ approximation under the cardinality constraint $|\mathcal{S}| = L^2$. We apply this greedy selection iteratively: after constructing each grid, the selected images are removed from the pool, and the next grid is computed using the remaining candidates until we obtain the desired number of grids G or exhaust the pool. This iterative submodular selection ensures that each grid captures fresh semantic structure while maintaining global distributional coverage.

1.4 Practical Considerations

Our formulation is designed to accommodate large candidate sets efficiently. Kernel computation can be efficiently computed using GPU with the number of images per class (about 1000) in less than a minute, while the greedy maximization remains extremely efficient due to incremental updates of marginal gains. This allows us to perform selection at scale while preserving the theoretical guarantees of submodular optimization.

2 Subjective Evaluation of Detail-Enhanced Reconstructions

To complement the quantitative comparisons presented in the main paper, we provide additional subjective results that examine the visual characteristics of our detail-enhanced reconstructions. Each 2×2 panel contains (top-left) the original 1024×1024 crop, (top-right) the distilled low-resolution representation upsampled to the same size, (bottom-left) the initial noise input used by InvSR, and (bottom-right) the detail-enhanced reconstruction produced by our method.

These panels illustrate how the proposed enhancement module leverages diffusion priors to restore high-frequency structures that are not explicitly preserved in the compact distilled representations. Across the examples, our method consistently recovers sharper edges, finer textures, and more coherent global structure compared to the distilled or noisy inputs. The enhancement stage also introduces minimal hallucination artifacts, maintaining semantic fidelity with respect to the original image. Overall, these qualitative results highlight the ability of our framework to bridge the gap between compressed grid representations and natural image statistics, producing reconstructions that retain both global composition and local detail.

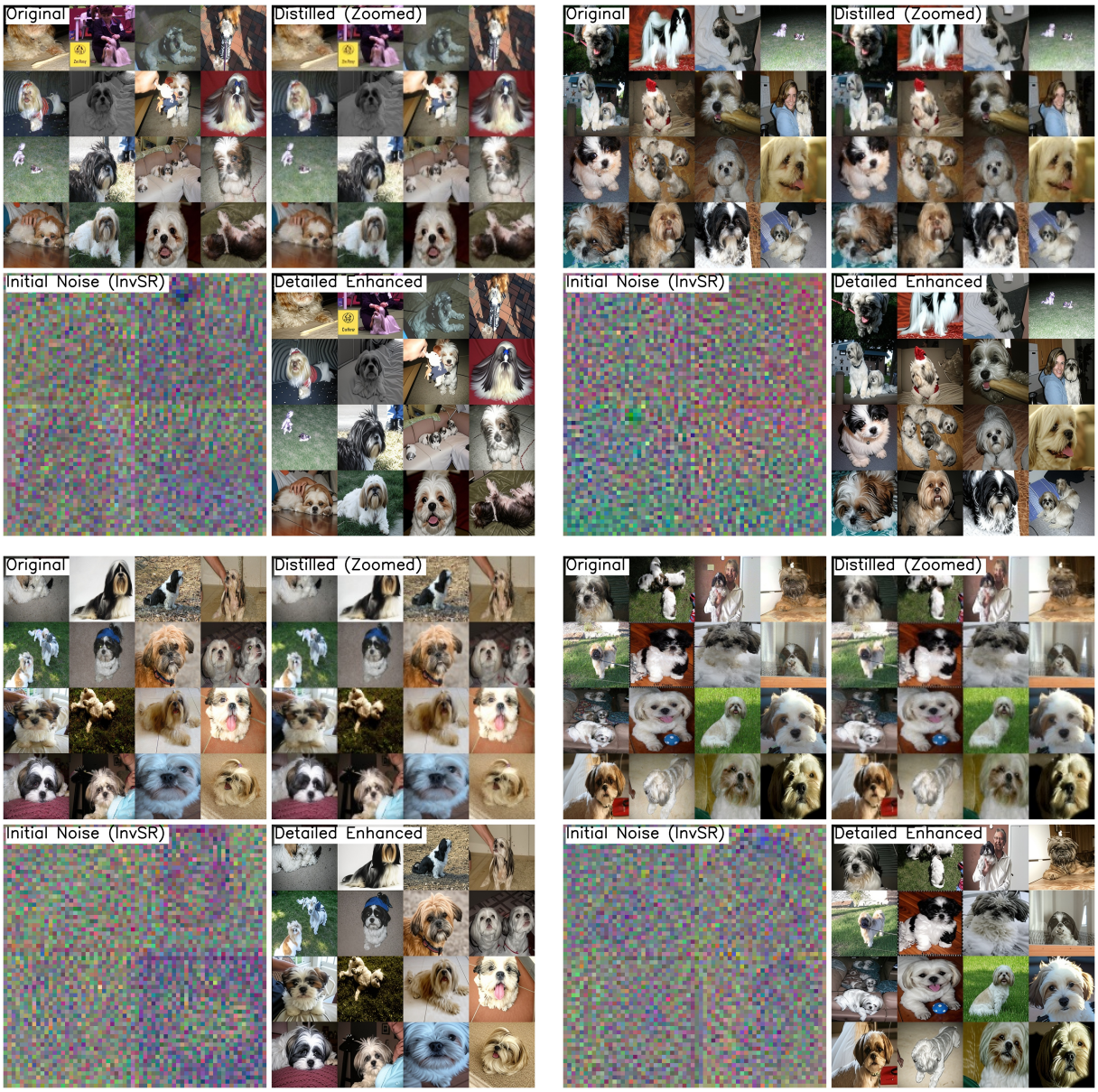

Figure 1: Subjective comparison of reconstructed detail-enhanced images. Each panel displays: Original (top-left), **Distilled (Zoomed/Bilinear)** (top-right), Initial Noise (InvSR) (bottom-left), and Detailed Enhanced result (bottom-right). Our method consistently restores fine structure while preserving global semantic fidelity. **Please zoom on images to check the enhanced details.**

3 Qualitative Comparison of Distilled Images with SOTA

Figure 2 presents a side-by-side comparison of images distilled using SRe^{2}L, Minimax, VLCP, and our proposed Grid Distillation method. Across classes, SRe^{2}L tends to produce low-frequency, texture-like patterns that convey coarse class-level signals but lack spatial structure. Minimax improves visual sharpness but frequently collapses to a narrow set of prototypes, resulting in limited intra-class diversity. VLCP preserves more detailed textures and yields cleaner images, yet often

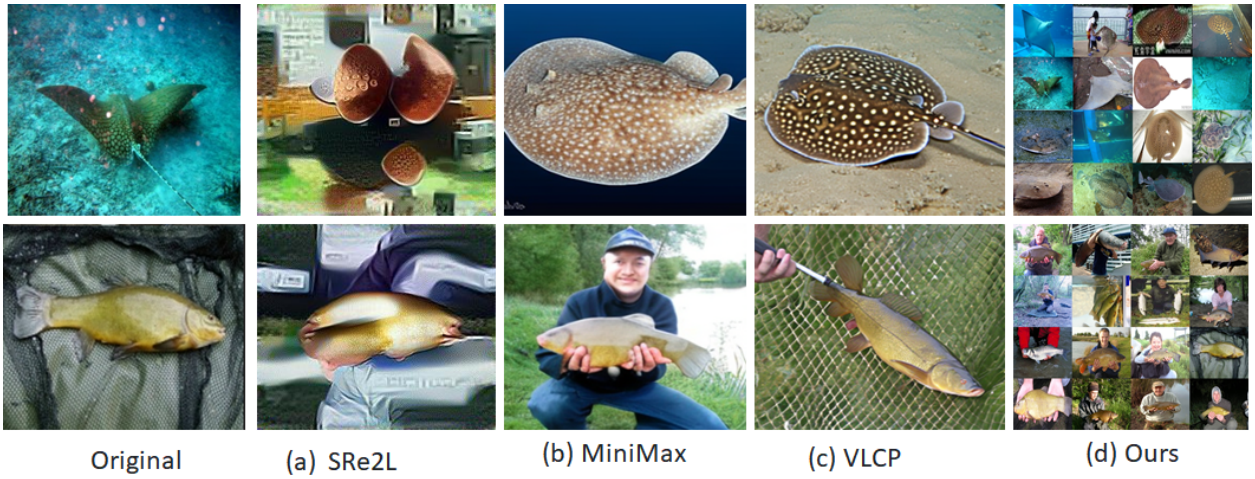

Figure 2: **Qualitative comparison of distilled images across methods.** We compare representative distilled samples generated by SRe^{2}L, Minimax, VLCP, and our Grid Distillation framework. SRe^{2}L often produces visually coarse patterns with limited semantic localization, while Minimax shows sharper appearance but tends to collapse onto a few characteristic modes. VLCP yields cleaner textures but still is limited by amount of information captured in an image. Our method produces visually coherent grid using clip's world knowledge, compositionally rich images with clearer grid boundaries and more balanced semantic coverage, reflecting the stronger representational properties enforced by grid-based spectral submodular selection.

underrepresents secondary object modes and fails to capture fine geometric variation.

In contrast, our method benefits from grid-level spectral coverage and CLIP-informed compositional selection, generating samples that maintain semantic breadth while recovering spatially consistent object layouts after diffusion reconstruction. These qualitative improvements align with our quantitative results, highlighting that structured grid selection paired with diffusion priors leads to both richer and more representative distilled imagery.

4 Grid-Aware Cropping

To ensure compatibility between our 1024×1024 distilled grids and standard 224×224 or 256×256 classifier inputs, we introduce a **Grid-Aware Cropping** strategy. Unlike naïve random cropping—which may cut across semantic partitions inside the grid—our method probabilistically aligns crops to the underlying $L \times L$ grid structure.

Given a 1024×1024 compositional detailed enhanced grid or bilinear up-sampling, we sample 256×256 patches using an alignment probability p. With probability p, the crop origin is restricted to multiples of the sub-grid size (e.g., 0, 256, 512, 768), preserving spatial boundaries and semantic consistency. With probability $1 - p$, we sample a fully random crop, encouraging robustness to shifts and local perturbations.

This simple biased-cropping mechanism significantly improves downstream discriminability, helping the model attend to semantically complete grid regions while still maintaining variability. We provide the implementation used in all experiments below.

```python
class GridAwareCrop:
    """
    A PyTorch transform that crops a 256x256 patch from a 1024x1024 grid image (Bilienar/Detail Enhanced).
    With probability `align_prob`, the crop is grid-aligned (starts at multiples of 256),
    and with (1 - align_prob) it's a random crop.
    """

    def __init__(self, crop_size=256, grid_size=1024, align_prob=0.6):
        self.crop_size = crop_size
        self.grid_size = grid_size
        self.align_prob = align_prob

    def __call__(self, img):
        if isinstance(img, torch.Tensor):
            img = F.to_pil_image(img)

        w, h = img.size
        if w != self.grid_size or h != self.grid_size:
            raise ValueError(f"Expected {self.grid_size}x{self.grid_size} image, got {w}x{h}")

        # Aligned vs random
        if random.random() < self.align_prob:
            start_x = random.choice(range(0, self.grid_size, self.crop_size))
            start_y = random.choice(range(0, self.grid_size, self.crop_size))
        else:
            start_x = random.randint(0, self.grid_size - self.crop_size)
            start_y = random.randint(0, self.grid_size - self.crop_size)

        return F.crop(
            img,
            top=start_y,
            left=start_x,
            height=self.crop_size,
            width=self.crop_size
        )

    def __repr__(self):
        return (f"{self.__class__.__name__}(crop_size={self.crop_size}, "
                f"grid_size={self.grid_size}, align_prob={self.align_prob})")
```

5 Ablation Supplement: Effect of Using Action Labels in Prompts for Grid Enhancement

We perform an ablation study on the UCF101 motion–modulated frames to analyze the impact of using action labels in the text prompt during grid enhancement. We compare two variants: (i) **Action-Aware Enhancement**, where the prompt explicitly includes the ground-truth action category, and (ii) **Action-Agnostic Enhancement**, where only a generic visual enhancement description is used.

Including the action label injects high-level semantic priors (e.g., "Apply Eye Makeup", "Playing Guitar", "Clean And Jerk"), allowing the enhancement model to emphasize on action relevant regions such as limb boundaries and object interactions. Without action labels, the model relies solely on local pixel statistics, often producing weaker motion sharpening and less discriminative textures.

Example Prompts. With Action Label Template:

```
f'High Quality 4x4 Grid, Person performing {action} action.
high-contrast, photo-realistic, 8k, ultra HD,
meticulous detailing, hyper sharpness,
perfect without deformations'
```

Without Action Label:

```
High Quality 4x4 Grid, high-contrast,
photo-realistic, 8k, ultra HD, meticulous detailing, hyper sharpness,
perfect without deformations
```

6 Detail-Enhanced Reconstructions Across Categories

In this section, we provide additional qualitative results showcasing detail-enhanced reconstructions produced by our method across a diverse set of image categories. Each category exhibits unique structural and textural characteristics, including fine-grained object shapes, natural textures, and complex background patterns. The presented images are the final high-resolution 1024×1024 outputs obtained after applying our detail enhancement module to the distilled representations.

Across all eight categories, our approach consistently restores high-frequency details that are otherwise absent from the low-resolution distilled inputs. The enhanced images demonstrate sharper edges, richer textures, and more coherent object boundaries while preserving the semantic content of the original scene. These subjective results highlight the generalization capability of our enhancement strategy and its ability to adapt to varying visual domains without introducing drastic artifacts or hallucinations. Overall, the restored outputs validate the effectiveness of our diffusion-guided refinement process across a broad range of visual semantics.

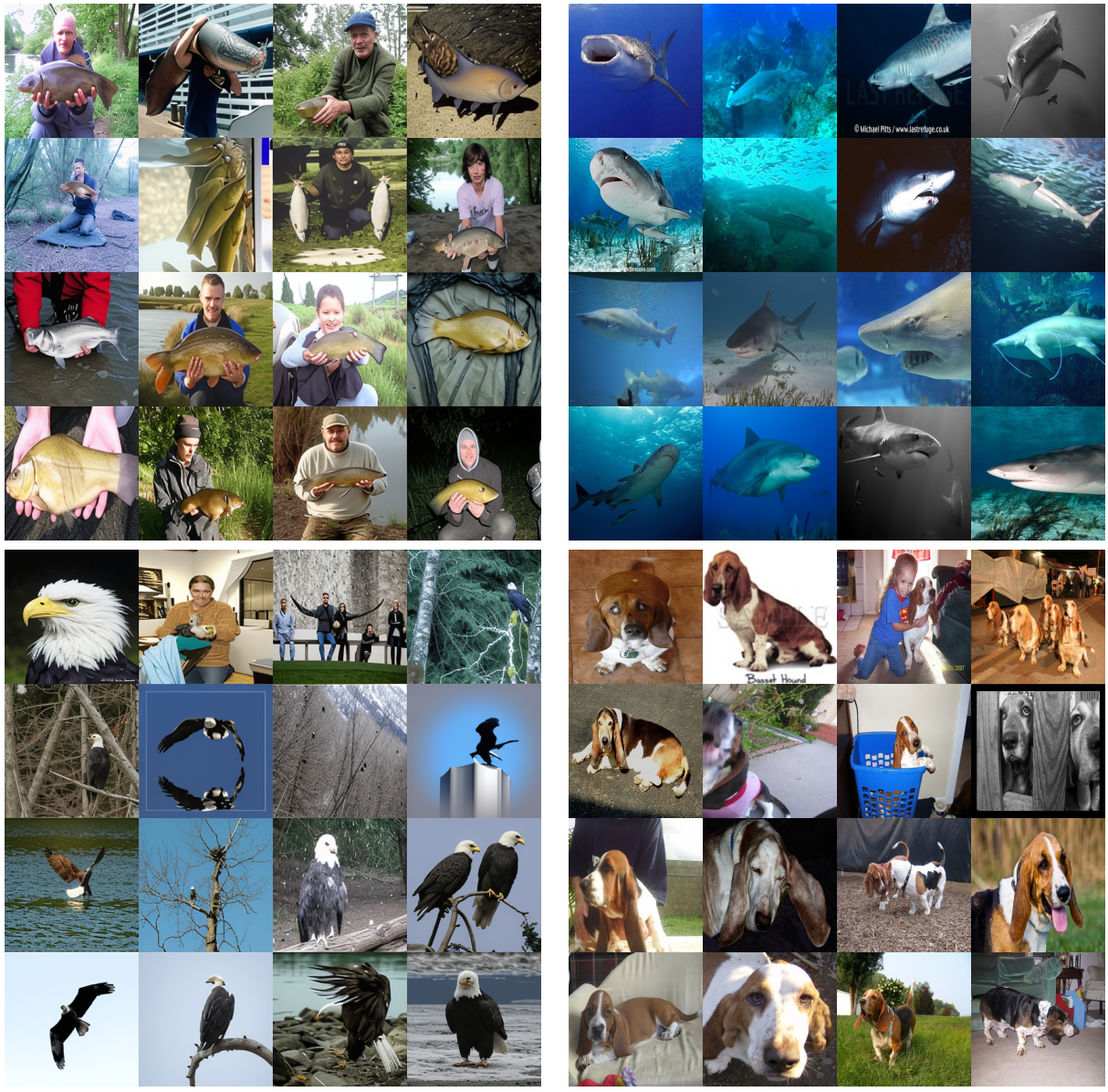

Figure 3: Detail-enhanced reconstructions across four visual categories. Our method restores fine textures and semantic structure across diverse domains, demonstrating strong generalization in the enhancement stage.

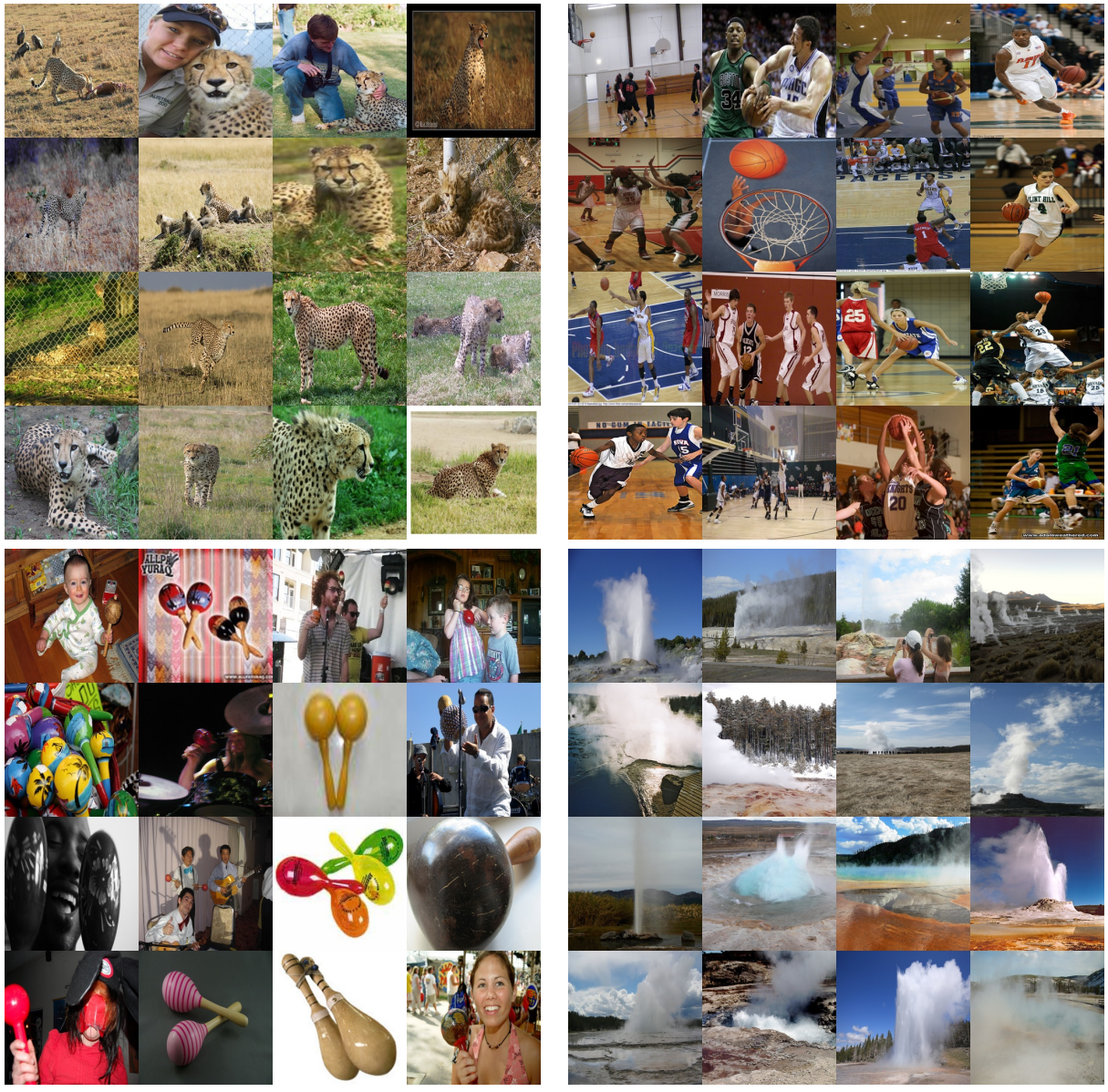

Figure 4: Detail-enhanced reconstructions across additional visual categories. Our method restores fine textures and semantic structure across diverse domains, demonstrating strong generalization in the enhancement stage.